Gearbox compound fault diagnosis using CEEMDAN feature extraction and a dual-attention multi-scale BiLSTM model

Lianxin Wu[1], Xiaojie Sun[2]
School of Intelligence Technology, Shanghai Institute of Technology, Shanghai, China
[1]Corresponding author
E-mail: [1]*906889823@qq.com*, [2]*sxjlm2003@163.com*

Received 29 January 2026; accepted 15 April 2026; published online 25 May 2026
DOI https://doi.org/10.21595/jve.2026.26068

Abstract. As a core component of mechanical transmission systems, the gearbox's operating state directly determines equipment reliability and industrial production safety. In actual working conditions, a single fault can easily evolve into a complex fault mode with multiple coupled faults. Traditional diagnostic methods face challenges such as insufficient feature extraction and low fault mode discrimination. To address this issue, an intelligent diagnostic model is proposed that integrates adaptive noise complete set empirical mode decomposition (CEEMDAN) feature extraction, multi-scale convolution, and a dual attention mechanism. First, CEEMDAN is used to decompose the vibration signal at multiple scales. After effective IMF filtering, time-domain, frequency-domain, fault-specific, and coupled interactive features are extracted to form a multi-dimensional feature set. Then, adaptive principal component analysis (PCA) is used to reduce the dimensionality to obtain a low-redundancy feature set. Subsequently, a diagnostic model containing multi-scale convolution, a bidirectional long short-term memory network (BiLSTM), and dual attention branches is constructed, and an improved loss function is combined to enhance the ability to distinguish complex fault features. Experimental results based on the Beijing Jiaotong University bogie gearbox bench dataset verify the effectiveness and robustness of the proposed method under complex fault modes, providing a reliable technical solution for gearbox fault diagnosis in industrial scenarios.

Keywords: gearbox, fault diagnosis, compound faults, attention mechanism, Feature fusion.

1. Introduction

The gearbox is the core component of the rail vehicle transmission system, and the movement and torque conversion are completed through the periodic meshing of the gear teeth, and its operating state directly affects the overall reliability of the equipment and the safety of industrial production [1]. In actual working conditions, gearboxes often face complex working conditions such as heavy loads, high speeds, and variable loads, and key components such as gears and bearings are prone to failures such as root cracks, tooth surface wear, broken teeth, and damage to the inner ring of the bearing [2]. More importantly, if a single fault is not diagnosed in time, it is very easy to induce the failure of multi-component association, forming a compound failure mode of multi-fault coupling. This type of compound fault leads to the presence of multi-source and multi-scale nonlinear features in the vibration signal, and the fault features are modulated and crossed with each other, which seriously increases the difficulty of fault diagnosis [3-4].

At present, gearbox fault diagnosis research mostly focuses on single fault detection [5], and traditional methods such as envelope analysis [6], wavelet transform [7], spectral kurtosis [8], and other time-frequency domain analysis techniques have achieved certain results in single fault diagnosis, but in multi-fault coupling scenarios, it is difficult to effectively separate superimposed fault features, resulting in a significant decrease in diagnostic accuracy. With the development of artificial intelligence technology, machine learning [9-10] and deep learning [11-12] methods have been gradually applied to the field of gearbox fault diagnosis. Zhou et al. proposed a

convolutional sparse code separation and diagnosis algorithm based on multi-scale convolutional kernel matching composite regularization, which has excellent fault separation ability under random noise and harmonic interference [13]. To improve the universality of diagnostic methods, a study proposed a general fault diagnosis framework for rotating machinery based on phase entropy. The phase entropy is used to effectively quantify the complexity and irregularity of the phase sequence of vibration signals. Combined with classifiers such as dual support vector machines, a general diagnostic process that does not rely on a large number of working condition labels and can adapt to different operating conditions is constructed [14]. To further integrate multi-source information and mine deep fault features, a multi-modal, multi-scale, and multi-level fusion quadrant entropy model was proposed [15]. A few-fault diagnosis model based on multi-scale perception and multi-level feature fusion image quadrant entropy (MPMFFIQE) can effectively and accurately diagnose mechanical faults in industrial applications with only a small number of training samples [16]. Li et al. proposed a capsule neural network with an improved feature extractor for intelligent identification of composite fault components [17]. Jia Shunyu et al. designed a multi-channel diagnostic method that combines random forests and evidence theory to achieve accurate identification of various single faults in compound faults [18]. Xie Fengyun proposed a diagnostic scheme based on multi-scale Weber dispersive entropy graph neural network, which effectively improved the state recognition effect of nonlinear and nonstationary signals [19]. However, the existing methods still have limitations: most rely on single decomposition or feature extraction techniques, which are difficult to fully capture the multi-scale interaction features of faults. The lack of targeted differentiation mechanism in the compound failure mode is prone to misdiagnosis [20].

To address the aforementioned challenges, this paper proposes an intelligent diagnostic model specifically designed for complex fault coupling scenarios. Its core innovations are reflected in three aspects:

First, a feature system dedicated to complex faults is constructed. Addressing the problem of overlapping and difficult-to-separate multi-source features in complex faults, this paper innovatively introduces fault-specific features and coupling interaction features into the traditional time-domain and frequency-domain feature spaces. Through multi-dimensional feature fusion, the model's ability to represent the inherent coupling relationships of complex faults is enhanced.

Second, a dual-attention branch mechanism is designed. Unlike traditional single-attention mechanisms, this paper proposes a dual-branch structure consisting of a general multi-head robust attention and a complex fault-specific attention. The former is responsible for the robust extraction of general fault features, while the latter focuses on the coupling features of complex faults, achieving a two-layer distinction of "general feature extraction + precise focusing of coupling features," filling the gap in existing methods for targeted modeling of fault coupling features.

Finally, the hybrid loss function is optimized. To solve the core problem of low discriminative power of complex fault patterns, this paper improves the multi-label cross-entropy loss and introduces a complex fault contrast loss and an attention entropy regularization term. By maximizing the feature differences between different composite faults through contrastive loss and enhancing feature focusing ability through regularization terms, the model can effectively learn the discriminative features of composite faults.

Based on a bench dataset, this paper systematically verifies the effectiveness of the proposed model in composite fault diagnosis tasks, and also validates the basic performance of the method through single fault diagnosis experiments. The research results can provide technical support for intelligent operation and maintenance of gearboxes in industrial scenarios.

2. Method principle

2.1. Basic principle of CEEMDAN

Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) is an

improved signal decomposition algorithm based on Ensemble Empirical Mode Decomposition (EEMD) and Complementary Ensemble Empirical Mode Decomposition (CEEMD). It addresses the shortcomings of traditional decomposition methods, such as mode mixing and error accumulation, by introducing independent additive Gaussian white noise in each decomposition round, dynamically offsetting the calculation errors of each Intrinsic Mode Function (IMF), and significantly improving mode orthogonality and decomposition stability [21]. The core idea is to decompose a non-stationary original signal into a series of physically meaningful Intrinsic Mode Functions (IMFs) and residual signals. The mathematical expression is:

$$x_{(t)} = \Sigma_{k=1}^{k} IMF_k(t) + r_k(t), \tag{1}$$

where $x(t)$ is the input signal; $IMF_k(t)$ is the k-th Intrinsic Mode Function; k is the decomposition order (10 in this paper); and $r_k(t)$ is the residual signal.

