scientific reports

Check for updates

OPEN

SCBM-Net: a multimodal feature fusion-based dual-channel method for bearing fault diagnosis

Qiang Liu, Weiyan Tong✉, Hongwei Bai & Shien Dong

To address the limitations of single-modal approaches in bearing fault diagnosis under complex operating conditions, this study proposes SCBM-Net—a novel deep learning model based on a dual-channel multimodal fusion architecture. The model innovatively combines Continuous Wavelet Transform (CWT) and Variational Mode Decomposition (VMD) to extract complementary features from time–frequency images and temporal signals, respectively. Specifically, the first channel employs a Swin Transformer to effectively model both local and global representations of CWT-based images through a hierarchical window-based attention mechanism. The second channel adopts a CNN-BiGRU-Attention network to dynamically capture temporal dependencies from intrinsic mode functions decomposed by VMD. Features from both channels are deeply fused using a Multimodal Compact Bilinear Pooling (MCB) module, enhancing fault feature representation and overall model robustness. Experimental results on the CWRU dataset show that SCBM-Net achieves an accuracy of 99.83% under clean conditions. Even under a few-shot learning setting with only 60 training samples per class, the model still maintains a high recognition accuracy of 98.64%, demonstrating strong generalization in low-data scenarios. On an imbalanced dataset, SCBM-Net exhibits stable performance for both majority and minority classes, achieving an average accuracy of 97.33%. In a generalization test on the SEU bearing dataset, the model achieves an accuracy of 98.33%, further validating its cross-platform and cross-domain robustness and transferability. Moreover, under severe noise interference at −10 dB, SCBM-Net retains a fault recognition accuracy of 80.67%, outperforming comparable models and demonstrating excellent noise robustness and practical applicability.

Keywords Bearing fault diagnosis, Parallel dual-channel model, Swin transformer, Multimodal fusion, Complex operating conditions

With the rapid development of industrial automation and intelligent manufacturing technologies, the reliability and stability of equipment operations have attracted increasing attention. As one of the most critical components in mechanical transmission systems, bearing fault diagnosis has become a key research topic in the field of industrial condition monitoring.

In recent years, traditional bearing fault diagnosis methods have mainly relied on time-domain, frequency-domain, and time–frequency analysis techniques[1]. Time-domain feature extraction methods, such as envelope analysis and statistical indicators, can reflect amplitude and energy variations in vibration signals but are often sensitive to noise interference[2]. Frequency-domain methods obtain spectral information through Fourier transform and are effective in identifying periodic faults, but they struggle to characterize the non-stationary nature of signals[3]. To address these limitations, time–frequency methods such as Continuous Wavelet Transform (CWT) and Empirical Mode Decomposition (EMD) have been widely adopted. For instance, Xu et al.[4] employed CWT for multi-resolution analysis of bearing fault signals, effectively extracting local fault features; Boudiaf et al.[5] employed Ensemble Empirical Mode Decomposition (EEMD) to decompose vibration signals and applied wavelet threshold denoising, which significantly improved diagnostic accuracy; however, the issue of mode mixing remained unresolved .

With the advancement of deep learning, Convolutional Neural Networks (CNNs) have achieved remarkable results in bearing fault diagnosis[6]. Ince et al.[7] proposed a one-dimensional CNN model that enables end-to-end classification of temporal vibration signals, enabling real-time fault detection. Janssens et al.[8] introduced a CNN framework that uses spectrum images as input and achieves higher robustness and accuracy than traditional methods through deep feature learning. Jia et al.[9] also achieved promising results with CNN-based feature extraction methods. However, most of these studies were conducted under low-noise laboratory conditions, and the challenges posed by noise interference in real-world scenarios have yet to be fully addressed.

School of Chemical Process Automation, Shenyang University of Technology, Liaoyang 111003, China. ✉email: tongweiyan@sut.edu.cn

For fault diagnosis in noisy environments, Variational Mode Decomposition (VMD) has been widely studied due to its capability to suppress mode mixing. Traditional VMD methods often rely on empirical or exhaustive approaches to select parameters, which are prone to local optima and may compromise decomposition quality and subsequent feature extraction accuracy. Chen et al.[10] enhanced signal clarity by selecting modal components based on the kurtosis criterion, while Du et al.[11] integrated wavelet thresholding to improve denoising performance; however, the issue of parameter sensitivity remains unresolved. To address this, Ma et al.[12] proposed the RIME-VMD method, which employs the RIME algorithm for global optimization and automatically determines the optimal parameter combination for VMD, significantly improving decomposition efficiency and accuracy while avoiding modal redundancy and feature omission. Wang et al.[13] further introduced a method that optimizes VMD parameters via SSA and utilizes RCMDE to extract multiscale complexity features of the signal, thereby enhancing robustness under complex and noisy conditions. Nevertheless, these approaches mostly rely on single-channel temporal modeling with VMD, which limits their ability to fully capture the spectral evolution and spatiotemporal coupling characteristics inherent in fault signals.

In recent years, multimodal information fusion has gradually become a key approach to improving diagnostic accuracy and robustness. Xiao et al.[14] proposed a multi-scale 1D CNN to fuse multi-channel features, enhancing the model's ability to distinguish between various fault types. Lin et al.[15] implemented cross-domain semi-supervised diagnosis based on meta-learning techniques, demonstrating the model's generalization capability across different datasets. However, most existing multimodal approaches still focus on shallow fusion at the input or decision level, failing to deeply integrate temporal and time–frequency domain information. Consequently, there remains a trade-off between representation accuracy and model interpretability in the current fusion mechanisms.

With the success of Vision Transformers (ViT) in image classification tasks, their hierarchical structure and global attention mechanism have also been introduced into bearing fault diagnosis[16]. Tang et al.[17] utilized a ViT-based model to extract features from CWT-generated time–frequency images, leveraging multi-head self-attention to capture spectral patterns at multiple scales, which enabled the effective recognition of weak fault signals. Ji et al.[18] integrated a sliding window attention mechanism into ViT to reduce computational complexity and achieved over 98% diagnostic accuracy on both the CWRU and Southeast University datasets. Furthermore, to enhance noise robustness, Deng[19] introduced a multi-head attention module that significantly improved the model's noise immunity and generalization capability .

Given the non-stationarity and long-range dependencies in one-dimensional vibration sequences, the bidirectional gated recurrent unit (BiGRU) has emerged as an effective sequential modeling technique by capturing contextual information from both past and future time steps. Zhang et al.[20] combined BiGRU with channel attention to perform spatiotemporal feature fusion and employed dual-channel attention (DCA) to extract weighted features from vibration signals, enabling effective diagnosis under complex conditions. Hou et al.[21] further extended the temporal attention mechanism to a multi-head design, allowing the model to simultaneously focus on the temporal evolution of multiple frequency bands, and validated its superiority under variable working conditions. However, using Transformer or BiGRU alone still limits the model to learning features from a single type of data, making it incapable of simultaneously capturing both image-based and temporal modality information.

Although the aforementioned methods have achieved progress in their respective domains, single-modal feature extraction and shallow fusion strategies still struggle to simultaneously capture both global and local information, as well as temporal and time–frequency characteristics. To address these limitations, this study proposes a bearing fault diagnosis method based on a parallel dual-channel model. In this framework, one channel utilizes Continuous Wavelet Transform (CWT) to convert raw vibration signals into time–frequency images, from which discriminative features are extracted using a Swin Transformer. The other channel decomposes the signal using Variational Mode Decomposition (VMD), and constructs a feature extraction network based on Convolutional Neural Networks (CNN), Bidirectional Gated Recurrent Units (BiGRU), and an attention mechanism to effectively model the intrinsic mode components.The features extracted from both channels are fused via Multimodal Compact Bilinear Pooling (MCB), enabling efficient representation of fault information, followed by classification through a fully connected layer.

To ensure the practical applicability of the model across various data sources, it is essential to verify its generalization capability in cross-scenario settings. For instance, Lin et al.[15] proposed a cross-domain semi-supervised bearing fault diagnosis method based on meta-learning and validated their model on multiple datasets. Inspired by this, the present study not only evaluates the diagnostic performance of the proposed model under different noise levels, but also conducts experiments on multiple rolling bearing datasets and imbalanced data distributions to comprehensively demonstrate its cross-domain generalizability. Accordingly, the necessity of this research is reflected in the following aspects.

