10.37965/jdmd.2026.1552

A Rolling Bearing Fault Diagnosis Method Based on Scaled Dot-Product Attention

Changgang Yan，zhichao Cong，Jinlong Wang

College of mechanical engineering, Dalian Jiaotong University, Dalian 116028, China

*Corresponding author: congzhichao1@163.com

Received 20 June 2026; Revised 13 August 2026; Accepted 18 August 2026; Published online 19 August 2026

Abstract: To tackle the persistent challenges of low bearing fault diagnosis accuracy—specifically the difficulty of extracting faint fault features under complex, variable operating conditions and strong background noise, as well as the tendency to lose deep temporal dependencies—this paper proposes a novel diagnosis method integrating FFT-VMD feature extraction with a Bi-TCN-Bi-GRU neural network. First, to overcome the heuristic parameter selection and frequency-band aliasing inherent in traditional Variational Mode Decomposition (VMD), the Fast Fourier Transform (FFT) is introduced. By extracting global spectral prior information, the FFT guides the VMD to perform adaptive decomposition and effective denoising of non-stationary vibration signals. Subsequently, a dual-path Bi-TCN-Bi-GRU diagnostic model is constructed. The dilated causal convolution mechanism of the Bidirectional Temporal Convolutional Network (Bi-TCN) is utilized to extract deep local spatial features, while the Bidirectional Gated Recurrent Unit (Bi-GRU) is integrated to deeply mine the forward and backward dynamic evolutionary dependencies within the sequence. Experimental results demonstrate that even under severe background noise with a signal-to-noise ratio (SNR) of -2 dB, the proposed method maintains an exceptional diagnostic accuracy of 98.21% on a self-built bearing dataset and 99.90% on the Southeast University (SEU) dataset. This study provides a highly robust and promising solution for enhancing bearing safety and predictive maintenance in complex industrial environments.

Keywords: rolling bearing fault diagnosis; variational mode decomposition; fast fourier transform; bidirectional gated recurrent unit; scaled dot-product attention;

I. INTRODUCTION

The operating condition of rolling bearings directly determines the safety and reliability of modern industrial rotating machinery. However, in real-world industrial scenarios, bearing vibration signals often exhibit strong nonlinear and nonstationary characteristics, and early, subtle fault features a re easily masked by complex alternating loads and strong background noise[1-3]. Traditional diagnostic methods based on the time-frequency domain or simple parameter analysis have proven inadequate for effectively separating fault impulse components from background noise under these complex operating conditions[4,5]. Consequently, exploring efficient, adaptive deep learning-based fault diagnosis methods under high-noise and non-stationary conditions has become a leading research frontier in this field[6−8].

In the subfield of front-end signal processing, Variational Mode Decomposition (VMD) has demonstrated significant noise-reduction advantages due to its adaptive, anti-aliasing mathematical framework[9-11]. However, the decomposition performance of VMD is constrained by the selection of parameters (the number of modes K and the penalty factor), and the arbitrary nature of parameter settings can easily lead to over-or under-decomposition[12]. Although recent studies(such as Wang et al.[13] and Ding et al.[14]) have attempted to introduce the Fast Fourier Transform(FFT) to obtain frequency-domain priors for guiding VMD, these methods still lack sufficient robustness and convergence efficiency—enhanced by their adaptive characteristics—when dealing with extremely variable operating conditions, and have failed to completely overcome the limitations of empirical parameter setting. To address this issue, this paper proposes a more direct and efficient FFT-guided adaptive VMD processing strategy, which directly quantifies and guides VMD parameters using global spectral priors, thereby achieving precise decoupling and denoising of the original non-stationary signal.

In the field of deep feature space mining, methods represented by temporal convolutional networks(TCN) and bidirectional gated recurrent units (Bi-GRU) have become mainstream for capturing multiscale spatial and temporal features[15, 16]. Although recent studies (such as Saghi et al. [17] and Xu et al. [18])have explored the potential of combining multiscale CNNs with Bi-GRU under non-stationary conditions, most current multimodal network designs still rely on simple concatenation or serial stacking of features. Such loose architectures struggle to effectively couple the local deep receptive fields of Bi-TCN with the global, long-range, forward-backward dynamic evolution logic of Bi-GRU in deep space [19,20].Consequently, under strong transient disturbances, existing models are highly prone to losing critical temporal dependencies. To address this, this paper innovatively constructs a dual-path feature fusion network—Bi-TCN-Bi-GRU—which effectively bridges the feature gap between local feature depth and global temporal logic present in single-network architectures.

Furthermore, under the interference of strong background noise (such as extremely low signal-to-noise ratios), weight imbalance in feature representations poses another core challenge that limits the clarity of the model's classification boundaries[21]. Although channel or spatial attention mechanisms have been widely adopted to suppress background noise [22,23], they often overlook cross-view interaction information within multiscale feature sequences, making them prone to feature distortion under strong noise. Scaled Dot-Product Attention (SDA), as a core component of the Transformer, can effectively capture cross-view interactions through dot-product similarity metrics [24-26];however,its targeted application in the fusion of multi-scale features under strong noise remains insufficient. This paper specifically introduces the SDA mechanism to calculate interaction weights between multiscale views, adaptively enhancing the components critical to fault detection and fundamentally maintaining the model's high-precision recognition performance under harsh conditions.

In summary, addressing the three core challenges in industrial field environments—insufficient adaptive guidance of VMD under complex and variable operating conditions, the difficulty of a single network in balancing deep spatial and bidirectional temporal dependencies, and feature expression imbalance under strong noise—this paper proposes a bearing fault diagnosis method based on FFT-VMD feature extraction and the synergistic optimization of Bi-TCN and Bi-GRU. Unlike the "module stacking" approach commonly found in existing research, the substantive innovation of this paper lies in proposing a synergistic design architecture tailored specifically for extreme operating conditions. The necessity of this tight coupling between modules is reflected in the progressive layers of physical logic: First, the front end relies on FFT-VMD to provide adaptive physical denoising, effectively preventing subsequent deep networks from overfitting to background noise; second, the middle layer utilizes Bi-TCN to capture local multiscale transient impacts, while Bi-GRU identifies global long-range degradation trends, thereby achieving intrinsic complementarity between spatio-temporal features; finally, the SDA in the back end serves as a critical coupling bridge, specifically designed to compute cross-view interaction weights in the dual spatio-temporal feature spaces, thereby adaptively amplifying fault-sensitive components under extremely low signal-to-noise ratios. The main contributions of this paper are as follows:

1.We propose an FFT-guided adaptive VMD processing strategy. By utilizing global spectral information obtained via FFT as prior guidance, we address the blindness of VMD in parameter tuning and achieve efficient decoupling and denoising of the original vibration signal[27].