2.2. Adaptive PCA

Principal component analysis (PCA) is a classical linear dimensionality reduction method that maps high-dimensional features to low-dimensional space through orthogonal transformation to maximize data variance in the new space, thereby retaining key information and eliminating redundancy [22]. In order to adapt to the high-dimensional characteristics of compound fault characteristics, this paper adopts an adaptive PCA strategy, and the specific steps are as follows:

(1) Missing value imputation:

Since outliers or missing values may be generated during signal decomposition and statistical feature calculation, in order to ensure that the covariance matrix can be constructed correctly, the mean imputation method is first used to handle the missing values generated during feature extraction to ensure data integrity.

$$x_{ij}^{*} = \begin{cases} x_{ij}, & \text{if not missing,} \\ \bar{x}_j, & \text{if missing,} \end{cases} \tag{2}$$

where $\bar{x}_j$ is the mean of the j-th feature.

(2) Standardization processing:

Different features have different dimensions. If the covariance matrix is calculated directly, large-scale features will dominate the principal component direction. Therefore, Z-score standardization is performed on the padded features to eliminate the impact of dimensional differences on model training:

$$Z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}, \tag{3}$$

where $\bar{x}_j$ is the feature mean; σ_j is the feature standard deviation.

After standardization, the mean of each feature is 0 and the variance is 1, causing the covariance matrix to degenerate into a correlation coefficient matrix, thus ensuring that each feature contributes fairly to the principal component direction.

(3) Construction and eigenvalue decomposition of the covariance matrix:

Construct the covariance matrix for the standardized matrix Z:

$$C = \frac{1}{n-1} Z^T Z. \tag{4}$$

Then, eigenvalue decomposition is performed:

Where $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_m)$, $U = [\mu_1, \mu_2, \ldots, \mu_m]$.

Sort by eigenvalue from largest to smallest:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m. \tag{5}$$

(4) Adaptive Principal Component Selection.

The cumulative variance contribution rate is defined as:

$$R_k = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{m} \lambda_i}. \tag{6}$$

Traditional PCA requires manual setting of the number of principal components k, while this paper adopts an adaptive strategy: setting the cumulative contribution rate threshold to 98 % and automatically selecting the minimum k value that satisfies R_k:

$$R_k \geq 0.98. \tag{7}$$

This strategy effectively removes low-contribution noise components while preserving the dominant frequency band energy, modulation structure, and interaction features in complex faults to the greatest extent possible. This reduces the input dimensionality and parameter size of subsequent deep networks, decreases the risk of overfitting, and improves training stability.

The final dimensionality-reduced data is as follows:

$$Y = ZW_k, \tag{8}$$

where $W_k = [\mu_1, \ldots, \mu_k]$.

2.3. Convolutional neural networks

Convolutional Neural Networks (CNNs) are deep learning models inspired by biological visual systems. They are based on multi-layer convolution and pooling operations, possessing parameter sharing, sparse connectivity, and strong inductive bias capabilities. They exhibit excellent performance in feature extraction and generalization, while having a more streamlined parameter set. This study employs a multi-scale one-dimensional convolutional structure, using three different sized convolutional kernels: a 7×1 kernel to capture low-frequency modulation and global structural information, a 5×1 kernel to enhance the expression of meshing frequencies and their harmonic features, and a 3×1 kernel to focus on characterizing high-frequency impacts and transient pulses. This progressive design of convolutional kernels, from large to small, enables the model to extract features layer by layer from macroscopic modulation structures to microscopic impact details, effectively alleviating the multi-band coupling problem in complex faults and improving feature decoupling and discrimination capabilities.

The basic structure alternately stacks the convolutional layer, the pooled layer and the fully connected layer: the convolutional layer slides in the local receptive field through the convolutional kernel, extracts the input features and introduces nonlinearity through the activation function; The pooling layer downsamples the feature map, which retains the salient features while reducing the dimension and enhancing the robustness of the model. The fully connected layer maps the extracted high-order features to the probabilities of each category of the sample to complete the classification task.

2.4. Bidirectional long short-term memory network (BiLSTM)

Long short-term memory networks (LSTMs) effectively solve the gradient disappearance problem of traditional recurrent neural networks (RNNs) through the gating mechanism of input gates, forgetting gates, and output gates, but they can only use historical time series information

and cannot capture future data dependencies. Bidirectional long short-term memory network (BiLSTM) realizes the synchronous extraction of bidirectional information from sequence data by constructing forward and reverse LSTM layers in parallel, and its core formula is as follows:

$$\overrightarrow{h_t} = f\left(\vec{w}_x \times X_t + \overrightarrow{h_{t-1}} \times \vec{w}_h + \overrightarrow{b_n}\right), \tag{9}$$

$$\overleftarrow{h_t} = f\left(\overleftarrow{w_x} \times X_t + \overleftarrow{h_{t-1}} \times \overleftarrow{w_h} + \overleftarrow{b_h}\right), \tag{10}$$

$$Y_t = f\left(\overrightarrow{w_y} \times \overrightarrow{h_t} + \overleftarrow{w_y} \times \overleftarrow{h_t} + b_y\right), \tag{11}$$

where X_t is the input of the time t; w_x and w_h are the weight matrix input to the hidden state and the weight matrix from the previous moment to the current hidden state, respectively. b_n and b_y are the currently hidden bias terms and the output bias terms, respectively.

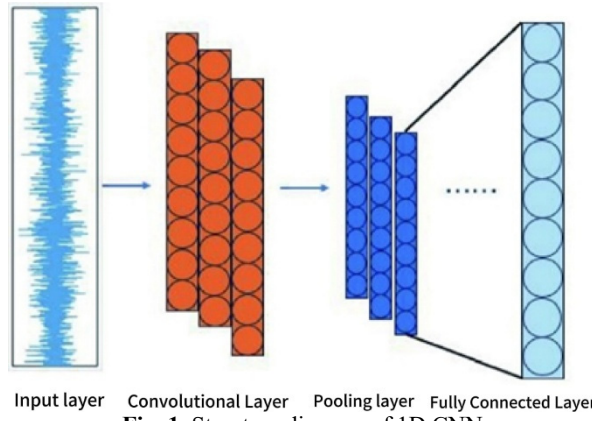

Fig. 1. Structure diagram of 1D CNN

2.5. Design of dual attention mechanism

To achieve precise focusing of composite fault features and improve the model's ability to identify complex coupled faults, this paper designs a dual-attention mechanism. This structure consists of two branches: General Multi-head Robust Attention and Compound Fault-specific Attention, achieving feature weighting and selection from two levels: general feature modeling and fault-specific feature enhancement.