(1) Multimodal Information Fusion: By fully leveraging the complementary advantages of CWT time–frequency images and VMD-decomposed signals, a deep fusion of image and sequential data is achieved.
(2) Feature Extraction Network Design: The Swin Transformer and CNN-BiGRU-Attention architectures are respectively employed to enable efficient extraction and modeling of features from different modalities.
(3) Compact Bilinear Pooling: The introduction of the MCB module effectively integrates features from the two channels, thereby enhancing both the accuracy and generalization capability of fault diagnosis.
(4) This paper proposes SCBM-Net, a dual-channel model for rolling bearing fault diagnosis, which effectively extracts fault features from non-stationary signals by leveraging both one-dimensional time-series data and two-dimensional time–frequency images. The model achieves accurate and reliable diagnosis performance. To evaluate its generalization capability, extensive experiments including ablation studies, cross-domain validation, few-shot learning, imbalanced data analysis, and noise robustness tests were conducted. The

results demonstrate that the proposed model exhibits superior performance across a range of challenging conditions.

Relevant theories

Continuous wavelet transform

The Continuous Wavelet Transform (CWT)[22] is a time–frequency analysis method widely used for signal processing, capable of simultaneously providing localized information in both the time and frequency domains. Unlike traditional Fourier Transform, CWT performs localized analysis by employing a set of scalable and translatable mother wavelets. This allows for more effective handling of non-stationary signals and transient features, which is particularly significant in practical applications such as bearing fault diagnosis.

Let$x(t)$be a continuous signal, and let$\psi(t)$be a mother wavelet that satisfies certain admissibility conditions (e.g., zero mean,$\int_{-\infty}^{+\infty}\psi(t)dt=0$), The Continuous Wavelet Transform (CWT) of$x(t)$is defined as:

$$W(a,b)=\frac{1}{\sqrt{a}}\int_{-\infty}^{+\infty}x(t)\psi^{*}\left(\frac{t-b}{a}\right)dt \tag{1}$$

Here, α is the scale parameter, which controls the dilation or compression of the wavelet function;β is the translation parameter, determining the position of the wavelet in the time domain; $\psi^{*}(\cdot)$denotes the complex conjugate of the mother wavelet; and$\frac{1}{\sqrt{a}}$is the normalization factor, used to ensure energy consistency across different scales.

The core idea of the CWT lies in computing the inner product between the signal and a family of self-similar wavelet functions, allowing the extraction of localized features at various scales and positions. Specifically, when the scale parameter$\alpha$is small, the wavelet is compressed, resulting in high-frequency resolution; conversely, a largerαstretches the wavelet, capturing low-frequency components. This multi-scale analysis capability makes CWT particularly suitable for detecting signal discontinuities, impacts, and other transient features, which often serve as critical indicators in mechanical fault diagnosis.

Theoretically, as long as the mother wavelet satisfies certain conditions (such as the admissibility condition), the original signal $x(t)$can be perfectly reconstructed from its CWT. The inverse transform is given by:

$$x(t)=\frac{1}{C_{\psi}}\int_{0}^{+\infty}\int_{-\infty}^{+\infty}\frac{1}{\sqrt{a}}W(a,b)\psi\left(\frac{t-b}{a}\right)\frac{dbda}{a^{2}} \tag{2}$$

Here,$C_{\psi}$is the wavelet admissibility constant, defined as:

$$C_{\psi}=\int_{0}^{+\infty}\frac{\left|\hat{\psi}(\omega)\right|^{2}}{\omega}d\omega \tag{3}$$

Where $\hat{\psi}(\omega)$ denotes the Fourier transform of the mother wavelet in the frequency domain.

In this study, the CWT is employed to transform one-dimensional vibration signals into two-dimensional time–frequency images. After comparative analysis, the mother wavelet selected is 'cmor100-1', where cmor refers to the complex Morlet wavelet. In this configuration, the parameter 100 specifies the bandwidth, and 1 denotes the center frequency. This selection enables effective capture of transient variations and localized frequency information within the signal, thereby providing a robust data representation for subsequent feature extraction using the Swin Transformer.

The resulting time–frequency images not only intuitively illustrate the time-varying characteristics of fault signals but also enhance the identification of subtle fault features. Figure 1 presents the CWT time–frequency images for different faulty bearings. As clearly illustrated in Fig. 1, the time–frequency images corresponding to different fault types exhibit significant differences, with well-defined structures and distinct features.

Variational mode decomposition

Signal processing is a critical component of fault diagnosis, and the application of effective signal processing techniques can significantly enhance diagnostic performance[23]. Variational Mode Decomposition (VMD)[24], as an adaptive time–frequency analysis method, is capable of handling nonlinear and non-stationary signals with high analytical precision.

VMD is a non-recursive and adaptive signal decomposition method. Its adaptiveness lies in its ability to determine the number of Intrinsic Mode Functions (IMFs) based on the characteristics of the signal. When the signal is decomposed into K IMFs, the corresponding constrained variational model can be formulated as follows:

$$\begin{aligned}&\min_{|u_k|,|\omega_k|}\left\{\vartheta_t\left[\left(\delta(t)+\tfrac{\mathrm{j}}{\pi t}\right)\cdot u_k(t)\right]\mathrm{e}^{-\mathrm{j}\omega k_2^{t2}}\right\}\\&\text{s.t.}\sum_{k=1}^{K}u_k=x(t)\end{aligned} \tag{4}$$

In the formulation above: – K is the number of modes to be decomposed (a positive integer); $x(t)$ denotes the input signal; uk represents the k-th Intrinsic Mode Function (IMF); wk denotes the center frequency of each IMF.

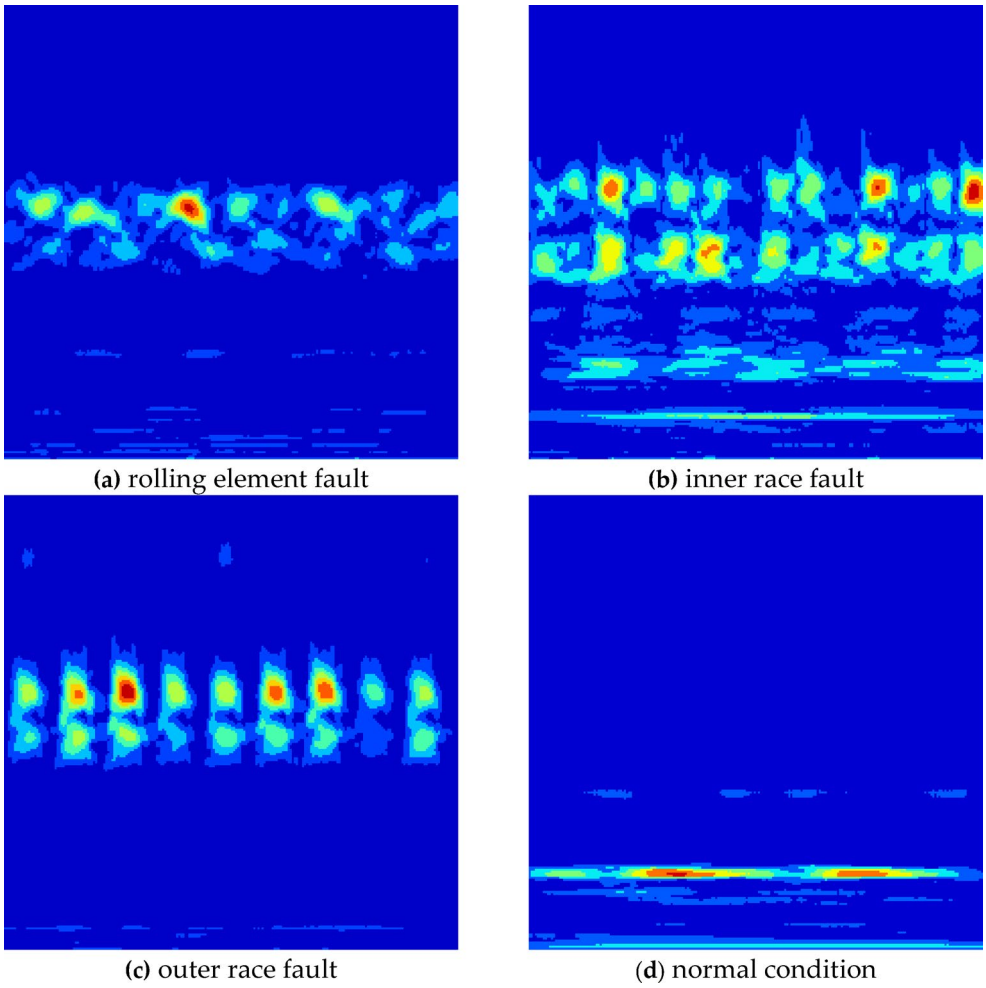


Fig. 1. CWT time-frequency images of different faulty bearings.