2.A dual-channel feature fusion network, Bi-TCN-Bi-GRU, was constructed. By using Bi-TCN to extract local, deep, multiscale spatial features and combining it with Bi-GRU to uncover the long-range temporal evolution logic of feature sequences, a comprehensive characterization of the complex feature space was achieved[28-30].

3.A Scaled Dot-Product Attention (SDA) mechanism was introduced. By calculating the interaction weights between multiscale views, the representation of fault-critical components was adaptively enhanced, significantly improving the model's robustness in high-noise environments (SNR =-2dB)[31,32].

4.Extensive cross-dataset validation. Systematic validation was conducted on our self-built dataset and the Southeast University (SEU) dataset. The results show that under mixed operating conditions and in high-noise environments, the accuracy of our method reached 98.21% and 99.90%, respectively, significantly outperforming existing ablation models and mainstream benchmark methods[33].

II. SYSTEM DESIGN AND OPTIMIZATION METHODS

A. Model Preprocessing Section and Solution of the VMD Variational Model

The core of VMD is to decompose the original signal into multiple intrinsic mode functions (IMFs) with finite bandwidths that are mutually independent. The core process for bandwidth quantification and decomposition is as follows: perform a Hilbert transform on each IMF, and combine it with an impulse function to obtain the one-sided spectrum and the decomposed signal; introducing the center frequency of each IMF (ω_n), shifting the analytical signal via an exponential term ($e^{-j\omega_n t}$) to move the spectrum to the baseband centered at (ω_n); calculating the time-domain partial derivatives and norms of the shifted analytical signal; and combining the objective of “minimizing the sum of bandwidths” with the constraint of “reconstructing the original signal through superposition,” ultimately formulating the variational problem for VMD decomposition.

$$\min_{\mu_i,\omega_i}\left\{\sum_i \left\|\partial_t\left[\left(\delta(t)+\frac{j}{\pi t}\right)^* \mu_i(t)\right]e^{-j\omega_i t}\right\|_2^2\right\} \quad (1)$$
$$\text{s.t.}\sum_i \mu_i = f$$

In the equation, $\mu_i(t)$ denotes the i th IMF to be solved (where = 1, 2, ..., N, and N is the total number of optimal modes in the VMD decomposition), $\delta(t)$ is the Dirac impulse function, ∂_t is the time-domain derivative operator, $\|\cdot\|_2^2$ is the squared L2 norm of the signal, and s.t. is the constraint.

By introducing a penalty term and a Lagrangian multiplier, the constrained variational problem is transformed into an unconstrained optimization problem.

$$L=\alpha\sum_{i=1}^{N}\left\|\partial_t\left(u_i^+(t)e^{-j\omega_i t}\right)\right\|_2^2+\left\|f(t)-\sum_{i=1}^{N}u_i(t)\right\|_2^2 +\langle\lambda(t), f(t)-\sum_{i=1}^{N}u_i(t)\rangle \quad (2)$$

Where L is the augmented Lagrangian function, $u_i^+(t)$ is the iteratively updated value of the th IMF component (i), α is the penalty term, $\lambda(t)$ is the Lagrangian multiplier, and $\langle,\rangle$ denotes the inner product operation.

Solve the above variational problem using the Alternating Direction Method of Multipliers (ADMM) by iteratively updating

$\hat{u}_i(\omega)$, ω_i, $\hat{\lambda}(\omega)$ and then transform it to the frequency domain via the Parseval isotropy transformation to obtain the optimal solution for the modal components as:

$$\hat{u}_i^z(\omega)=\frac{\hat{f}(\omega)-\sum_{i=1, i\neq n}^{N}\hat{u}_i^{z-1}(\omega)+\frac{\hat{\lambda}^{z-1}(\omega)}{2}}{1+2\alpha\left(\omega-\omega_n^{z-1}\right)^2} \quad (3)$$

Where $\hat{u}_i^z(\omega)$ denotes the frequency-domain solution of the i th IMF component in the z th iteration, z is the iteration number, the superscript $z-1$ represents the computational result of the $z-1$ th iteration, $\hat{f}(\omega)$ is the frequency-domain form of the original vibration signal, ω_i denotes the center frequency corresponding to the i th IMF, and $\hat{\lambda}(\omega)$ is the frequency-domain form of the Lagrange multiplier.

The center frequency ω_n is updated as follows:

$$\omega_i^z=\frac{\int_0^\infty \omega\left|\hat{u}_i^z(\omega)\right|^2 d\omega}{\int_0^\infty\left|\hat{u}_i^z(\omega)\right|^2 d\omega} \quad (4)$$

The Lagrangian $\hat{\lambda}(\omega)$ is updated as:

Table 1. VMD parameters

Parameter	Value	Parameter	Value
α	2000	DC Component	0
τ	0	Convergence Tolerance	1×10^{-7}
Initialization Method	1	N	2–10

10.37965/jdmd.2026.1552

$$\hat{\lambda}^{z}(\omega)=\hat{\lambda}^{z-1}(\omega)+\tau\left(\hat{f}(\omega)-\sum_{i=1}^{N}\hat{u}_{i}^{z}(\omega)\right) \quad (5)$$

where τ is the update step size of the Lagrangian multiplier, $\hat{f}(\omega)-\sum_{i=1}^{N}\hat{u}_{i}^{z}(\omega)$ is the signal reconstruction error, and $\left|\hat{u}_{i}^{z}(\omega)\right|^{2}$ is the spectral energy density of the i th IMF component in the z th iteration.

The specific VMD parameters are shown in Table 1

B. Model Theory Section

Expanded causal convolutions build upon causal convolutions by adding an expansion operation to establish temporal constraints. Specifically, they ensure that only past information is used to predict the future, while guaranteeing that no feature information from past time steps is omitted. By introducing an expansion factor, interval sampling can be performed on the time series, achieving the broadest coverage with the fewest convolutional layers. Here, padding refers to the padding neurons. The convolution output at any given time, $y=(y_0,y_1,y_2,\ldots,y_T)$, depends only on the inputs preceding the current time, $x=(x_0,x_1,x_2,\ldots,x_T)$. The structure of the dilated causal convolution is shown in Figure 1.