Let the input feature sequence be:

$$X = [x_1, x_2, \ldots, x_n] \in R^{n \times d}, \tag{12}$$

where n is the feature length and d is the feature dimension.

(1) General Multi-head Robust Attention.

To improve the model's ability to represent fault features at different scales, this paper introduces a multi-head attention mechanism. Multiple attention heads can learn feature weights from different subspaces, thereby enhancing the model's robustness to complex signal features.

The weights of the i-th attention head are calculated as follows:

$$a_i = \mathrm{Softmax}\left(W_{2,i} \cdot \tanh\left(W_{1,i}X + b_{1,i}\right) + b_{2,i}\right), \tag{13}$$

where: $W_{1,i}$, $W_{2,i}$ are the weight matrices of the i-th attention head; $b_{1,i}$, $b_{2,i}$ are bias terms; and $\tanh(\cdot)$ is a non-linear activation function.

The weighted features obtained for each attention head are:

$$F_i = a_i \odot X, \tag{14}$$

where $\odot$ represents the element-wise weighting operation.

To reduce the bias influence that a single attention head may bring, this paper adopts a mean fusion strategy to fuse the four attention heads:

$$F_{gen} = \frac{1}{H}\sum_{i=1}^{H} F_i, \tag{15}$$

where $H = 4$ represents the number of attention heads.

This module effectively learns the global dependencies of common fault features, improving the model's stable recognition capability for multiple types of fault features.

(2) Composite Fault-Specific Attention.

Since composite faults often contain coupled features of multiple single faults, relying solely on general attention is insufficient for precise differentiation. Therefore, this paper designs a composite fault-specific attention branch to enhance features for different fault types.

First, the G_x fault attention branch and the G5 fault attention branch are constructed respectively:

$$a_{gx} = \mathrm{Softmax}\left(W_{gx}X + b_{gx}\right), \tag{16}$$
$$a_{g5} = \mathrm{Softmax}\left(W_{g5}X + b_{g5}\right). \tag{17}$$

The corresponding weighted features are:

$$F_{gx} = a_{gx} \odot X, \tag{18}$$
$$F_{g5} = a_{g5} \odot X, \tag{19}$$

where a_{gx} represents the G_x fault feature weight; a_{g5} represents the G5 fault feature weight.

To further characterize the coupling relationship between complex faults, this paper constructs an interactive attention branch:

$$a_{int} = \mathrm{Softmax}\left(W_{int}\left[F_{gx}, F_{g5}\right] + b_{int}\right), \tag{20}$$

where $[F_{gx}, F_{g5}]$ represent feature concatenation operations.

The final complex fault feature is represented as:

$$F_{spec} = \alpha F_{gx} + \beta F_{g5} + \gamma a_{int} \odot X, \tag{21}$$

where:

$$\alpha + \beta + \gamma = 1, \tag{22}$$

is used to control the contribution ratio of the three types of attention weights.

(3) Dual Attention Fusion Output.

The final model fuses general attention features with fault-specific attention features:

$$F_{out} = F_{gen} + F_{spec}. \tag{23}$$

This structure can simultaneously capture general fault pattern features, specific fault features, and the coupling relationship of complex faults, thereby significantly improving the model's ability to identify complex faults.

2.6. Improve the mixed loss function

To enhance the model's ability to distinguish between composite faults, this paper designs an improved hybrid loss function, consisting of three parts: multi-label binary classification cross-entropy loss, composite fault contrast loss, and attention entropy regularization term.

Different composite categories (such as G1+G5 and G2+G5) are prone to feature aliasing when sharing G5 components. To strengthen the discriminative margin between different composite categories, a boundary constraint based on class response differences is introduced.

Suppose there are two types of composite faults A and B, with key discriminative dimensions c_A and c_B, respectively. Define the average logits of the two classes of samples in their corresponding dimensions:

$$\bar{z}_A = \frac{1}{|S_A|}\sum_{i\in S_A} z_{i,c_A}, \quad \bar{z}_B = \frac{1}{|S_B|}\sum_{i\in S_B} z_{i,c_B}, \tag{24}$$

where S_A, S_B represent the sample sets of the two composite fault classes, respectively.

The inter-class response difference is defined as:

$$\Delta_{AB} = |\bar{z}_A - \bar{z}_B|. \tag{25}$$

To ensure a minimum separable margin γ (set to 0.6 in this paper) between different composite classes, a hinge-like contrast loss is constructed:

$$L_{contrast}^{AB} = \max(0, \gamma - \Delta_{AB}). \tag{26}$$

A penalty is applied when the inter-class difference is less than the margin threshold; otherwise, the loss is 0.

The summation is applied pairwise for all composite classes:

$$L_{contrast} = \sum_{(A,B)\in\mathcal{P}} \max\,(0, \gamma - |\,\bar{z}_A - \bar{z}_B|), \tag{27}$$

where $\mathcal{P}$ represents the set of composite fault class pairs.

Theoretically, this loss is equivalent to constructing a margin-based separation constraint in the output space, ensuring that different composite classes maintain a minimum margin in the discriminative subspace, thereby enhancing inter-class separability.

Let the output weights of the general multi-head robust attention be:

$$\alpha_i = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{iT}), \tag{28}$$

satisfying:

$$\sum_{t=1}^{T} \alpha_{it} = 1, \tag{29}$$

its information entropy is defined as:

$$H(\alpha_i) = -\sum_{t=1}^{T} \alpha_{it} \log(\alpha_{it} + \varepsilon), \tag{30}$$

where ε is the numerically stable term.

Averaging the batch of samples:

$$L_{attn1} = \frac{1}{N}\sum_{i=1}^{N} H(\alpha_i) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} \alpha_{it} \log(\alpha_{it} + \varepsilon). \tag{31}$$

The smaller the entropy, the more concentrated the attention distribution, meaning the model is more focused on key feature regions.

Let the weights of the composite fault-specific attention be:

$$\beta_i = (\beta_{i1}, \beta_{i2}, \ldots, \beta_{iT}). \tag{32}$$

Similarly, define its entropy as:

$$L_{attn2} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} \beta_{it} \log(\beta_{it} + \varepsilon). \tag{33}$$

Since the specific attention focuses more on the interaction region between G_x and G5, its entropy constraint helps to suppress weight dispersion and improve feature decoupling ability.

Complete hybrid loss function:

$$L_{total} = L_{BCE} + \alpha \cdot L_{contrast} - \beta \cdot (L_{attn1} + L_{attn2}), \tag{34}$$

where L_{BCE} is the cross-entropy loss of multi-label binary classification, which is used for basic fault classification tasks; The $L_{contrast}$ is the comparative loss of compound faults, with a weight α of 0.15, which improves the discrimination by maximizing the feature differences of different compound faults. L_{attn1} and L_{attn2} are the entropy regular terms of general attention and compound fault attention, respectively, with a weight β of 0.02, which enhances the feature focusing ability by minimizing attention entropy.