To obtain the optimal solution to the variational problem, the Lagrangian function is introduced, leading to the augmented Lagrangian expression:

$$L(\{u_k\},\{\omega_k\},\lambda) = \alpha \sum_k \left\| \vartheta_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) \cdot u_k(t) \right] e^{-i\omega_k t} \right\|_2^2 + \left\| x(t) - \sum_k u_k(t) \right\|_2^2 + \lambda(t) \cdot x(t) - \sum_k u_k(t) \tag{5}$$

α is the penalty parameter, which helps ensure reconstruction accuracy under noisy conditions; λ is the Lagrange multiplier.

Equation (5) is solved using the Alternating Direction Method of Multipliers (ADMM) to find the saddle point of the Lagrangian function, which corresponds to the optimal solution of the constrained variational model.

Both the penalty parameter αand the number of modes K significantly affect the decomposition results. While αprimarily influences the precision of the decomposition, an inappropriate choice of K may lead to modal components that do not correspond well to the actual characteristics of the signal, thereby degrading the effectiveness of subsequent analysis. In this study, K is empirically set to 4 by comparing the center frequencies obtained under different values of K with the frequency content of the original signal.

Convolutional neural network

Convolutional Neural Networks (CNNs) are a class of deep neural networks specialized in extracting local features and are widely used in image processing and signal analysis. In recent years, researchers have increasingly applied CNNs to time-series data to enhance feature extraction capabilities. Essentially, CNNs use filters to extract features from raw data and generate feature vectors, and employ activation functions to solve classification or regression tasks.

In this study, one-dimensional convolution is adopted to extract local features from each mode decomposed by VMD. The convolution operation captures localized time–frequency patterns, while pooling reduces noise and dimensionality, thereby providing enriched feature representations for subsequent sequence modeling.

Let the input signal be *x* (with a length of *n*), the one-dimensional convolutional kernel be *w* (with a length of *k*), and the bias term be *b*, Then, the convolution output at position *i* is given by:

$$y[i] = f\left(\sum_{j=0}^{k-1} w[j] \cdot x[i+j-p] + b\right) \tag{6}$$

$f(\cdot)$ denotes the Rectified Linear Unit (ReLU) non-linear activation function; p is the padding size, which is selected $p = \frac{k-1}{2}$ to preserve the same output length as the input.

The pooling layer is formulated as:

$$y[i] = \max_{j \in \mathcal{R}(i)} x[j] \tag{7}$$

$\mathcal{R}(i)$denotes the pooling region (receptive field) centered at position *i*.

In this study, after multiple layers of convolution and pooling, the feature maps extracted by the CNN are utilized for subsequent sequence modeling. The purpose is to transform the rich local features embedded in the raw signal into higher-level representations.

Bidirectional gated recurrent unit

The Gated Recurrent Unit (GRU)[25] is an improved variant of the traditional Recurrent Neural Network (RNN)[26], designed to mitigate the vanishing gradient problem commonly encountered in long sequences by introducing gating mechanisms. The Bidirectional GRU (BiGRU) further enhances the model by simultaneously processing the sequence in both forward and backward directions, allowing it to capture contextual information from both past and future time steps. This bidirectional structure significantly improves the network's ability to extract and model sequential dependencies.

In this study, BiGRU is employed to perform temporal modeling on the sequential features extracted by the CNN. This enables the network to effectively capture the dynamic variations and long-term dependencies within the vibration signals.

For a single GRU unit, let the current input be xt and the hidden state from the previous time step be $ht-1$. The computation steps are as follows:

(1) Update Gate:

$$z_t = \sigma\left(W_z x_t + U_z h_{t-1} + b_z\right) \tag{8}$$

Where Wz is the input weight matrix, Uz is the recurrent weight matrix, and bz is the bias vector of the update gate. The function $\sigma(\cdot)$ denotes the sigmoid activation, whose output lies in the range [0, 1], controlling the trade-off between retaining the previous information and incorporating new input.

(2) Reset Gate:

$$r_t = \sigma\left(W_r x_t + U_r h_{t-1} + b_r\right) \tag{9}$$

where Wr, Ur, and br are the input weights, recurrent weights, and bias vector of the reset gate, respectively.

(3) Candidate Hidden State:

$$\tilde{h}_t = \tanh\left(W_h x_t + U_h\left(r_t \odot h_{t-1}\right) + b_h\right) \tag{10}$$

Wh: Input weight matrix for the candidate hidden state; Uh: Recurrent weight matrix for the candidate hidden state; bh: Bias vector for the candidate hidden state; $\odot$: Element-wise multiplication operator; $\tanh(\cdot)$Hyperbolic tangent activation function, which introduces nonlinearity and maps the output to the range $[-1, 1]$.

(4) Hidden State Update:

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \tag{11}$$

This equation blends the previous hidden state $ht-1$ and the candidate hidden state $\tilde{h}_t$ according to the update gate zt, thereby determining how much of the past information is preserved.

In the Bidirectional GRU (BiGRU), both forward and backward GRUs are computed at each time step:

Forward GRU: Processes the sequence in the original time order to compute the forward hidden state $\vec{h}_t$.

Backward GRU: Processes the sequence in reverse time order to compute the backward hidden state $\overleftarrow{h_t}$.

Finally, the forward and backward hidden states are concatenated at each time step to form the combined hidden representation:

$$h_t^{bi} = \left[\vec{h}_t; \overleftarrow{h}_t\right] \in \mathbb{R}^{2d_h} \tag{12}$$

This concatenation ensures that the hidden representation at each time step contains information from both the past (forward direction) and the future (backward direction), thereby capturing comprehensive temporal features of the sequence.

Swin transformer

The Swin Transformer[16] is a hierarchical vision transformer designed to address the limitations of traditional transformers in terms of computational complexity and their inefficiency in processing high-resolution images. The Swin Transformer divides an input image into several non-overlapping windows and performs local self-attention within each window. By employing a hierarchical architecture, it progressively merges features across layers, which not only reduces computational cost but also enables the extraction of multi-scale information. Within each local window, the Swin Transformer computes self-attention to capture local contextual relationships. This localized attention mechanism significantly lowers the computational burden while preserving critical local features.

To overcome the limitations imposed by fixed window partitions, the Swin Transformer introduces a shifted window mechanism, in which the windows are shifted across adjacent layers. This strategy allows for cross-window interactions, thereby facilitating feature fusion across different spatial regions. Similar to conventional transformers, the Swin Transformer employs a multi-layer perceptron (MLP) following the self-attention modules. Additionally, residual connections and layer normalization are incorporated to enhance model stability and overall performance.

For the feature representation within a local window $X \in \mathbb{R}^{N\times d}$ (where N denotes the number of patches in the window and d represents the feature dimension), linear projections are first applied to obtain the Query (Q), Key (K), and Value (K) matrices:

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V \tag{13}$$

Where $W^Q, W^K, W^V \in \mathbb{R}^{d\times d}$ are learnable weight matrices.

The attention scores are computed and normalized as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{14}$$

Equation (15) performs self-attention within each local window. Specifically, the similarity between queries and keys is calculated using the dot product, scaled by $\sqrt{d}$ to mitigate the effect of large dot product values, and normalized using the Softmax function. The resulting weights are then used to compute a weighted sum of the value matrix V, yielding the output feature representations.

To capture more comprehensive information, the Swin Transformer typically employs a multi-head self-attention mechanism. This involves executing the self-attention operation in multiple distinct subspaces in parallel, then concatenating the results and projecting them through a linear transformation:

$$\text{MultiHead}(X) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{15}$$

Each attention head is computed as:

$$\text{head}_i = \text{Attention}(XW_i^Q, XW_i^K, XW_i^V) \tag{16}$$

Where W^O is the output projection matrix.