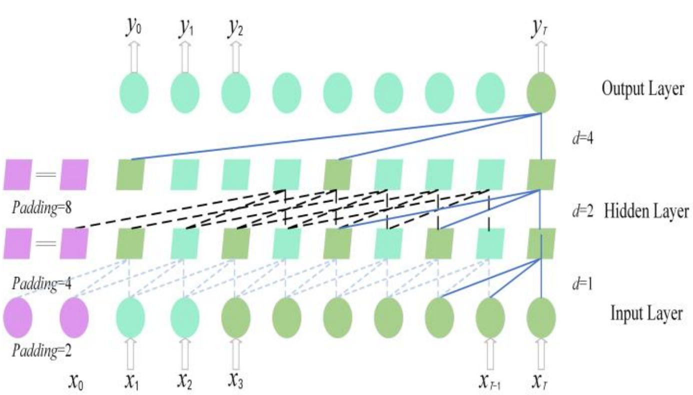


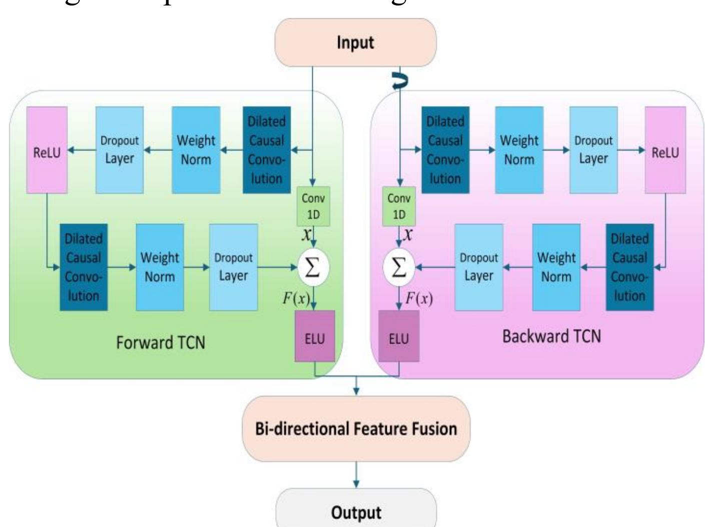
Figure 1. Expanded Causal Convolution Structure

The expanded causal convolution at time step t, using the filter $f=(f_0,f_1,f_2,\ldots,f_{k-1})$ as the convolution kernel, is expressed as follows:

$$A(t)=\sum_{k-1}^{i=0}f(i)x_{t-i\cdot d} \quad (6)$$

where: k is the size of the convolution kernel; $x_{t-i\cdot d}$ is the time series after interval sampling; d is the expansion factor, which generally increases as a power of 2

The core functionality of Bi-TCN is to extract multiscale local temporal dependency features and long-range contextual correlations from preprocessed time-series vibration signals. The preprocessed time-series feature sequences are fed into the Bi-TCN network as input, where bidirectional causal convolution, hole convolution, and nonlinear activation operations are employed to achieve multiscale feature extraction and bidirectional encoding of temporal information. This process enhances features and reduces dimensionality while preserving the complete temporal structure, thereby obtaining more discriminative fault-sensitive features. These features provide critical support for subsequent temporal modeling and high-precision fault pattern recognition. The Bi-TCN feature extraction process for a single sample is shown in Figure 2.



10.37965/jdmd.2026.1552

Figure 2. BiTCN Feature Extraction Process

GRU is an improved architecture of LSTM, consisting of a reset gate r and an update gate z. The update gate z controls the proportion of the previous hidden state h_{t-1} transmitted to the current state, while the reset gate r determines the amount of information to be ignored from the historical states. Both gate signals are jointly computed from the state h and the input x as follows:

$$\begin{aligned} z_t &= \sigma[W_z(h_{t-1}, x_t)], \\ r_t &= \sigma[W_r(h_{t-1}, x_t)], \\ \tilde{h}_t &= \tanh[W_{\tilde{h}}(r_t * h_{t-1}, x_t)], \\ h_t &= (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \end{aligned} \tag{7}$$

Where: σ is the sigmoid function; W_z , W_r and $W_{\tilde{h}}$ are weights; $\tilde{h}$ is the current time step; h_t is the current hidden state output; z_t is the update gate; $z_t \in (0,1)$; the closer z_t is to 0, the more information from the current time step is retained.

Bi-GRU consists of a forward GRU and a reverse GRU operating in parallel. This architecture simultaneously processes the forward and reverse information flows of the input sequence, generating context-aware vector representations through a feature fusion mechanism. The Bi-GRU structure is shown in Figure3.

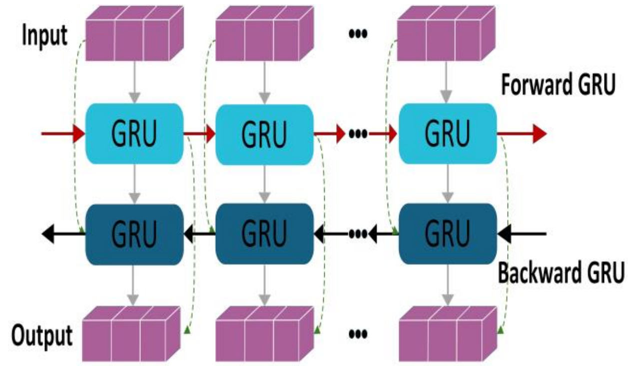


Figure 3. BiGRU Feature Extraction Process

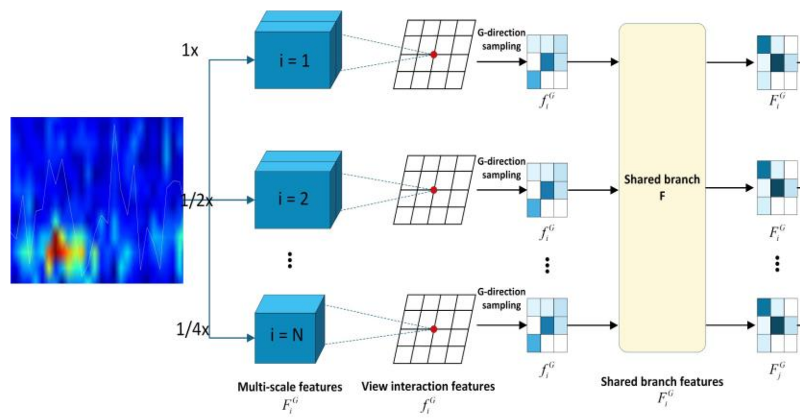


Figure 4. Scaled Dot-Product Attention Mecha

C. Scaled Dot-Product Attention Mechanism

Figure 4 illustrates the structure of the fusion dot-product module (FDCM), whose core lies in introducing view interaction information between different views as a shared branch.