3. Experimental design

3.1. Experimental data and preprocessing

To verify the effectiveness of the proposed method, the Beijing Jiaotong University BJTU-RAO bogie transmission system fault simulation test bench dataset was used for experimental validation [23]. This dataset was obtained through a fault simulation experiment on the subway train bogie transmission system and contains multi-sensor data for 51 health states. In this experiment, gearbox fault data were selected, covering six basic states: G0 (normal state), G1 (tooth root crack), G2 (tooth surface wear), G3 (tooth missing), G4 (tooth breakage), and G5 (bearing inner ring fault), as well as four compound fault states: G1 G5, G2 G5, G3 G5, and G4 G5. The data acquisition parameters were: sampling frequency 64 kHz, motor speed 20 Hz, and load 0 kN.

In order to balance feature integrity and computational efficiency, the sliding window technology is used to segment the original vibration signal: the window length is set to 2048 points, covering multiple gear meshing cycles; The step size is set to 1024 points, with an overlap rate of 50 %. For small sample problems, 3x data augmentation is performed on each raw sample, with specific strategies including:

1) Amplitude scaling: randomly scaling the signal amplitude to 0.85-1.15 times the original amplitude;

2) Gaussian noise addition: Add Gaussian white noise with a standard deviation of 1 % of the

standard deviation of the original signal to enhance the model's anti-interference ability.

The hierarchical sampling strategy is used to ensure the consistent distribution of various fault samples, and the dataset is divided into training set, verification set and test set according to a ratio of 7:1:2.

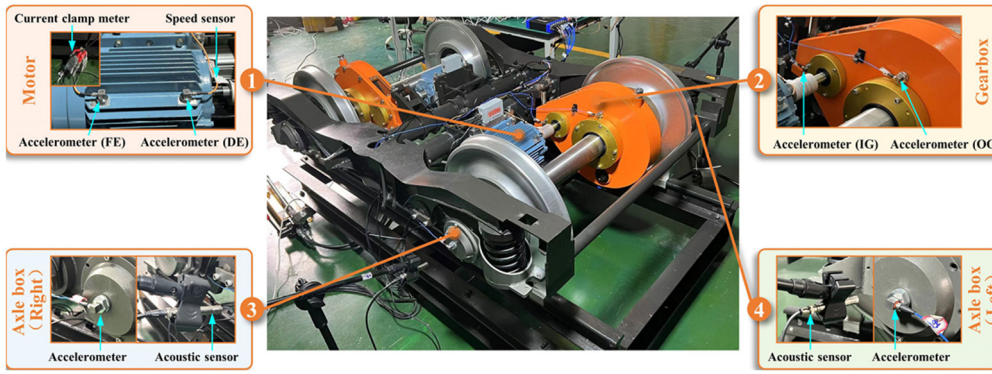

Fig. 2. Metro train bogie drive system fault simulation test bench

3.2. Feature extraction process

In order to comprehensively capture the multi-dimensional features of compound faults, a four-level feature extraction system is constructed, and the specific steps are as follows:

1) CEEMDAN Decomposition and Effective IMF Selection: This paper determines the optimal number of decomposition layers through sensitivity experiments. A comparison of 6-12 layer decompositions reveals that a 10-layer decomposition yields the best overall performance. Therefore, this paper performs a 10-layer CEEMDAN decomposition on the preprocessed vibration signal, obtaining 10 IMF components. Based on the matching criteria of energy proportion, correlation coefficient, and fault frequency, two effective IMF components containing key fault characteristic information are selected from the 10 IMFs to eliminate noise and spurious components, thereby improving feature effectiveness.

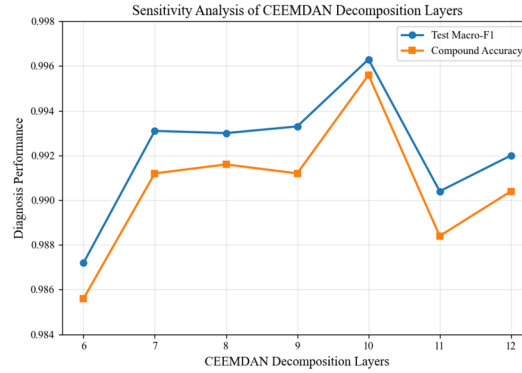

Fig. 3. Results of the layer number sensitivity experiment

2) Basic feature extraction: For the two selected effective IMF components, 14-dimensional time-domain features (mean, standard deviation, root mean square, skewness, kurtosis, peak factor, etc.) and 14-dimensional frequency-domain features (spectral energy, spectral centroid, spectral entropy, the first three peak frequencies and amplitude, etc.) are extracted for each component, resulting in a total of 56 basic features;

3) Fault-specific feature extraction: For composite fault modes, 7-dimensional fault-specific features are extracted from the original signal and key IMF, including 4-dimensional features

specific to G1-G4 (spectral peak width coefficient, harmonic distribution entropy, impulse factor, pulse duration) and 3-dimensional features specific to G5 (fault frequency energy ratio, sideband energy ratio, impulse response spectrum peak value).

4) Coupling interaction feature extraction: extract the two-dimensional composite fault interaction features (G_x and G5 frequency-energy ratio, composite fault entropy) to describe the nonlinear characteristics of multi-fault coupling;

5) Global feature extraction: 10-dimensional global features are directly extracted from the original signal, including global time-domain features (mean, root mean square, peak factor, etc., 5 dimensions) and global frequency-domain features (meshing frequency energy, fault frequency ratio, etc., 5 dimensions), to supplement the overall information that may be lost during the decomposition process.

6) Feature Fusion and Adaptive PCA Dimensionality Reduction: The extracted 56-dimensional basic features, 7-dimensional specific features, 2-dimensional interaction features, and 10-dimensional global features are fused to form an original feature set with a total dimension of 75. Finally, adaptive PCA dimensionality reduction is performed on this 75-dimensional feature set, setting the cumulative variance contribution rate threshold to 98 % to automatically determine the number of principal components, resulting in a final low-redundancy feature set for subsequent model training.