After the attention operation in each Transformer layer, residual connections and layer normalization are applied:

$$\hat{X} = \text{LN}(X + \text{MultiHead}(X)) \tag{17}$$

This is followed by a two-layer Multi-Layer Perceptron (MLP) module:

$$X' = \text{LN}(\hat{X} + \text{MLP}(\hat{X})) \tag{18}$$

The MLP typically consists of two fully connected layers interleaved with a non-linear activation function, enhancing the model's representational capacity.

Multimodal compact bilinear pooling

Multimodal Compact Bilinear Pooling (MCB)[27] enables efficient approximation of outer product interactions between multiple modalities by projecting the input features into a higher-dimensional space using randomized mappings and convolution in the frequency domain. While originally proposed for visual question answering,

MCB can be effectively applied to bearing fault diagnosis due to its ability to capture high-order correlations between heterogeneous modalities, such as time–frequency representations (from CWT) and sequential signal features (from VMD). By modeling cross-modal interactions, MCB enhances fault-discriminative information that may not be apparent when modalities are considered separately, thus providing domain-specific adaptation.

Assuming the input feature vectors $x \in \mathbb{R}^{d_x}$ and $y \in \mathbb{R}^{d_y}$, and a target fusion dimension d, the MCB process comprises four main steps:

Random hashing and sign mapping
Two randomized hash functions $h_x, h_y : [1, d_x] \to [1, d]$are defined to map original feature indices to the target fusion space:

$$hx(i) \sim Uniform\{1, d\}, hy(j) \sim Uniform\{1, d\} \tag{19}$$

Corresponding sign functions $s_x, s_y : [1, d_x] \to \{-1, +1\}$are generated independently:

$$sx(i) \sim Bernoulli(0.5), sy(j) \sim Bernoulli(0.5) \tag{20}$$

These parameters control how the original features are randomly distributed across the fused space. A larger fusion dimension d reduces approximation error, thereby improving the quality of cross-modal interactions.

(1) Count Sketch Projection:

The input features are projected into the fusion space using Count Sketch:

$$\tilde{x} = \mathrm{CountSketch}(x, h_x, s_x), \quad \tilde{y} = \mathrm{CountSketch}(y, h_y, s_y) \tag{21}$$

This step approximates the outer product between *x* and *y*without explicitly computing the $d_x \times d_y$matrix, preserving second-order interactions efficiently.

(2) Frequency-Domain Convolution.

Apply Fast Fourier Transform (FFT) to $\tilde{x}$and $\tilde{y}$:

$$X = \mathrm{FFT}(\tilde{x}), \quad Y = \mathrm{FFT}(\tilde{y}) \tag{22}$$

Element-wise multiplication in the frequency domain yields:

$$Z = X \odot Y \tag{23}$$

This operation corresponds to convolution in the Count Sketch space and encodes cross-modal correlations while maintaining computational tractability.

(3) Inverse Transformation and Real Component Extraction.

Apply inverse FFT and take the real part to obtain the fused feature vector:

$$z = \mathrm{Re}(\mathrm{FFT}^{-1}(Z)) \tag{24}$$

The resulting vector $z \in \mathbb{R}^d$approximates the full second-order interactions between the original feature vectors. This fused representation enhances fault-discriminative information by integrating complementary insights from time–frequency images and sequential signal features while significantly reducing memory and computation costs.

By explicitly stating the role of random projections and sign mappings, MCB provides a theoretically grounded and domain-adapted method for multimodal fusion in bearing fault diagnosis, balancing approximation accuracy and computational efficiency.

SCBM-net method

CWT Channel: To effectively capture the time-frequency characteristics of vibration signals, the image branch employs Continuous Wavelet Transform (CWT) to convert raw one-dimensional signals into two-dimensional time-frequency representations. Specifically, the complex Morlet wavelet function (cmor100-1) is utilized to perform CWT on the input signal, generating a time-frequency image (TFI) of size 224 × 224. This image reflects the energy distribution of the signal across different time and scale domains, enabling robust characterization of non-stationary behavior and modal diversity. The generated TFI is then fed into a Swin Transformer-based feature

extractor. This network consists of four hierarchical stages (Stage 1–4), each comprising patch partitioning (with a patch size of 4×4), linear embedding (embedding dimension = 96), window-based multi-head self-attention (window size = 7), and multi-scale feature aggregation. The final output is a global image-level feature vector with a dimensionality of 768, which serves as input for subsequent multimodal fusion.

VMD Channel: To enhance the model's capability in representing local mode components within non-stationary signals, the sequence branch applies Variational Mode Decomposition (VMD) followed by a lightweight hybrid modeling module based on CNN and BiGRU. VMD decomposes each input signal into $K=4$ Intrinsic Mode Functions (IMFs), which are stacked into a multi-channel input tensor of shape (4, 1024). This structure preserves distinct frequency components and facilitates the modeling of complementary modal information.

The multi-channel tensor is processed through a feature extraction backbone consisting of three one-dimensional convolutional layers followed by a single-layer bidirectional gated recurrent unit (BiGRU). The convolutional block performs convolution, non-linear activation (ReLU), and max pooling operations to capture localized temporal patterns. The resulting sequence is fed into the BiGRU module, which models bidirectional temporal dependencies. An attention mechanism is then applied to perform weighted aggregation over the BiGRU outputs, producing a sequence-level feature vector with a dimensionality of 128. This architecture is designed to effectively extract and integrate temporal features for industrial fault diagnosis tasks.

To further evaluate the computational cost of each module in the dual-channel fault diagnosis framework, we computed the parameter counts and floating-point operations (FLOPs) for Swin Transformer, CNN, BiGRU, Attention, and the MCB fusion module. The results are summarized in Table 1.

It is evident that the Swin Transformer constitutes the majority of both parameters and FLOPs, while the VMD branch modules contribute relatively minor computational load. This distribution is consistent with the design rationale: the CWT branch, processed by the Swin Transformer, extracts rich hierarchical spatial features from time-frequency images, which are critical for accurate fault discrimination. Meanwhile, the VMD branch provides complementary temporal features through CNN, BiGRU, and attention, requiring relatively less computation. Therefore, the Swin Transformer dominating the overall computational complexity is reasonable and aligned with its primary role in capturing the most informative fault-related patterns.

The overall fault diagnosis procedure of the SCBM-Net model is illustrated in Algorithm 1 and Fig. 2. The detailed steps are as follows:

Step 1: The original vibration signals are preprocessed using a sliding window technique. Overlapping time series segments of equal length are generated with a fixed stride, ensuring temporal continuity between samples. The dataset is then divided into training and testing sets according to a predefined ratio to maintain consistency and balance.

Step 2: Continuous Wavelet Transform (CWT) is applied to the preprocessed time series signals, converting one-dimensional vibration data into two-dimensional time-frequency representations. CWT effectively preserves the time-frequency dependency of the signals, maintains the integrity of the features, and introduces spatial characteristics that enhance classification performance. Additionally, the relatively small size of the CWT-generated images reduces storage and computational costs.

Step 3: Variational Mode Decomposition (VMD) is employed to decompose the raw signals into a set of Intrinsic Mode Functions (IMFs). This adaptive decomposition enhances the discriminability of fault-related features and effectively suppresses noise interference. VMD allows for the extraction of key features across multiple temporal scales, improving the separability of different modal components in the feature space.

Step 4: A multimodal dataset is constructed, ensuring the consistency of inputs between the two channels. The time-frequency images generated by CWT and the sequential tensor data obtained from VMD are aligned via a strict sample ID matching mechanism. This guarantees synchronization of input data across both channels and enables joint feature extraction from time-frequency images and time-domain signals.

Step 5: A parallel dual-channel deep learning model is constructed. In Channel 1, a Swin Transformer is used to extract features from the CWT time-frequency images. Its hierarchical representation structure and shifted window self-attention mechanism allow effective modeling of both local and global information, enhancing the network's capability to express image features. In Channel 2, a CNN-BiGRU-Attention network is employed to extract features from the VMD-decomposed sequential signals. CNN captures local temporal patterns, BiGRU models temporal dependencies in both forward and backward directions, and the attention mechanism adaptively emphasizes key fault-related features while suppressing irrelevant information.

Step 6: The features extracted from both channels are fused to enhance fault classification performance. A Multimodal Compact Bilinear (MCB) pooling module is used to deeply integrate features from both channels, achieving complementary enhancement of multimodal information. Adaptive average pooling and dimensionality reduction are applied to reduce computational complexity. In the final classification stage, a fully connected layer followed by a Softmax classifier is used to identify fault types. The effectiveness and robustness of the model are validated through experimental evaluation metrics and visualization analyses.