Specifically, the Fusion Dot-Product Module consists of a three-layer structure, with view interaction information shared between layers. The computational process is as follows: First, in the direction of observation d , bilinear sampling is performed on the multiscale feature information extracted by the multiscale feature extraction network to obtain the view interaction feature information F, which serves as the shared branch; then, based on the multi-scale features F and the view interaction features F_m , the shared features (F_i^G, F_j^G) are computed. The formula for F_m is as follows: Its specific composition consists of the following two parts:

$$FDCM:\left(F_i^G,F_j^G\right)\rightarrow\left(f_i^G,f_j^G\right),\forall i,j\in\{1,2,\ldots,N\},i\neq j,$$
$$F_m=\left\{\cos\left(f_i^G,f_j^G\right)\mid i,j\in\{1,2,\ldots,N\},i\neq j\right\} \qquad (8)$$

Here, FDCM denotes the fusion dot product module; F_i^G represents the multiscale features of the i th image in the sampling direction G; f_i^G represents the multiscale features of f after being processed by the fusion dot product module.

D. Overall Fault Diagnosis Workflow and Data Flow

The overall architecture for bearing fault diagnosis proposed in this paper, based on FFT-VMD and BiTCN-BiGRU-SDA, is shown in Figure 5. The model consists of three core stages: signal preprocessing, dual-branch feature extraction, and SDA attention-based fusion classification. The specific data flow process is as follows:

1. Signal Preprocessing Stage: The acquired raw non-stationary time-domain vibration signal $X\in R^{1\times N}$ is first subjected to an FFT transform to obtain the global spectral distribution, which is used as prior information to determine the optimal number of decomposition modes K for VMD.VMD adaptively decouples the raw signal into K finite-bandwidth IMF components and reconstructs them into a multi-channel time-frequency feature matrix $S_{IMF}\in R^{K\times N}$, thereby achieving preliminary filtering of strong background noise.

2. Dual-branch feature extraction stage: The feature matrix S_{IMF} is fed in parallel into two complementary feature extraction branches: The Bi-TCN branch utilizes the multiscale receptive fields of dilated causal convolutions to extract local deep-space features of the signal in parallel from both the left-to-right and right-to-left directions, outputting the temporal feature

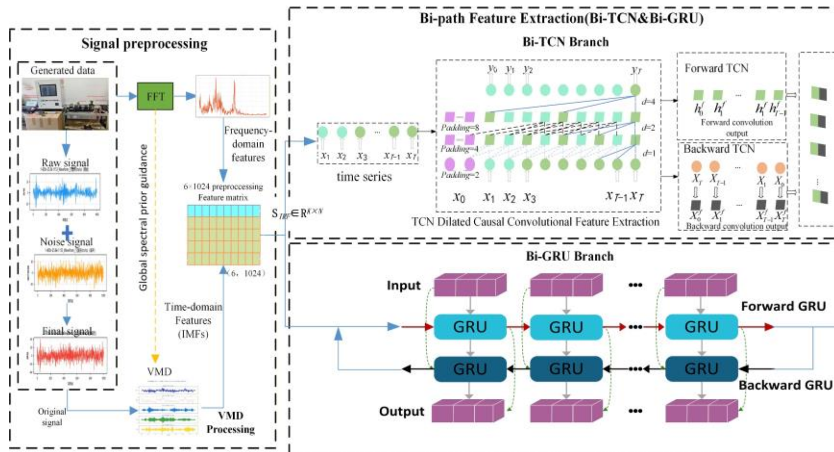


Figure 5. Full Model Flowchart

representation $F_{TCN}\in R^{L\times d}$; the Bi-GRU branch uses forward and backward gated recurrent units to deeply explore the long-range dynamic evolutionary dependencies in the sequence,outputting the temporal correlation representation $F_{GRU}\in R^{L\times d}$.

3. SDA Attention Fusion and Classification Stage: The feature matrices output by the two branches are combined to form a cross-view feature set. A Scaled Dot-Product Attention (SDA) mechanism is introduced, where F_{TCN} and F_{GRU} serve as the Query and Key, respectively, to compute a dot-product similarity matrix. Attention weights are adaptively assigned to dynamically enhance feature neurons sensitive to fault impacts. Finally, the weighted-fused features undergo Global Average Pooling (GAP) and a fully connected layer (FC), and are mapped through the Softmax activation function to

output classifications for various bearing fault diagnoses.

III. Bearing Failure Experiment Design

To verify the model's accuracy and robustness, we used both a self-built dataset and a publicly available dataset from Southeast University for validation.

A. Basic Structure of the Bearing Test Machine

The experiments rely on a self-built bearing failure simulation test platform; the detailed list of mechanical components and the on-site layout are shown in Figure 6.In terms of physical architecture, the system is supported by a sturdy welded base and consists of a drive motor, a coupling, a main shaft, and a test bearing housing equipped with the bearing under test, all connected in series. The core control and data acquisition process is as follows: the host computer (laptop) sets the target rotational speed and sends it to the PLC control cabinet on the left, which then drives the rotor system to operate. To simulate real-world engineering loads, a bolted loading mechanism is mounted on top of the test bearing housing on the right side of the platform. Combined with a pressure sensor and the force display controller below, this mechanism enables the application and monitoring of a stable radial load. The faint mechanical vibrations generated during operation are captured by a high-sensitivity accelerometer, then converted from analog to digital via an ART data acquisition card, and recorded on the computer for data storage and analysis.

Figure 6. Bearing Testing Machine

During actual operation of rolling bearings, critical components such as the inner ring, outer ring, and rolling elements may sustain localized damage due to long-term alternating loads, contact fatigue, and wear. To simulate real bearing failure conditions as closely as possible, this study used wire cutting to create inner ring and rolling element failures, respectively, and employed an angle grinder to simulate outer ring spalling. Additionally, normal bearings with no obvious damage were selected as control samples. Figure 7 shows photographs of a normal bearing and three types of typical failed bearings.

10.37965/jdmd.2026.1552

Table 2. Classification Results of the Self-Built Bearing Failure Dataset

Operating Condition	Failure Type	Label	Training Set	Validation set	Test set
	Inner Ring Failure	0	279	78	39
Low Speed Heavy Load	Outer Ring Failure	1	279	78	39
	Rolling Element Failure	2	279	78	39
	Normal Condition	3	279	78	39
	Inner Ring Failure	4	279	78	39
Medium speed Medium load	Outer Ring Failure	5	279	78	39
	Rolling Element Failure	6	279	78	39
	Normal Condition	7	279	78	39
	Inner Ring Failure	8	279	78	39
High Speed Light Load	Outer Ring Failure	9	279	78	39
	Rolling Element Failure	10	279	78	39
	Normal Condition	11	279	78	39

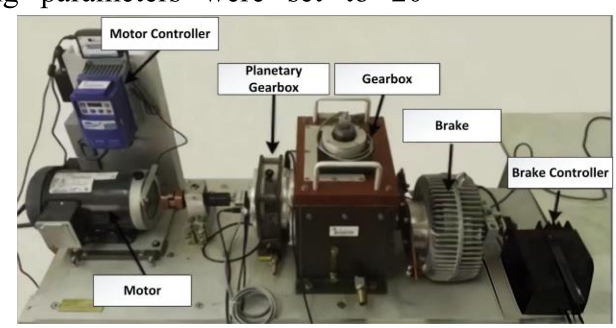
Figure 7. Photographs of a normal bearing and typical failed bearings