Table 1. Compound fault model network structure parameters

Module name	Submodule / layer type	Core parameters
Input layer	Data reshaping	Input dimensions: Features after PCA dimensionality reduction → Reshape to (N, n_channels, seq_len)
Convolution module	First convolution layer	Conv1d: input channels = n_channels, output channels = 64, kernel size = 7, padding = 3 BatchNorm1d(64); ReLU; MaxPool1d(2)
	Second convolutional layer	Conv1d: input channels = 64, output channels = 128, kernel size = 5, padding = 2 BatchNorm1d(128);ReLU; MaxPool1d(2)
	Third convolutional layer	Conv1d: input channels = 128, output channels = 256, kernel size = 3, padding = 1 BatchNorm1d(256);ReLU; MaxPool1d(2)
SE channel attention module	SEBlock	Number of channels = 256, compression ratio = 16 AdaptiveAvgPool1d (1)+Two fully connected layers+Sigmoid
Bidirectional LSTM module	LSTM	Input dimension = 256, hidden layer dimension = 128, number of layers = 2, bidirectional = True, dropout = 0.2, batch_first = True
General multi-head attention	Multi-head robust attention	Hidden dimension = 256, projection dimension = 64, number of heads = 4; each layer: Linear (256→64) → Tanh → Linear (64→1)
Composite fault-specific attention	Compound fault attention	Hidden dimension = 256, projection dimension = 64 Gx/G5 branch: Linear (256→64) → Tanh → Linear (64→1) Interaction branch: Linear (512→64) → Tanh → Linear (64→1)
Classifier	First fully connected layer	Linear(512→256); ReLU; Dropout(0.25)
	Second fully connected layer	Linear(256→192); ReLU; Dropout(0.25)
	Output layer	Linear (192→num_outputs)

3.3. Diagnostic model construction and overall architecture

The overall architecture of the intelligent diagnostic model proposed in this paper is shown in Fig. 4. Its data processing flow is as follows: First, the original vibration signal is decomposed by CEEMDAN, subjected to multi-dimensional feature extraction, and reduced by adaptive PCA to

obtain a low-dimensional feature vector. This vector is then reshaped and input into a three-layer multi-scale one-dimensional convolutional module (with kernel sizes of 7×1, 5×1, and 3×1) to extract deep fault features across multiple frequency bands. Subsequently, an SE channel attention module is introduced to adaptively recalibrate the feature channels output by the convolutional layers, strengthening key feature channels. Next, the feature sequence is input into a two-layer BiLSTM network to capture its temporal dependencies. The output of the BiLSTM is fed in parallel into two attention branches: a general multi-head robust attention branch and a composite fault-specific attention branch. The output features of the two branches are concatenated to form the final fault feature representation. Finally, this feature representation is fed into a classifier consisting of three fully connected layers to output multi-label diagnostic results.

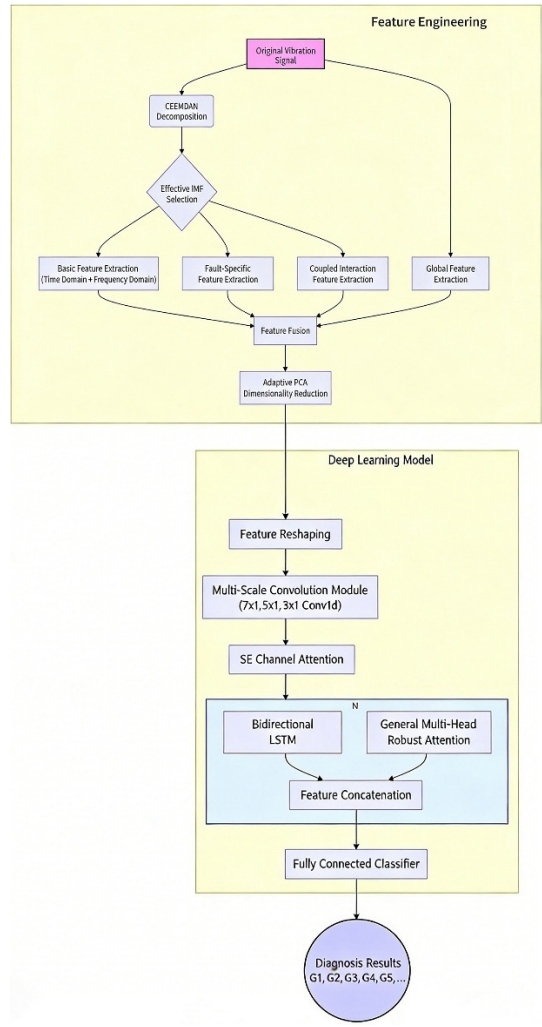

Fig. 4. Overall model architecture diagram

3.4. Model training and weight analysis

Model training uses the AdamW optimizer with a learning rate of 3e-4 and a weight decay of 5e-5. Training is conducted for 100 epochs with an early stopping strategy (patience = 15) to prevent overfitting, and mixed precision training is employed to improve computational efficiency.

To determine the optimal values of the composite fault loss weight α and the attention entropy regularization term weight β, this paper conducts grid search experiments on two sets of hyperparameters under fixed network structure and training parameters. The parameter ranges are set as follows: $\alpha \in [0.05, 0.25]$, $\beta \in [0.01, 0.05]$, where α controls the weight of the composite fault loss in the total loss function to enhance the model's learning ability for composite fault samples; β constrains the entropy value of the attention distribution to prevent excessive concentration of attention weights, thereby improving the model's generalization ability.

The Macro-F1 scores, composite fault recognition accuracy, and optimal validation set F1 results of the model under different parameter combinations are shown in the Figs. 5-6.

Experimental results show that when $\alpha = 0.15$ and $\beta = 0.02$, the model achieves a Macro-F1 score of 0.9986 on the test set, while maintaining a composite fault identification accuracy of 1.0 on the validation set. Compared to other parameter combinations, this setting achieves the best balance between overall diagnostic performance and composite fault identification capability.

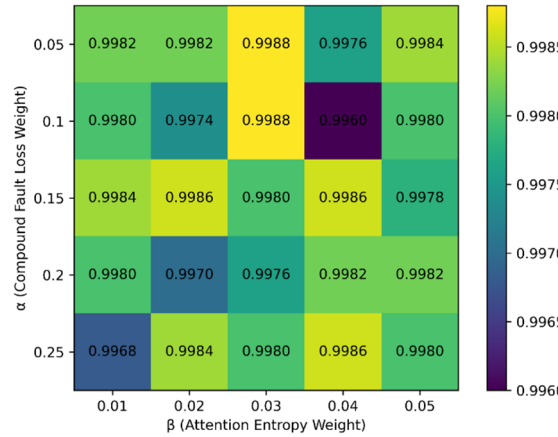

Fig. 5. Heatmap of macro-F1 under different hyperparameter combinations of α and β

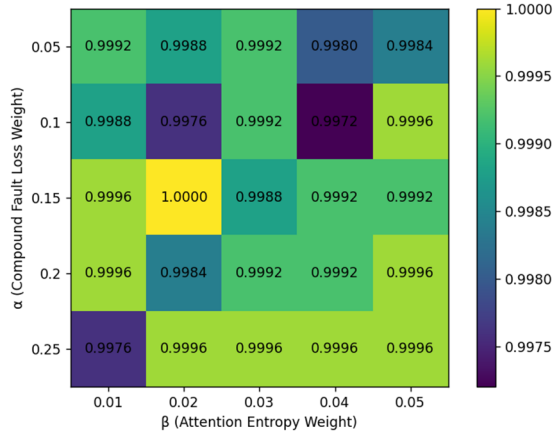

Fig. 6. Heatmap of composite fault accuracy under different hyperparameter combinations of α and β

4. Results and discussion

4.1. Single fault diagnosis results

To verify the model's basic performance, experiments were first conducted on a single fault dataset. Experimental results show that the model's training and validation F1 curves converged

rapidly and synchronously without significant overfitting, and the loss curve decreased smoothly. The confusion matrix shows that the diagnostic accuracy for various single faults is at a high level. Overall, the model demonstrates excellent accuracy and robustness in gearbox single fault diagnosis tasks, achieving a Macro-F1 score of 0.9868 and a perfect match rate of 0.9860, which can meet the practical needs of intelligent diagnosis of single gearbox faults in industrial scenarios.