Module	Parameters	FLOPs
Swin transformer	86.87 M	15466.9 M
CNN	31.3 K	19.66 M
BiGRU	494.6 K	1346.37 M
Attention	98.7 K	134.48 M
MCB	0	57.34 K

Table 1. Model complexity analysis of each module (parameters and FLOPs).

```
# Input:
#   D = {x1, x2, ..., xn}        # Raw vibration signals with associated labels
# Output:
#   y_pred_list                  # Predicted fault labels for test samples

# Step 1: Signal Segmentation
segments = []
for each signal x in D do
# Apply sliding window segmentation (window size: 1024, overlap: 50%)
segs = SlidingWindowSplit(x, window_size=1024, overlap=0.5)
assign unique identifiers and labels to segs
append segs to segments
end for

# Step 2: Dataset Partitioning
(D_train, D_test) = StratifiedSplit(segments, train_ratio=0.7)

# Step 3: Time-Frequency Channel via CWT
function CWT_Channel(sample):
# Perform Continuous Wavelet Transform using the complex Morlet wavelet (cmor100-1)
TFI = ContinuousWaveletTransform(sample, wavelet='cmor100-1')  # Shape: (224, 224)

# Extract deep image features using Swin Transformer
F_img = SwinTransformerFeatureExtractor(TFI)
- Patch Partitioning (size: 4×4)
- Linear Embedding (dim: 96)
- Hierarchical attention across 4 stages (window-based MHSA, window size: 7)
- Final output dimension: 768
return F_img
end function

# Step 4: Temporal Modeling Channel via VMD
function VMD_Channel(sample):
# Decompose the signal into K=4 intrinsic mode functions using VMD
IMFs = VariationalModeDecompose(sample, K=4, alpha, tau)
T = StackChannels(IMFs)  # Shape: (4, 1024)

# Extract temporal features using lightweight CNN + BiGRU + Attention architecture
F_seq = CNN_BiGRU_Attention(T)
- CNN Block:
Conv1D → ReLU → MaxPooling
Output shape: (16, 512)
```

Algorithm 1. SCBM-net fault diagnosis.

Theoretical justification of SCBM-Net

To provide a rigorous theoretical foundation for the proposed SCBM-Net and explain why its architectural design achieves superior performance in bearing fault diagnosis, this section presents a detailed analysis from three complementary perspectives: feature complementarity, dual-channel design effectiveness, and the representational advantages of multimodal compact bilinear (MCB) fusion.

```
- BiGRU Block:
Single-layer bidirectional GRU
Output shape: (batch_size, 128)
- Attention Mechanism:
Apply self-attention to aggregate BiGRU outputs
Output dimension: 128
return F_seq
end function

# Step 5: Feature Fusion via Compact Bilinear Pooling
function FuseFeatures(F_img, F_seq):
# Fuse image and sequence features: F_img ∈ ℝ^768, F_seq ∈ ℝ^128
Z = CompactBilinearPooling(F_img, F_seq, output_dim=128)
Z_pooled = GlobalAveragePooling(Z)  # Final fused feature: ℝ^128
return Z_pooled
end function

# Step 6: Classification Layer
function Classify(Z):
logits = FullyConnectedLayer(Z, num_classes)  # Linear projection:
128 → num_classes
y_pred = Softmax(logits)
return y_pred
end function

# Step 7: Model Training
initialize model parameters Θ
for epoch = 1 to N_epochs do
for each batch B in D_train do
for each sample s in B do
F_img = CWT_Channel(s.data)
F_seq = VMD_Channel(s.data)
Z = FuseFeatures(F_img, F_seq)
y_pred = Classify(Z)
loss = CrossEntropy(y_pred, s.label)
BackpropagateAndUpdate(Θ, loss)
end for
end for
end for

# Step 8: Inference and Evaluation
y_pred_list = []
for each sample s in D_test do
F_img = CWT_Channel(s.data)
F_seq = VMD_Channel(s.data)
Z = FuseFeatures(F_img, F_seq)
y_pred = Classify(Z)
append y_pred to y_pred_list
end for

# Compute overall prediction accuracy
accuracy = ComputeAccuracy(y_pred_list, true_labels)

End Algorithm
```

Fig. . (continued)

Complementarity of CWT and VMD features

Let $x(t)$ denote the raw vibration signal. The Continuous Wavelet Transform (CWT) represents the signal as:

$$W_x(a,b) = \int_{-\infty}^{\infty} x(t)\psi_{a,b}^*(t)\,dt \tag{25}$$

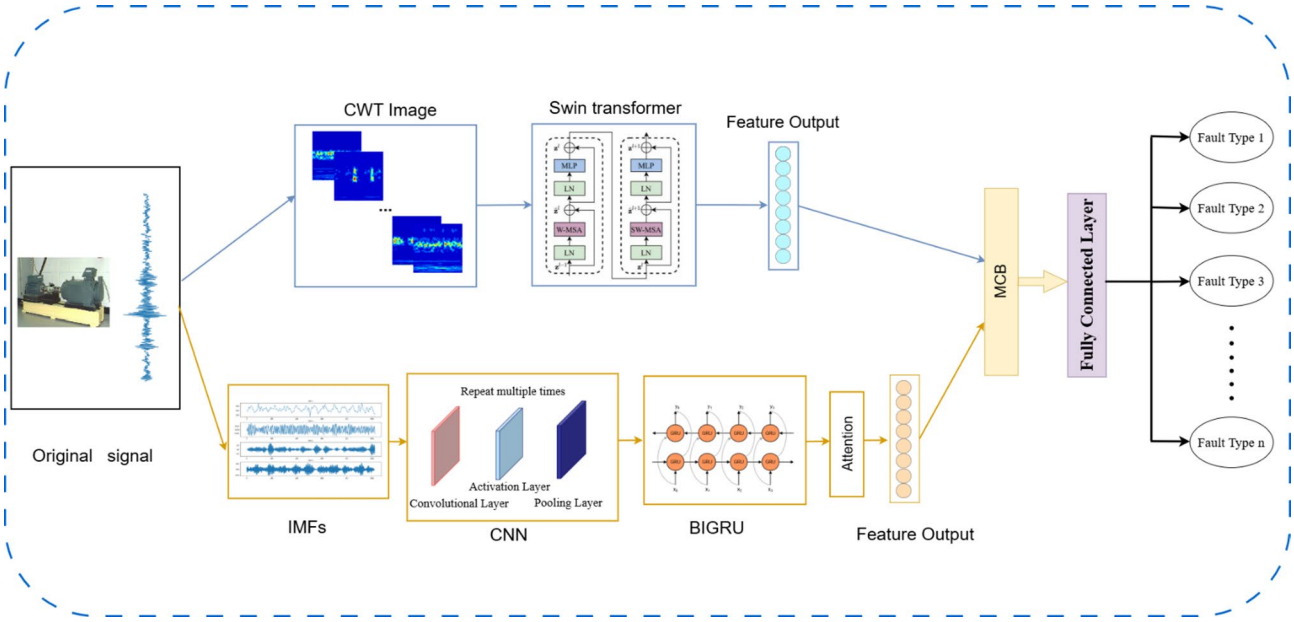


Fig. 2. Flowchart of the SCBM-Net Model.

where a and b are the scale and translation parameters, and $\psi_{a,b}(t)$ is the wavelet basis function. Small values of a emphasize high-frequency transient components, whereas large a capture low-frequency oscillatory trends. Hence, CWT effectively encodes localized transient and impact features, which are essential for identifying fault events in bearings.

On the other hand, Variational Mode Decomposition (VMD) decomposes $x\,(t)$ adaptively into K intrinsic mode functions (IMFs):

$$x(t) = \sum_{k=1}^{K} u_k(t), \quad u_k(t) \sim \text{narrow-band signal with center frequency } \omega_k. \tag{26}$$

Lower-order IMFs represent high-frequency impulsive patterns, while higher-order IMFs capture low-frequency oscillatory behavior. VMD thus provides an alternative representation of the signal dynamics, focusing on intrinsic oscillatory modes and long-range temporal dependencies.

From a representational perspective, CWT and VMD features are complementary: CWT emphasizes local transient information, while VMD captures global oscillatory patterns. This justifies the dual-channel design: single-modality models cannot simultaneously encode local impacts and global oscillatory dynamics, whereas a dual-channel architecture integrates both sources of information, theoretically enhancing discriminative capacity.