B. Data Acquisition and Design of the Experimental Software Environment

The test utilized vibration signals from Type 31306 tapered roller bearings to construct the experimental dataset. The bearing operating states included four categories: normal, inner ring failure, outer ring failure, and rolling element failure. To simulate the operating characteristics of bearings under various industrial conditions, three sets of speed–load combinations were established: 800 r/min + 3000 N(low speed, heavy load), 1100 r/min + 2000 N (medium speed, medium load), and 1400 r/min + 1000 N (high speed, light load), resulting in a total of 12 operating condition labels.

construction, a time-series window of length 1,024 and a 50% sliding overlap rate were used. The data was randomly divided into training, validation, and test sets in a 7:2:1 ratio; the training and test sets were derived from the same original signal sequence. To simulate the imbalanced nature of industrial field samples, random sample missing values were introduced into the dataset, and Gaussian white noise was added to verify the model's resistance to interference. The distribution of samples across each operating condition is detailed in Table 2.

The training parameters were set to 20 epochs, a batch size of 32, and a learning rate of 0.0003. Training and testing were conducted on a 64-bit Windows 11 system with CUDA version 12.1; the deep learning framework used was PyTorch, and the programming language was Python 3.9.21. The hardware configuration consists of an Intel Core i7-14650HX CPU, an RTX 4060 Laptop GPU, and 16 GB of RAM.



10.37965/jdmd.2026.1552

Table 3. Classification Results of the SEU Bearing Failure Dataset

Operating Condition	Failure Type	Label	Training Set	Validation set	Test set
	Inner Ring Failure	0	7161	2046	1023
1200r/min	Outer Ring Failure	1	7161	2046	1023
0N	Rolling Element Failure	2	7161	2046	1023
	Combined Failure	3	7161	2046	1023
	Normal Condition	4	7161	2046	1023

C. Southeast University Public Dataset

To further validate the generalization capability of the method proposed in this paper across different data domains, we selected the publicly available dataset from the Simulation Vibration Laboratory at Southeast University to conduct cross-dataset validation experiments. Compared to the in-house dataset, this dataset exhibits significant differences in terms of the test platform, sensor layout, sampling configuration, and fault types, thereby effectively testing the model's ability to adapt to data domain shifts.

This dataset was collected from the Simulation Vibration Laboratory at Southeast University. The experimental test platform consists of a drive motor, motor controller, planetary gearbox, reduction gearbox, brake, and brake controller, as shown in Figure 8.

Figure 8. Bearing Testing platform

brake controller. A total of seven 608A11 type vibration sensors were installed to capture multi-directional vibration data from the rolling bearings at the interfaces between the motor and the planetary gearbox and reduction gearbox.This paper uses signals from the drive-end vibration channel (de channel) with a sampling frequency of 5120 Hz for the experiments. The dataset covers five categories of bearing operating conditions: healthy, inner ring failure, outer ring failure, rolling element failure, and composite failure. The failure types range from single localized damage to multi-site composite damage, providing a comprehensive reflection of typical failure modes in actual in-service bearings.

The experimental operating speed was set to 1,200 r/min; the specific dataset is shown in Table 3. This operating condition overlaps with but is not entirely identical to the speed range of the self-built dataset, which helps evaluate the model's adaptability across different speed ranges.To ensure consistency with the in-house dataset and facilitate a fair comparison, the sample construction followed the same preprocessing workflow as described in Section IIIB: samples were constructed by sliding time-series windows on one-dimensional vibration signals and randomly split into training, validation, and test sets in a 7:2:1 ratio; to simulate the strong

10.37965/jdmd.2026.1552

background noise environment typical of industrial settings, Gaussian white noise with a signal-to-noise ratio of -2 dB was uniformly added to the test samples. The model architecture, hyperparameters, and training settings were consistent with those used in the experiments on the in-house dataset; no additional hyperparameter tuning was performed specifically for this dataset.

IV. Model Training and Results Analysis

A. Ablation Experiments and Comparative Experiments

To verify the superior performance of the model designed in this paper for the bearing fault diagnosis task, we conducted validation studies using forward ablation experiments under the same experimental environment and data preprocessing workflow: Using the basic single-layer BiTCN model as the baseline, we progressively introduced the BiGRU module and the scaled dot-product attention mechanism. We then employed a backward ablation strategy to remove the BiGRU module and the ablation attention mechanism, respectively. Through quantitative comparisons, we demonstrated the performance differences among the models, verifying the core role of BiGRU's temporal modeling capabilities and the attention mechanism in enhancing fault diagnosis performance. The specific results are shown in Table 4.

Table 4. Experimental Results of Ablation Experiments and the Proposed Model

Model	Accuracy/%	Precision/%	Recall(%)
BiTCN-SDA	78.57	82.57	77.12
BiTCN-BiGRU	78.12	83.03	77.93
BiTCN- BiGRU-SDA	98.21	98.48	98.06

To verify the overall performance of the model designed in this paper and assess the balance between improved fault diagnosis accuracy and computational burden, we conducted a statistical analysis of the number of trainable parameters and the total computation time over 20 cycles for all models involved in the ablation experiments. The comparison results are shown in Table 5:

Table 5. Comparison of the Number of Parameters and Total Time Consumed Over 20 Cycles for Ablation Models

Model	Number of Parameters/k	Total Time Consumed for 20 Cycles/s
BiTCN-SDA	89.4	89
BiTCN-BiGRU	107.8	104
BiTCN- BiGRU-SDA	130.4	118

As shown in Table 5, both the number of trainable parameters and the training time for 20 cycles for each model in the ablation experiments exhibit an increasing trend consistent with structural complexity. The base model, BiTCN-SDA, has 89.4k parameters and takes 89s to train; the full model, BiTCN-BiGRU-SDA, has 130.4k parameters and takes 118s to train. The data indicate that although the introduction of the BiGRU and SDA modules incurs some computational overhead, the overall time cost increases only marginally, successfully achieving a good balance between improved fault diagnosis accuracy and reasonable computational load.

To demonstrate the advantages and advanced nature of the model proposed in this paper in the field of bearing fault diagnosis, as well as to eliminate random interference and evaluate the model's comprehensive engineering performance, we selected the domain-general benchmark model CNN, the mainstream time-series model CNN-BiLSTM, and the attention-enhanced model CNN-Transformer-CA—all under the same experimental environment and with similar parameter counts—to conduct side-by-side comparison experiments with the model proposed in this paper. All comparison experiments were repeated independently five times under identical hardware and dataset conditions, with the results shown in Table 6.