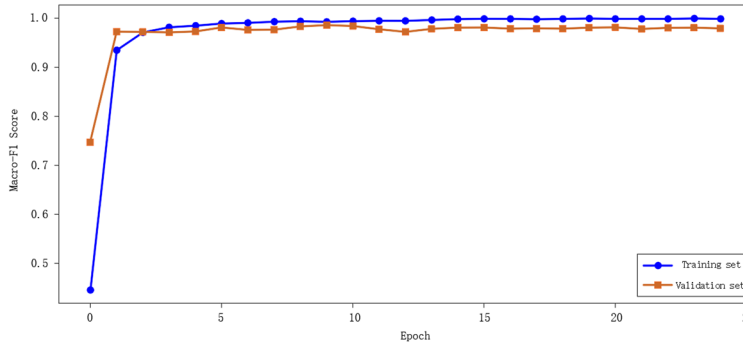

Fig. 7. Gearbox single fault F1 curve

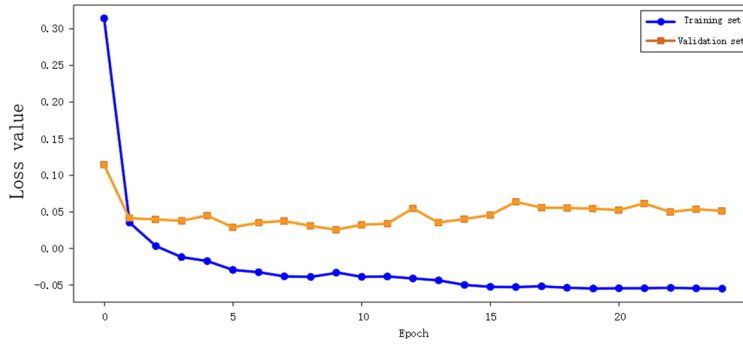

Fig. 8. Loss curve of single gearbox failure

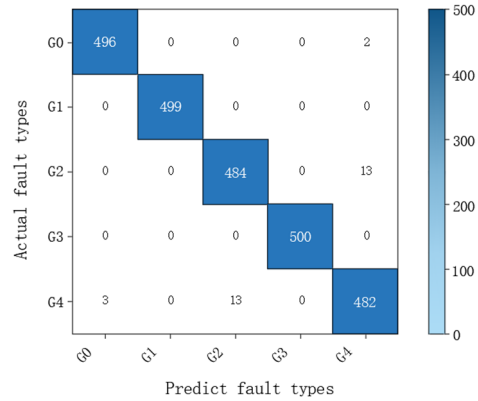

Fig. 9. Confusion matrix for single gearbox fault

4.2. Compound fault diagnosis results

The proposed model was validated on a dataset containing four types of composite faults (G1+G5, G2+G5, G3+G5, G4+G5).

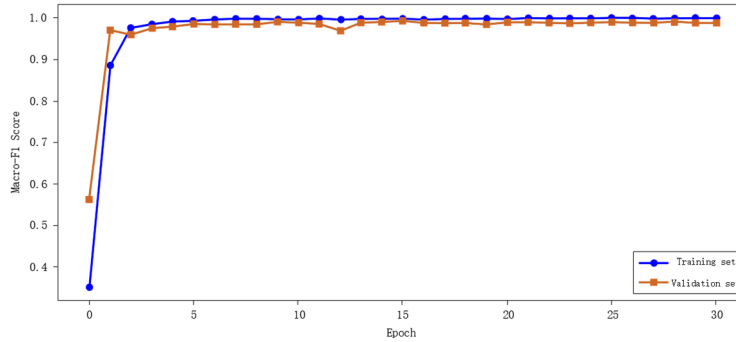

Fig. 10. Gearbox compound fault F1 curve

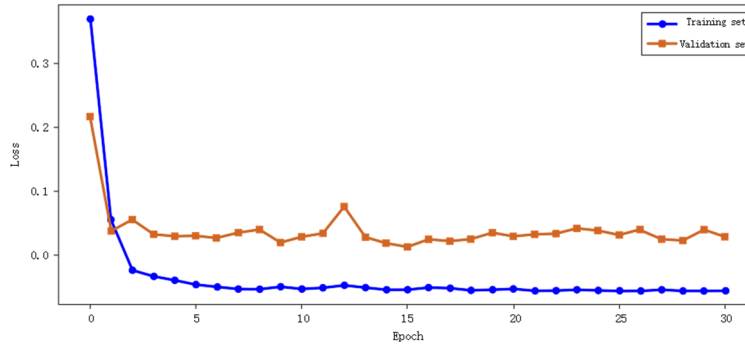

Fig. 11. Gearbox compound fault loss curve

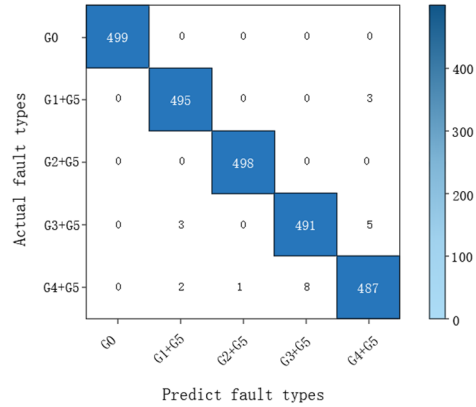

Fig. 12. Gearbox compound fault confusion matrix

From the F1 curve, it can be seen that the Macro-F1 score of the training set quickly converges to more than 0.99 within 5 Epoch, and the F1 synchronization of the verification set is stable around 0.98. In the loss curve, the training set loss continues to decrease with the Epoch iteration and approaches –0,05, and the validation set loss is stable at about 0.05, reflecting the convergence and stability of the training process. Combined with the confusion matrix analysis, the accuracy of various composite faults is at a high level, which verifies the effectiveness of the "exclusive attention mechanism of compound fault interaction features", and the model can effectively distinguish the multi-fault superposition mode by fusing single fault and coupling features. In summary, the model shows excellent accuracy and robustness in the composite fault diagnosis task, with a Macro-F1 score of 0.9918 and a perfect matching rate of 0.9896, which can meet the intelligent diagnosis needs of gearbox multi-fault coupling state in industrial scenarios.