Representational advantage of MCB fusion

A key innovation of SCBM-Net is the use of Multimodal Compact Bilinear (MCB) pooling, which captures second-order interactions between heterogeneous features. Let $\mathbf{f}_{\mathrm{CWT}} \in \mathbb{R}^{d_1}$ and $\mathbf{f}_{\mathrm{VMD}} \in \mathbb{R}^{d_2}$ denote the feature vectors extracted from the CWT and VMD channels, respectively. The fused representation is approximated as:

$$f_{\mathrm{MCB}} \approx f_{\mathrm{CWT}} \otimes f_{\mathrm{VMD}} \tag{27}$$

where $\otimes$ denotes the vector outer product. Unlike simple concatenation, MCB encodes nonlinear cross-modal dependencies, improving feature separability in the joint latent space.

Specifically, for sample , the fusion can be expressed as:

$$(f_{\mathrm{MCB}})_i = \mathcal{F}^{-1}\left(\mathcal{F}(\mathrm{CS}(f_{\mathrm{CWT}})) \cdot \mathcal{F}(\mathrm{CS}(f_{\mathrm{VMD}}))\right) \tag{28}$$

where $\mathrm{CS}(\cdot)$denotes Count Sketch projection, $\mathcal{F}$ and $\mathcal{F}^{-1}$are Fourier transform and its inverse, and the pointwise multiplication in the frequency domain implements convolution. This mechanism systematically captures second-order cross-modal interactions, which theoretically increases discriminative power in complex fault scenarios.

Feature redundancy and complementarity

While CWT and VMD provide distinct perspectives of the same vibration signal, it is theoretically necessary to characterize the relationship between their extracted features in terms of redundancy and complementarity.

Let $\mathbf{f}_{CWT} \in \mathbb{R}^{d_1}$ and $\mathbf{f}_{VMD} \in \mathbb{R}^{d_2}$ denote the feature vectors obtained from the CWT and VMD channels, respectively.

Feature Redundancy: Redundancy reflects the extent of overlapping information between modalities. It can be quantitatively defined using normalized mutual information:

$$R(\mathbf{f}_{CWT}, \mathbf{f}_{VMD}) = \frac{I(\mathbf{f}_{CWT}; \mathbf{f}_{VMD})}{\min\{H(\mathbf{f}_{CWT}), H(\mathbf{f}_{VMD})\}} \tag{29}$$

where $I(\cdot;\cdot)$denotes the mutual information and $H(\cdot)$the Shannon entropy. A large *R* value indicates high redundancy, meaning that both feature sets carry similar information.

Feature Complementarity: Complementarity describes the additional information gained when combining features from both modalities. It can be measured by:

$$C(\mathbf{f}_{CWT}, \mathbf{f}_{VMD}) = H(\mathbf{f}_{CWT}, \mathbf{f}_{VMD}) - \max\{H(\mathbf{f}_{CWT}), H(\mathbf{f}_{VMD})\} \tag{30}$$

where $H(\mathbf{f}_{CWT}, \mathbf{f}_{VMD})$ is the joint entropy. If$C > 0$, the fused representation encodes more information than any single modality, demonstrating the presence of complementary characteristics.

These definitions provide a rigorous theoretical basis for the dual-channel design: CWT emphasizes localized transient dynamics while VMD encodes oscillatory modes, and their fusion is expected to maximize *C* while controlling *R*. This justifies the use of MCB pooling to exploit second-order interactions while mitigating redundant correlations.

Uniqueness of the dual-channel architecture

SCBM-Net's advantage does not arise merely from combining existing modules, but from a carefully designed sample-aligned dual-channel architecture:

CWT Channel: The Swin Transformer models hierarchical spectral structures in CWT images, providing global–local contextual representation of fault transients.

VMD Channel (CNN-BiGRU-Attention): CNN extracts local IMF patterns, BiGRU captures long-range temporal dependencies, and the attention mechanism adaptively highlights the most fault-relevant modes.

MCB Fusion Layer: Achieves second-order interactions in the feature space, theoretically enhancing representational capacity and improving separability.

The precise alignment between CWT images and their corresponding VMD decompositions ensures that complementary information is maximally leveraged, which cannot be achieved by naive feature concatenation or independent unimodal models.

Theoretical insight and methodological significance

In summary, SCBM-Net's theoretical advantage and uniqueness can be outlined as follows:

Integrates complementary information from CWT and VMD features, capturing both local transients and global oscillatory behavior.

Utilizes MCB pooling to encode second-order cross-modal interactions, enhancing feature separability and discriminative capacity.

Employs a sample-aligned dual-channel architecture with attention-guided selection of informative VMD modes, further improving fault-relevant feature representation.

This design is theoretically justified: the dual-channel architecture systematically exploits complementary signal representations, and MCB fusion captures nonlinear cross-modal dependencies. These aspects provide clear theoretical reasoning for the superior empirical performance of SCBM-Net over simple module combinations or unimodal approaches.

Experimental validation

Data source

To evaluate the effectiveness of the proposed fault diagnosis method, experiments were conducted using the publicly available bearing dataset provided by Case Western Reserve University (CWRU), USA. This dataset includes bearing vibration signals collected under various fault conditions and operating scenarios. The experimental testbed is illustrated in Fig. 3.

In this study, vibration signals were collected from the drive-end sensor under an operational condition with a motor speed of 1772 RPM and a sampling frequency of 12 kHz. The investigated bearing faults include three typical defect types: inner race, outer race, and ball faults, with defect diameters of 0.1778 mm, 0.3556 mm, and 0.5334 mm, respectively. The detailed information on the dataset, including samples from normal bearings and ten different fault categories, is summarized in Table 2.

All experiments were conducted using the PyTorch deep learning framework. The hardware configuration used for training and evaluation comprises an Intel(R) Core(TM) i9-14900 K CPU at 3.20 GHz, an NVIDIA GeForce RTX 4090 GPU, and 64 GB of RAM. To ensure the reliability and consistency of the experimental results, each experiment was repeated with different random seeds, and the average performance was reported.

The model training time per epoch was approximately 7 s. Memory consumption is moderate, with each batch of 32 samples requiring approximately 3–4 GB of GPU memory. The proposed method is amenable to parallelization across multiple GPUs, which enables potential scalability to larger datasets. Inference is sufficiently fast for real-time industrial fault diagnosis applications, demonstrating that the method is both computationally efficient and practically deployable.

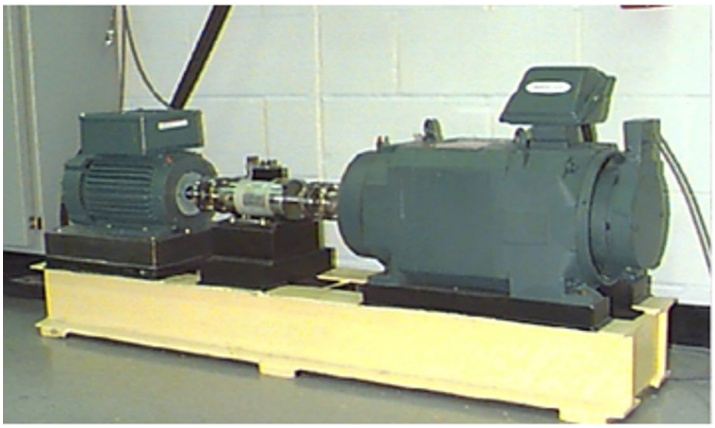


Fig. 3. CWRU bearing vibration dataset experimental platform.

Fault type	Label	Damage diameter (mil)	Training set	Test set
N	0	–	140	60
IR007	1	7	140	60
B007	2	7	140	60
OR007	3	7	140	60
IR014	4	14	140	60
B014	5	14	140	60
OR014	6	14	140	60
IR021	7	21	140	60
B021	8	21	140	60
0R021	9	21	140	60

Table 2. Public bearing dataset collected by CWRU.

To eliminate the possibility of cherry-picking, all critical hyperparameters were optimized through systematic grid search combined with validation performance monitoring. Candidate values for the learning rate were set to {1e-5, 1e-4, 1e-3}, while batch sizes of {16, 32, 64} were tested. The optimal configuration was determined based on the highest average F1-score on the validation set.