Table 6. Comparative Experimental Results and Results of the Proposed Model

Model	Accuracy/%	Precision/%	Recall/%
CNN	82.72±0.62	81.78±0.58	81.82±0.71
CNN-BiLSTM	88.91±0.51	89.71±0.49	88.56±0.63
CNN-Transformer-CA	92.87±0.43	93.11±0.41	92.11±0.52
BiTCN-BiGRU-SDA	97.67±0.28	97.72±0.35	97.91±0.35

The BiTCN-BiGRU-SDA model proposed in this paper significantly outperforms existing mainstream comparison algorithms in all three core classification metrics: accuracy, precision, and recall. Compared to the second-best CNN-Transformer-CA model, the diagnostic accuracy improved by 4.80%; compared to the classic benchmark CNN model, accuracy improved by 14.95 percentage points. Furthermore, the standard deviation of our model is significantly lower than that of the comparison models, fully validating its high robustness and training stability under complex operating conditions.

B. Training Accuracy and Loss Curves

To further validate the learning stability and generalization ability of the proposed model during the training process, we plotted the training accuracy and loss curves for each model. These curves allow for an intuitive analysis of the model's convergence efficiency, training stability, and risk of overfitting. The training accuracy curve and training loss curve are shown in the figure 9.

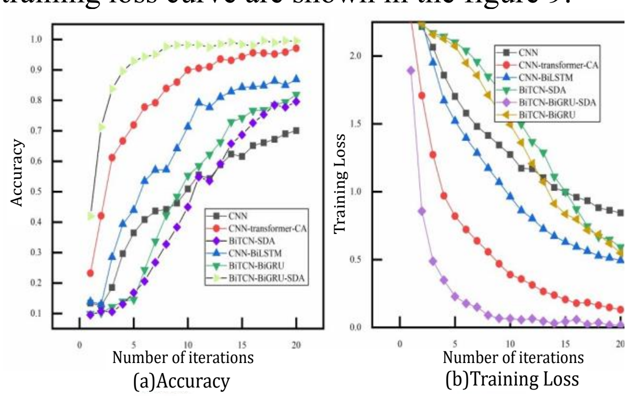
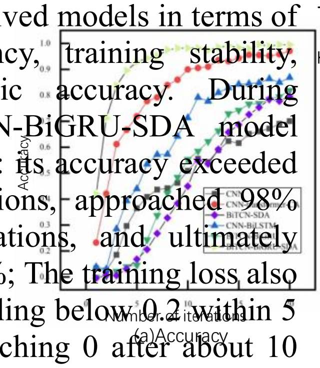


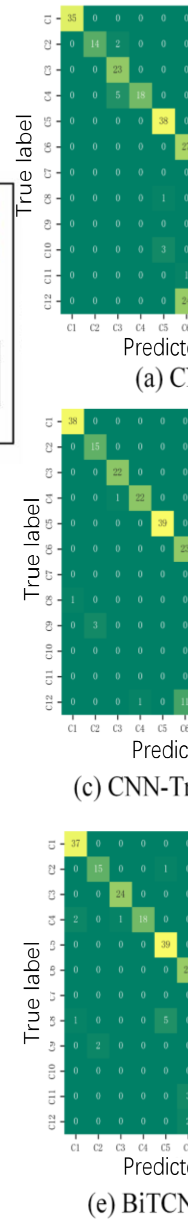
Figure 9. Training Accuracy and Loss Curves

As shown in Figure 9(a) (test accuracy curve) and Figure 9(b) (training loss curve), the proposed BiTCN-BiGRU-SDA model comprehensively outperforms all comparison and ablation-derived models in terms of convergence efficiency, training stability, and final diagnostic accuracy. During training, the BiTCN-BiGRU-SDA model converged the fastest: its accuracy exceeded 90% within 5 iterations, approached 98% after about 10 iterations, and ultimately stabilized at over 99%; The training loss also decreased rapidly, falling below 0.2 within 5 iterations and approaching 0 after about 10 iterations, with a smooth curve throughout and no significant fluctuations, demonstrating exceptional training stability. The baseline CNN model converged the slowest, achieving an accuracy of only about 70% after 20 training iterations, while the loss remained consistently high at over 0.8;The training performance of the remaining ablation models (CNN-BiLSTM, CNN-Transformer-CA, BiTCN-SDA, and BiTCN-BiGRU) progressively improved with architectural optimization, validating the synergistic and additive effects of the Bi-GRU temporal modeling module and the SDA attention mechanism on model performance.



C. Confusion Matrix Analysis and Feature Clustering Analysis

In bearing fault diagnosis, the confusion matrix is a core quantitative evaluation tool. Its value lies in overcoming the limitations of a single overall accuracy metric by quantifying the performance of individual fault class diagnosis through the distribution of true positives, false positives, and false negatives for each fault type. The following presents the results and analysis from this experiment, as shown in Figure 10.