ISSN PRINT 1392-8716, ISSN ONLINE 2538-8460

4.3. Ablation experiment

To verify the independent contribution of each core module in the proposed method to the composite fault diagnosis performance, this section designs ablation experiments. Using a controlled variable method, the CEEMDAN decomposition module, the dual attention branch, and the improved hybrid loss function are removed sequentially. The diagnostic performance of each variant model is compared under the same dataset and training strategy, thereby quantifying the actual role of each module.

The baseline model removes all core modules, retaining only the basic 1D-CNN and BiLSTM structures, and uses the standard BCE loss function to directly extract features from the original signal. Variant models: with ceemdan adds only the CEEMDAN multi-scale decomposition module to the baseline model. With attention adds only the dual attention branch to the baseline model. With loss uses only the improved hybrid loss function to the baseline model. The full model is the complete model proposed in this paper, containing all core modules.

The key performance indicators of each model are shown in the Table 2.

Table 2. Impact of each module on the model

Configuration name	Test macro-F1	Perfect match rate	Composite fault accuracy
Baseline	0.9823	0.9805	0.9811
With ceemdan	0.9865	0.9850	0.9865
With attention	0.9885	0.9875	0.9885
With loss	0.9905	0.9892	0.9905
Full model	0.9918	0.9896	0.9928

Ablation experiments show that each core module proposed in this paper contributes positively to the model performance. The complete model achieves optimal results on all metrics, fully demonstrating the effectiveness and synergistic effect of the module combination.

4.4. Comparative experiments

To verify the effectiveness of the proposed composite fault attention model, SVM, basic CNN, basic LSTM, and several advanced diagnostic methods (VMD-CNN, RSSD-CYCBD, ResNet-RFA) were selected as comparison models, and comparative experiments were conducted on the same dataset. All deep learning models used the same optimizer parameters, number of training epochs, and early stopping strategy; the SVM employed the RBF kernel function and a OneVsRest multi-label strategy. The experimental results show that the compound fault accuracy of the model in this paper is superior to all the compared deep learning models, demonstrating stronger feature extraction and fault pattern differentiation capabilities. Although the Macro-F1 value of SVM is slightly higher than that of the model in this paper, this is because SVM, as a traditional machine learning model, has stronger generalization ability in small-sample, high-dimensional data scenarios. The preprocessed experimental data is more linearly separable, and its simple structure can better balance the precision and recall of various fault types. The proposed model, being more complex in structure, may have slight overfitting for some small-sample fault categories during training, resulting in a slightly lower Macro-F1 value. However, SVM performs worse than the proposed model in composite fault accuracy and relies on manual feature engineering, making it difficult to capture the deep correlations among multiple faults in composite faults and hard to adapt to practical scenarios. The proposed model can automatically learn multi-scale features of composite faults through a deep network, making it more suitable for practical fault diagnosis needs. Therefore, the model in this paper was chosen over the SVM model.

4.5. Variable operating conditions experiment results

To verify the robustness of the model, variable operating conditions experiments were

conducted to test the model performance under different motor speeds and lateral loads. The results are shown in Table 4. When the motor speed is 20 Hz and the load is 0 kN, the model performs optimally, with a Macro-F1 score of 0.9944. As the speed increases and the load grows, the model performance slightly decreases, but the Macro-F1 score remains above 0.97, indicating that the model has good adaptability under complex operating conditions.

Table 3. Performance comparison of different models

Model	Macro-F1	Compound Fault Accuracy
SVM	0.9960	0.9908
CNN	0.9850	0.9836
LSTM	0.9786	0.9736
VMD-CNN	0.9912	0.9872
RSSD-CYCBD	0.9890	0.9824
ResNet-RFA	0.9848	0.9799
This article's model	0.9944	0.9928

Table 4. Model performance under different operating conditions

Operating conditions	Motor speed / lateral load	Macro-F1	Compound fault accuracy
1	20 Hz/0 kN	0.9944	0.9928
2	40 Hz/0 kN	0.9827	0.9796
3	60 Hz/0 kN	0.9744	0.9700
4	20 Hz/+10 kN	0.9787	0.9744
5	40 Hz/+10 kN	0.9771	0.9728
6	60 Hz/+10 kN	0.9740	0.9683
7	20 Hz/-10 kN	0.9821	0.9792
8	40 Hz/-10 kN	0.9711	0.9671
9	60 Hz/-10 kN	0.9785	0.9744

5. Conclusions

Aiming at the problems of insufficient feature extraction and low discrimination of fault mode in gearbox composite fault diagnosis, a deep learning model combining CEEMDAN feature extraction and dual attention mechanism is proposed, and the effectiveness and robustness of the method are verified by experiments, and the main conclusions are as follows:

1) Based on CEEMDAN's multi-scale decomposition and effective IMF screening, the original feature set constructed by combining time-domain, frequency-domain, fault-specific, and coupled interactive features can comprehensively cover the core information of both single and complex faults. Adaptive PCA dimensionality reduction effectively suppresses redundancy, laying a solid foundation for efficient model training.

2) A hybrid architecture of multi-scale convolution + BiLSTM + dual attention enables multi-scale extraction of fault features, temporal dependency modeling, and precise focusing. In particular, the dual attention mechanism ensures robustness of feature extraction and distinguishability of complex faults, significantly improving the recognition accuracy of complex faults.

3) An improved hybrid loss function (multi-label binary classification cross-entropy + complex fault contrast loss + attention entropy regularization term) effectively enhances the distinguishability of fault features. Combined with data augmentation and early stopping strategies, it further improves the model's generalization ability and training stability.

4) The proposed method has been verified by bench experiments, and the project landing path is clear, which can be directly adapted to the vibration signal acquisition system with 64 kHz sampling frequency, providing a high-precision and high-robustness intelligent diagnosis solution for the predictive maintenance of gearboxes in industrial scenarios.

Future research can further expand the direction: first, the introduction of transfer learning technology to improve the adaptability of the model under small samples and variable working

conditions; second, optimize the lightweight design of feature extraction and model architecture to meet the deployment needs of embedded devices; The third is to combine multi-sensor data fusion (such as vibration, temperature, and acoustic signals) to further improve the diagnostic reliability under complex working conditions.

Acknowledgements

This paper is supported by the Shanghai Science and Technology Program Project (21210750300), the Shanghai Municipal Education Commission's "Special Project for the Plan to Promote Scientific Research Paradigm Reform and Empower Discipline Advancement through Artificial Intelligence" (AIZX-3), and the Jiangsu Provincial Market Supervision Administration's Science and Technology Program Project (KJ2026017).