Model training was performed with the Adam optimizer and weight decay for regularization. A StepLR scheduler was applied to reduce the learning rate by a factor of 0.1 every 10 epochs, thereby facilitating stable convergence. The batch size was set to 32, which provided a good balance between convergence stability and GPU memory efficiency. Early stopping with a patience of 5 epochs on the validation accuracy was used to prevent overfitting, and dropout ($p = 0.2$) was applied in the BiGRU layers. The cross-entropy loss was employed as the optimization objective. Training proceeded for up to 50 epochs, and the model parameters that achieved the best validation accuracy were preserved for testing.

To ensure reproducibility, all experiments were conducted with fixed random seeds applied to PyTorch, NumPy, and Python's random module, and deterministic behavior was enforced by disabling non-deterministic CUDA operations. The main implementation environment included Python 3.10, PyTorch 2.6.0 with CUDA 12.4, and torchvision 0.21.0 with CUDA 12.4. All critical training and model parameters, including optimizer settings, learning rate schedule, batch size, number of epochs, early stopping criteria, regularization techniques, model dimensions, and random seed, are summarized in Table 3, providing sufficient information for replication of all reported results.

VMD data processing

The continuous one-dimensional time-series vibration signals are segmented into multiple samples of fixed length, each containing 1024 data points, using a sliding window approach with a 50% overlap rate. Each sample is assigned a unique global sample ID and its corresponding fault label. To ensure data balance across different fault categories, 200 samples are randomly selected from each fault type. These are further divided into 140 training samples and 60 testing samples per category, ensuring that the dataset satisfies the experimental requirements.

In this study, the Variational Mode Decomposition (VMD) algorithm is employed to preprocess the original signals, which are subsequently converted into PyTorch tensors. For each sample, the VMD algorithm adaptively decomposes the signal into a series of Intrinsic Mode Functions (IMFs). Taking the IR007 fault signal as an example, the time-domain decomposition results are illustrated in Fig. 4(a). The original signal is decomposed

Category	Parameter	Value/range
Optimizer	Adam	lr = 0.0001, weight decay = 1e − 5
Learning rate schedule	StepLR	Step size = 10, gamma = 0.1
Batch size	–	32
Epochs	–	50
Early stopping	Validation accuracy	Patience = 5
Loss function	–	CrossEntropyLoss
Dropout	BiGRU layers	0.2
Random seed	–	42

Table 3. Model training parameters and hyperparameter values.

into multiple IMFs (IMF1–IMF4), each exhibiting distinct oscillatory characteristics corresponding to different frequency components. Lower-order IMFs primarily capture high-frequency transient impact features, highlighting sharp fault pulses, whereas higher-order IMFs reflect more stable low-frequency trends, aiding in the representation of global vibration patterns and slow-varying features. This multiscale time-domain decomposition enables a clear separation of fault-related information across various frequency bands, thus enhancing the input quality for downstream feature extraction tasks.

Figure 4(b) presents the frequency-domain decomposition of the VMD-processed signal. Each IMF shows concentrated spectral energy in specific frequency bands: IMF1 and IMF2 exhibit prominent peaks in the high-frequency range, capturing transient impact signatures, while IMF3 and IMF4 concentrate energy in the mid- and low-frequency bands, representing broader structural vibrations and global trends. Compared with the spectrum of the original signal, the decomposed components exhibit cleaner spectral characteristics with reduced noise interference, thereby improving the discriminability of fault features.

Figure 4(c) compares the original signal, the reconstructed signal, and the residual. The reconstructed signal (red dashed line) closely overlaps with the original signal (black solid line), while the residual (green line) shows minimal amplitude with no apparent periodic patterns. This indicates that VMD decomposition effectively preserves the principal signal components while suppressing noise, thereby achieving high-fidelity signal reconstruction.

In summary, the VMD algorithm, through its multimodal decomposition strategy, enhances fault-related features across both time and frequency domains. It successfully retains the global trend while extracting local transient details, providing rich and high-quality input features for subsequent fault classification and identification.

To rigorously justify the selection of the decomposition mode number K and evaluate the quality of VMD decomposition, we conducted a systematic analysis using both quantitative metrics and physical validation.

The selection of the decomposition mode number K is critical for ensuring that Variational Mode Decomposition (VMD) can effectively separate the intrinsic mode functions (IMFs) corresponding to different vibration components in bearing signals. To systematically justify the choice of K, we evaluated K = 2,3,4,5,6 and quantified the decomposition quality using reconstruction error (RE) and orthogonality index (OI). In addition, the classification performance of the VMD-based channel model—including accuracy, precision, recall, and F1 score—was assessed for each K. The results are summarized in Table 4.

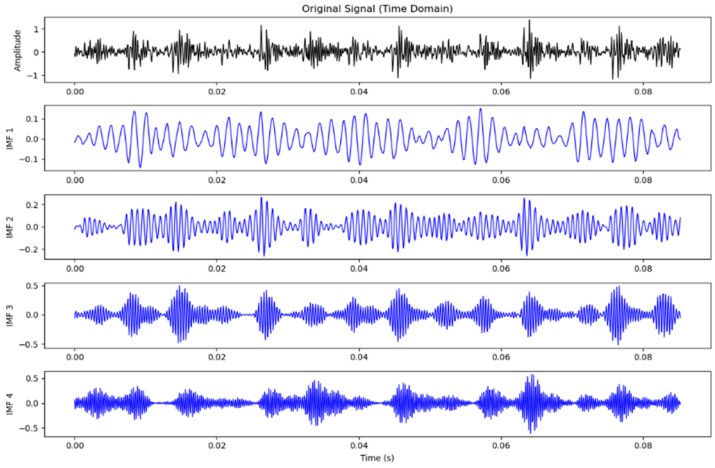
Analysis of Table 4 shows that K = 4 achieves the best trade-off between decomposition fidelity and orthogonality, while simultaneously maximizing classification accuracy. Lower K values result in insufficient mode separation, leading to overlapping frequency components and reduced classification performance. Conversely, higher K values introduce redundant IMFs without improving classification, and can even slightly degrade performance due to mode splitting.

The penalty parame $\alpha$and convergence tolerance $\tau$are critical hyperparameters in VMD that control the trade-off between mode bandwidth and reconstruction fidelity. We performed a grid search with $\alpha \in \{1000, 2000, 5000\}and\tau \in \{10^{-6}, 10^{-7}, 10^{-8}\}$, and evaluated the resulting decomposition quality (RE, OI) and classification performance. The results are summarized in Table 5.

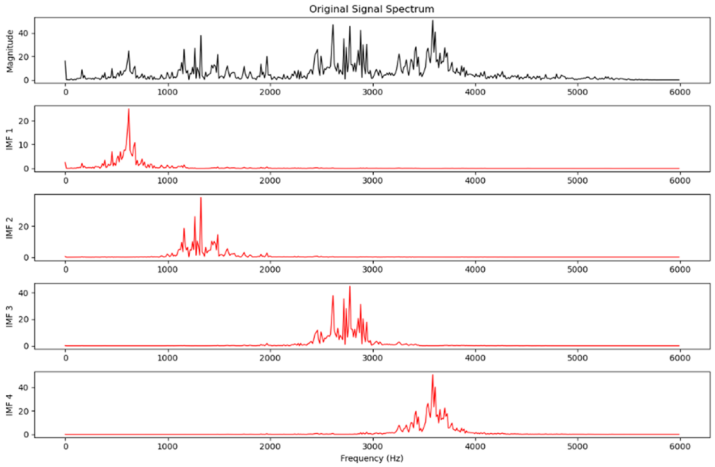
The optimal combination of α=2000 and $\tau = 1 \times 10^{-7}$ yields both low reconstruction error and high orthogonality while achieving the highest classification accuracy (97.50%) and F1-score (0.9732). These results demonstrate that proper tuning of α and τ is essential to ensure that each IMF captures a physically meaningful component rather than numerical noise.

To validate the physical significance of the extracted intrinsic mode functions (IMFs), a representative analysis was conducted using the de_7_inner bearing signal, which is known to contain an inner race fault with a theoretical characteristic frequency of 161 Hz. Figure 5 presents the envelope spectra of the four IMFs obtained via the VMD algorithm with parameters$K = 4$, α = 2000, and $\tau = 1 \times 10^{-7}$. Each IMF exhibits distinct frequency components corresponding to different vibration modes of the bearing, and the theoretical fault frequency (161 Hz), marked by a black dashed line, aligns closely with the dominant peaks across the four IMFs. The clear separation of frequency components demonstrates the effectiveness of the modal decomposition and confirms that the extracted IMFs capture physically meaningful fault-related signals.