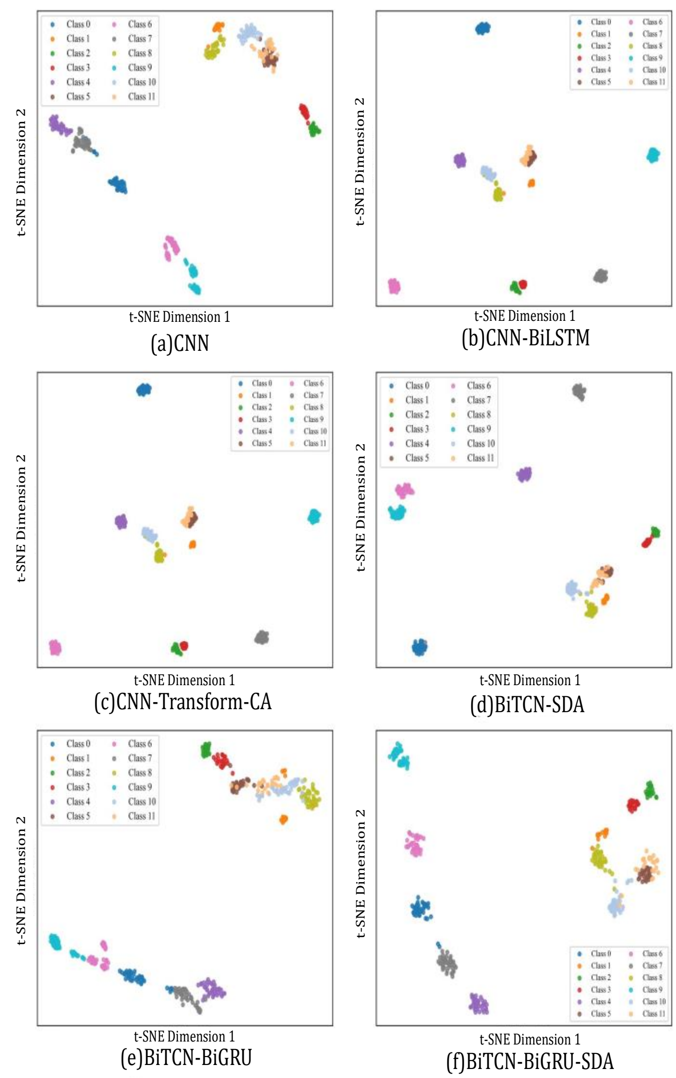


10.37965/jdmd.2026.1552

Figure 10. Confusion Matrices for Different Models

Compared with mainstream benchmark models in the field, such as CNN, CNN-BiLSTM, and CNN-Transformer-CA, the model proposed in this paper has the highest proportion of diagonal elements in the confusion matrix, with the fewest misclassified samples in the off-diagonal regions and the most concentrated distribution. The number of correctly identified samples remains high for the vast majority of failure categories, with only a very small number of misclassifications occurring among a few failure categories with extremely high feature similarity. The maximum number of misclassifications per category is far lower than that of all comparison models, achieving balanced and accurate identification of multi-category bearing failures.

In the bearing fault diagnosis task, the t-SNE feature clustering plot visually illustrates the feature distinctiveness of different bearing states through the clustering distribution of feature points. The following presents the results and analysis from this experiment, as shown in the figure11.



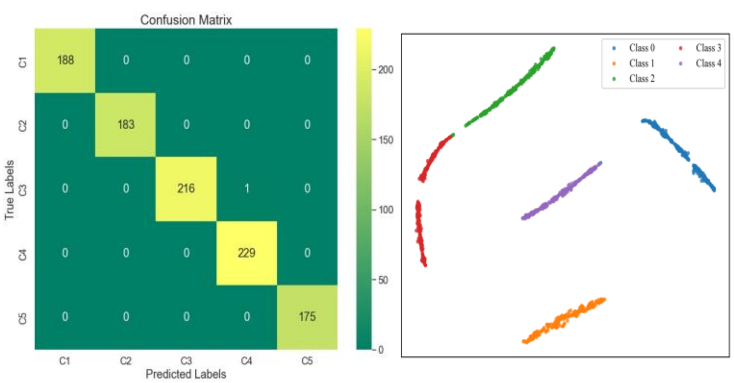
10.37965/jdmd.2026.1552

Figure 11. t-SNE Feature Clustering Results for Different Models

Comparative experiments further validate that, compared to all mainstream baseline and ablation-derived models, the BiTCN-BiGRU-SDA model proposed in this paper exhibits the most compact feature clusters, the lowest intra-class dispersion, the clearest inter-class boundaries, and the least category confusion. It thoroughly resolves the issue of feature overlap among similar faults, achieving precise separation and classification of bearing fault features.

D. Model Application

To validate the generalization capability of the BiTCN-BiGRU-SDA model, training and testing analyses were conducted under identical conditions using the Southeast University bearing dataset. The confusion matrix and clustering analysis results are shown in the figure12.



(a)Confusion Matrix (b)Feature Clustering

Figure 12. Confusion Matrix and Feature Clustering on the Southeast University Dataset

As shown in the confusion matrix in Figure 12(a), the model achieved a diagnostic accuracy of 99.90% on the Southeast University dataset, with precision and recall of 99.91% and 99.87%, respectively. The prediction results for each category are highly concentrated along the diagonal; the vast majority of samples were correctly classified. Only a very small number of misclassifications occurred between categories with high feature similarity, with no more than one misclassification per category, and the recognition performance across categories was relatively balanced. It is worth noting that the composite faults in the dataset exhibit overlapping features involving damage to multiple components—such as the inner ring, outer ring, and rolling elements—and exhibit feature coupling with single-fault modes; however, the model was still able to effectively distinguish them from each single-fault category, producing only one misclassification. This further validates the SDA attention mechanism's ability to capture multi-scale, cross-view interactive features.

As shown by the t-SNE feature clustering results in Figure 12(b), the feature clusters for the five sample categories are compact, with low intra-class dispersion and clear inter-class boundaries; there is no significant feature overlap between the different failure categories, and the spatial distribution of sample points exhibits distinct clustering characteristics. Compared with the clustering results from the self-built dataset in Section **IV C**, the model also achieved precise separation of fault features on the Southeast University dataset, indicating that the features extracted by the model possess strong domain-independent discriminative power and did not degrade significantly due to changes in the test platform, sensor configuration, or fault type.

The above results demonstrate that the method proposed in this paper can maintain high-precision diagnostic performance on completely different public datasets, with feature discriminative power comparable to that observed in the experiments on the self-constructed dataset. This confirms that the model possesses good cross-dataset generalization capability and engineering practical value under complex, variable operating conditions and in environments with strong background noise.

V. Conclusions

To address the challenge of low accuracy in bearing fault diagnosis models under complex operating conditions and strong background noise, we propose a bearing fault diagnosis method based on scaled point attention. We developed the BiTCN-BiGRU-SDA model and validated its reliability using both a self-built bearing dataset and the Southeast University dataset. The main conclusions are as follows:

(1)To address the low accuracy of mainstream models under complex operating conditions and strong background noise, we utilized VMD decomposition to extract time-domain features from the raw signal and FFT to extract frequency-domain features. We proposed the BiTCN-BiGRU-SDA bearing fault diagnosis model, which integrates time- and frequency-domain features, thereby resolving the low accuracy issue of mainstream models under such conditions.

(2)In a high-noise environment with a signal-to-noise ratio of -2 dB and a standard deviation of 2.58, the proposed BiTCN-BiGRU-SDA model achieved diagnostic accuracies of 98.21% and 99.90% on the self-built bearing dataset and the Southeast University bearing dataset, respectively.

(3)Designed to address the practical challenges of complex, variable operating conditions and strong background noise in industrial settings, the model presented in this paper possesses strong interference resistance and generalization capabilities, making it suitable for the online monitoring and diagnosis of bearing failures in rotating machinery at industrial sites. Future research may further explore the model's adaptability in transfer learning scenarios across different equipment and operating conditions, as well as lightweight optimization schemes to meet the requirements of real-time deployment at the edge.

(4)Given the characteristic of continuously varying operating speeds in industrial settings, the method proposed in this paper demonstrates good potential adaptability under variable-speed conditions: FFT-VMD uses global spectral priors to adaptively determine the number of decomposition modes without relying on fixed rotational speed priors, while the SDA mechanism can adaptively enhance fault-sensitive features in response to fluctuations in the signal-to-noise ratio. Existing experiments have covered three discrete speed conditions—low, medium, and high—and have been validated across datasets using the Southeast University dataset, preliminarily confirming the model's generalization capability across operating conditions. Due to limitations in the controllability of test bench conditions, experimental validation of continuous speed ramping processes has yet to be conducted; subsequent work will further evaluate its online diagnostic performance based on actual variable-speed operation data.