Data availability

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

Author contributions

Lianxin Wu proposed the conceptual methodology, created the model, wrote the paper. Xiaojie Sun reviewed and edited, revised the paper.

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] W. Wang, "Research on composite fault diagnosis method of rotating machinery based on multi-task hybrid expert," (in Chinese), *Yanshan University*, 2025, https://doi.org/10.27440/d.cnki.gysdu.2025.001048

[2] L. Lin, "Research on general methods for prediction and fault diagnosis of rotating machinery," (in Chinese), *Dalian University of Technology*, 2021, https://doi.org/10.26991/d.cnki.gdllu.2021.000359

[3] X. Zhao et al., "A gearbox compound fault recognition method based on triplet loss," (in Chinese), *Vibration and Shock*, Vol. 40, No. 5, pp. 46–54, 2021, https://doi.org/10.13465/j.cnki.jvs.2021.05.007

[4] W. Teng, X. Ding, H. Cheng, C. Han, Y. Liu, and H. Mu, "Compound faults diagnosis and analysis for a wind turbine gearbox via a novel vibration model and empirical wavelet transform," *Renewable Energy*, Vol. 136, pp. 393–402, Jun. 2019, https://doi.org/10.1016/j.renene.2018.12.094

[5] C. Peeters, J. Antoni, and J. Helsen, "Blind filters based on envelope spectrum sparsity indicators for bearing and gear vibration-based condition monitoring," *Mechanical Systems and Signal Processing*, Vol. 138, p. 106556, Apr. 2020, https://doi.org/10.1016/j.ymssp.2019.106556

[6] F. Cheng, L. Qu, W. Qiao, C. Wei, and L. Hao, "Fault diagnosis of wind turbine gearboxes based on dfig stator current envelope analysis," *IEEE Transactions on Sustainable Energy*, Vol. 10, No. 3, pp. 1044–1053, Jul. 2019, https://doi.org/10.1109/tste.2018.2859764

[7] M. Wen, H. Wang, and G. Zhu, "Application of wavelet transform and deep residual shrinkage network in gearbox fault diagnosis," (in Chinese), *Mechanical Science and Technology*, Vol. 43, No. 5, pp. 790–797, 2024, https://doi.org/10.13433/j.cnki.1003-8728.20230054

[8] J. Zhou, "Research on fault diagnosis method of wind power gearbox based on improved spectral kurtosis and CNN," (in Chinese), *Wuhan University*, 2022, https://doi.org/10.27379/d.cnki.gwhdu.2022.003427

[9] Q. Wang, "Research on gearbox fault diagnosis technology based on machine learning," (in Chinese), *Qingdao University*, 2023, https://doi.org/10.27262/d.cnki.gqdau.2023.002845

[10] K. Horváth and A. Zelei, "Gearbox fault diagnosis using industrial machine learning techniques," *Engineering Proceedings*, Vol. 79, No. 1, p. 36, 2024, https://doi.org/10.3390/engproc2024079036

[11] H. Ahmad et al., "Deep learning-based fault diagnosis of planetary gearbox: A systematic review," *Journal of Manufacturing Systems*, Vol. 77, pp. 730–745, Dec. 2024, https://doi.org/10.1016/j.jmsy.2024.10.004

[12] C. Chen et al., "Research on planet gearbox fault diagnosis algorithm combining deep learning and transfer learning," *Machine Tools and Hydraulics*, Vol. 53, No. 10, pp. 40–49, 2025.

[13] H. Zhou et al., "Research on multi-scale composite sparse compound fault diagnosis of gearboxes," (in Chinese), *Vibration, Measurement and Diagnosis*, Vol. 43, No. 2, pp. 215–222, 2023, https://doi.org/10.16450/j.cnki.issn.1004-6801.2023.02.002

[14] Z. Wang et al., "A generalized fault diagnosis framework for rotating machinery based on phase entropy," *Reliability Engineering and System Safety*, Vol. 256, p. 110745, 2025, https://doi.org/10.1016/j.ress.2024.110745

[15] Z. Wang et al., "Multi-modal multi-scale multi-level fusion quadrant entropy for mechanical fault diagnosis," *Expert Systems with Applications*, Vol. 281, p. 127715, Jul. 2025, https://doi.org/10.1016/j.eswa.2025.127715

[16] Z. Wang et al., "Few-shot fault diagnosis for machinery using multi-scale perception multi-level feature fusion image quadrant entropy," *Advanced Engineering Informatics*, Vol. 63, p. 102972, Jan. 2025, https://doi.org/10.1016/j.aei.2024.102972

[17] G. Li, L. He, Y. Ren, X. Li, J. Zhang, and R. Liu, "Compound fault diagnosis of planetary gearbox based on improved LTSS-BoW model and capsule network," *Sensors*, Vol. 24, No. 3, p. 940, 2024, https://doi.org/10.3390/s24030940

[18] S. Jia et al., "Multi-channel gearbox compound fault diagnosis based on the integration of RF and D-S evidence theory," (in Chinese), *Vibration and Shock*, Vol. 43, No. 13, pp. 115–125, 2024, https://doi.org/10.13465/j.cnki.jvs.2024.13.013

[19] F. Xie et al., "Research on gearbox compound fault diagnosis based on multi-scale Weber dispersion entropy graph neural network," (in Chinese), *Journal of Electronic Measurement and Instrumentation*, Vol. 39, No. 9, pp. 244–253, 2025, https://doi.org/10.13382/j.jemi.b2407930

[20] Z. Ding, F. Li, X. Xu, and H. Shao, "The spatiotemporal band-gated modal decomposition method and its application in compound fault diagnosis of gearbox," *Advanced Engineering Informatics*, Vol. 69, No. PB, p. 103880, Jan. 2026, https://doi.org/10.1016/j.aei.2025.103880

[21] W. Ma et al., "Application of AOA-CEEMDAN and fusion features in gearbox fault diagnosis," *Mechanical and Electrical Engineering*, Vol. 41, No. 5, pp. 817–826, 2024.

[22] L. Chen et al., "Gear fault diagnosis method based on RF-PCA-improved SVM model," (in Chinese), *Measurement and Control Technology*, Vol. 42, No. 8, pp. 15–21, 2023, https://doi.org/10.19708/j.ckjs.2023.08.003

[23] A. Ding, Y. Qin, B. Wang, L. Guo, L. Jia, and X. Cheng, "Evolvable graph neural network for system-level incremental fault diagnosis of train transmission systems," *Mechanical Systems and Signal Processing*, Vol. 210, p. 111175, Mar. 2024, https://doi.org/10.1016/j.ymssp.2024.111175

Xiaojie Sun, Doctor of Engineering, associate professor, mainly engaged in research on rail vehicle inspection and monitoring, and active control.

Lianxin Wu, Master's student, research interests include deep learning and fault diagnosis.