References

[1] J. Sun et al., "A robust approach to multi-sensor fusion for fault diagnosis using a convolutional neural network," J. Dyn. Monit. Diagn., vol. 1, pp. 103–110, 2022.

[2] W. Zhang et al., "A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environments and different working loads," Mech. Syst. Signal Process., vol. 100, pp. 439–453, 2018.

[3] H. Wang et al., "Fault diagnosis of bearings based on multi-sensor information fusion and a 2D convolutional neural network," IEEE Access, vol. 9, pp. 23717–23725, 2021.

[4] Z. Cao, J. Dai, and C. Chang, "Bearing fault diagnosis with frequency sparsity learning," IEEE Trans. Instrum. Meas., vol. 71, p. 3527210, 2022.

[5] R. Lu et al., "A novel fault diagnosis method based on NEEEMD-RUSLP feature selection and BTLSTSVM," IEEE Access, vol. 11, pp. 113965–113994, 2023.

[6] S. Neupane and J. Seok, "Bearing fault detection and diagnosis using the Case Western Reserve University dataset with deep learning approaches: A review," IEEE Access, vol. 8, pp. 93155–93178, 2020.

[7] M. Cerrada et al., "A review on data-driven fault severity assessment in rolling bearings," Mech. Syst. Signal Process., vol. 99, pp. 169–196, 2018.

[8] R. Zhao et al., "Deep learning and its applications to machine health monitoring," Mech. Syst. Signal Process., vol. 115, pp. 213–237, 2019.

[9] Y. Wang, R. Markert, and W. Zheng, "Research on variational mode decomposition and its application in detecting rub-impact faults in rotor systems," Mech. Syst. Signal Process., vol. 60, pp. 243–251, 2015.

[10] K. Dragomiretskiy and D. Zosso, "Variational Mode Decomposition," IEEE Trans. Signal Process., vol. 62, pp. 531–544, 2014.

[11] J. Zhou, M. Xiao, and G. Ji, "Rolling Bearing Fault Diagnosis Based on WGWOA-VMD-SVM," Sensors, vol. 22, p. 6281, 2022.

[12] X. Zhang, Q. Miao, and L. Wang, "A parameter-adaptive VMD method based on the grasshopper optimization algorithm for analyzing vibration signals from rotating machinery," Mech. Syst. Signal Process., vol. 108, pp. 58–72, 2018.

[13] H. Wang, C. Yang, and A. K. Nandi, "Bearing fault diagnosis based on optimized variational mode decomposition and 1D convolutional neural networks," Meas. Sci. Technol., vol. 32, p. 104007, 2021.

[14] J. Ding, L. Huang, and Y. Wang, "Spectrum-guided variational mode decomposition and its application to mechanical fault diagnosis," Measurement, vol. 156, p. 107604, 2021.

[15] J. Wang et al., "Remaining useful life prediction based on adaptive shrinkage processing and temporal convolutional network," Sensors, vol. 22, p. 9088, 2022.

[16] J. Liao et al., "Multi-scale residual neural network with enhanced gated recurrent unit for fault diagnosis of rolling bearings," Meas. Sci. Technol., vol. 35, p. 055008, 2024.

[17] T. Saghi et al., "Bearing fault diagnosis based on multi-scale CNN and bidirectional GRU," Vibration, vol. 6, p. 2, 2023.

[18] K. Xu et al., "Multimodal fault diagnosis of rolling bearings based on GRU-ResNet-CBAM," Machines, vol. 13, p. 318, 2026.

[19] X. Li et al., "Rolling bearing fault diagnosis via temporal-graph convolutional fusion," Sensors, vol. 25, p. 3894, 2025.

[20] X. Deng et al., "A multi-level fusion framework for bearing fault diagnosis using multi-source information," Processes, vol. 13, p. 2657, 2025.

[21] S. Wang, C. Wang, and B. Luo, "Research on bearing fault diagnosis based on VMD-RCMWPE feature extraction and WOA-SVM-optimized multidataset fusion," Sensors, vol. 25, p. 5139, 2025.

[22] M. Zhao, M. Kang, and M. Pecht, "Multiple wavelet coefficients fusion in deep residual networks for fault diagnosis," IEEE Trans. Ind. Electron., vol. 66, pp. 4696–4706, 2019.

[23] X. Liu et al., "Lightweight gearbox fault diagnosis under high noise based on improved multi-scale depthwise separable convolution and efficient channel attention," Sensors, vol. 26, p. 1196, 2026.

[24] L. Zhang et al., "A bearing fault diagnosis method with implementation of multiscale time–frequency feature fusion and bidirectional information interaction," IEEE Sens. J., vol. 25, pp. 23740–23756, 2025.

[25] Y. Zou et al., "Rolling bearing fault diagnosis based on cross-attention fusion WDCNN and BILSTM," Comput. Mater. Contin., vol. 83, pp. 4699–4723, 2025.

[26] J. Guo et al., "MD-Transformer: a noise-oriented lightweight bearing fault diagnosis model," Meas. Sci. Technol., vol. 36, p. 116111, 2025.

[27] H. Li et al., "An optimized VMD method and its applications in bearing fault diagnosis," Measurement, vol. 166, p. 108185, 2020.

[28] Z. Zhang et al., "Application of a residual structure time convolutional network based on an attention mechanism in predicting the remaining useful life interval of bearings," Sensors, vol. 24, p. 4132, 2024.

[29] Y. Deng et al., "Switch open-circuit fault diagnosis of the Vienna rectifier using the Transformer–BiTCN network with improved snow geese algorithm optimization," Electronics, vol. 13, p. 3655, 2025.

[30] H. Bao et al., "A novel signal resolution enhancement method based on CBAM-ResNet for intelligent bearing fault diagnosis," Meas. Sci. Technol., vol. 35, no. 9, p. 096132, 2024.

[31] L. Guo et al., "Fault diagnosis method of rolling bearings based on a 1D multi-channel improved convolutional neural network in a noisy environment," Sensors, vol. 25, p. 2286, 2025.

[32] H. Peng, W. Desmet, and K. Gryllias, "RMA-CNN: A residual mixed-domain attention CNN for bearing fault diagnosis and its time-frequency domain interpretability," J. Dyn. Monit. Diagn., vol. 2, p. 156, 2023.

[33] W. Li et al., "Dual-channel parallel multimodal feature fusion for bearing fault diagnosis," Machines, vol. 13, p. 950, 2025.