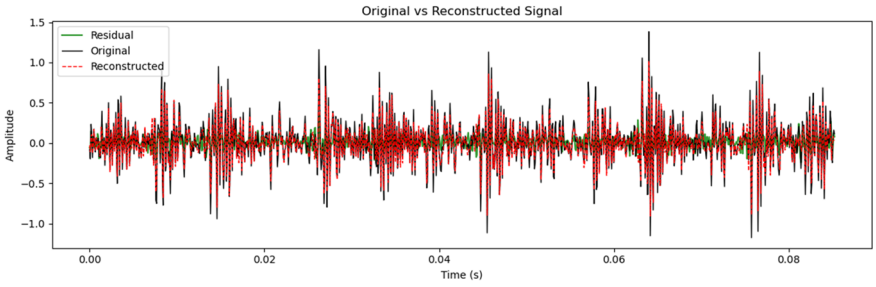
The above analysis demonstrates that choosing $K = 4$with optimized parameters α and τ yields IMFs of high quality and clear physical interpretability. This configuration not only minimizes reconstruction error and enhances mode orthogonality, but also facilitates superior classification performance, as the downstream



(a) Time-domain decomposition of the vibration signal using VMD



(b) Frequency-domain representation of the decomposed VMD signal



(c) Comparison between the original signal, reconstructed signal, and residual

Fig. 4. VMD decomposition of the vibration signal and its frequency-domain representation.

CNN-BiGRU-Attention model can effectively extract discriminative features from the well-separated IMFs. Consequently, the VMD parameter selection is justified both quantitatively, through reconstruction error and orthogonality index, and physically, through envelope spectrum validation, thereby providing a robust and reliable basis for bearing fault diagnosis.

K	RE	OI	Accuracy (%)	Precision	Recall	F1-score
2	0.532234	0.002951	85.83%	0.8676	0.8583	0.8453
3	0.302973	0.113116	91.83%	0.9223	0.9183	0.9147
4	0.260675	0.282948	97.50%	0.9726	0.9750	0.9732
5	0.208848	0.254858	94.67%	0.9496	0.9467	0.9446
6	0.172572	0.337792	94.67%	0.9497	0.9467	0.9442

Table 4. Performance comparison at different modal numbers K.

α	τ	RE	OI	Accuracy (%)	Precision	Recall	F1-score
1000	10^{-6}	0.185133	0.339019	94.00%	0.9442	0.9400	0.9347
1000	10^{-7}	0.184928	0.339040	94.33%	0.9508	0.9433	0.9409
1000	10^{-8}	0.184867	0.339051	95.33%	0.9544	0.9533	0.9516
2000	10^{-6}	0.261097	0.282827	96.83%	0.9695	0.9683	0.9677
2000	10^{-7}	0.260675	0.282948	97.50%	0.9726	0.9750	0.9732
2000	10^{-8}	0.260630	0.282974	96.50%	0.9651	0.9650	0.9643
5000	10^{-6}	0.393508	0.090837	97.00%	0.9698	0.9700	0.9697
5000	10^{-7}	0.393401	0.090806	96.50%	0.9660	0.9650	0.9638
5000	10^{-8}	0.393400	0.090808	96.67%	0.9672	0.9667	0.9662

Table 5. Performance comparison at different parameters $\alpha$and τ.

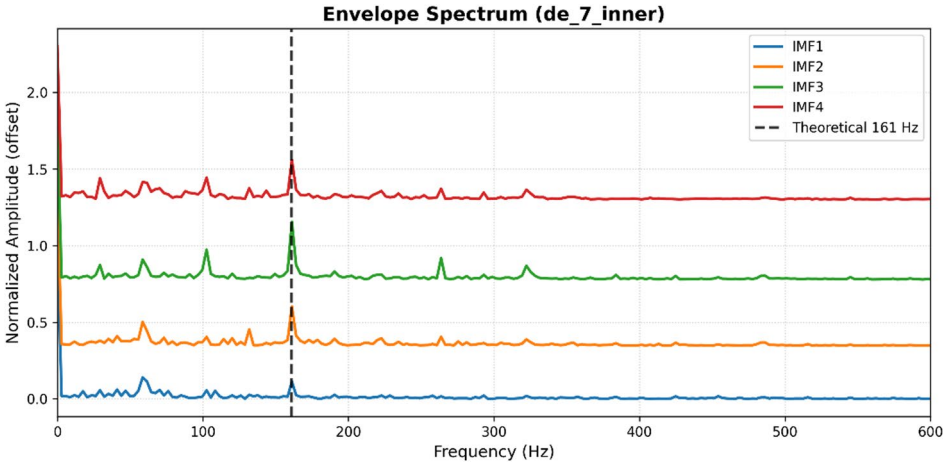


Fig. 5. Example envelope spectrum for de_7_inner operating condition.

CWT data processing

The vibration signal samples in the dataset are transformed into two-dimensional time–frequency images using the Continuous Wavelet Transform (CWT) algorithm. These images are then used as the input data for the Swin Transformer network. As illustrated in Fig. 6, the time–frequency images corresponding to the ten different bearing conditions exhibit notable variations in texture patterns and energy distributions. These visual differences provide the fault diagnosis model with critical discriminative information, thereby enhancing classification accuracy and robustness to noise.

In scenarios involving similar fault types or complex noise backgrounds, the refined time–frequency features extracted by CWT offer more informative representations. These features facilitate multi-level convolution and attention mechanisms within the model, enabling the network to acquire deeper discriminative cues and ultimately achieve accurate identification of fault modes.

To rigorously validate the selection of the mother wavelet and its associated parameters for rolling bearing fault diagnosis, a comparative study was conducted. Specifically, six representative wavelets were evaluated: complex Morlet wavelets (cmor40-1, cmor100-1, cmor160-1), Daubechies wavelet (db8), Mexican Hat wavelet (mexh), and the classical Discrete Wavelet Transform (DWT) approach. All experiments employed the same dataset, a consistent sliding window segmentation strategy, and fixed training configurations to ensure comparability.

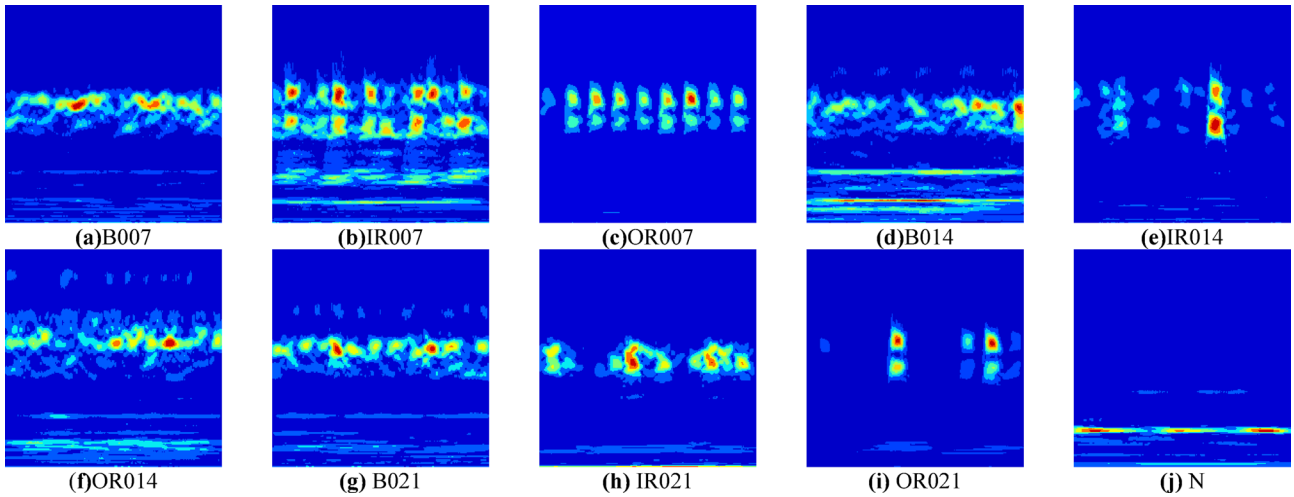


Fig. 6. Two-dimensional time-frequency images.

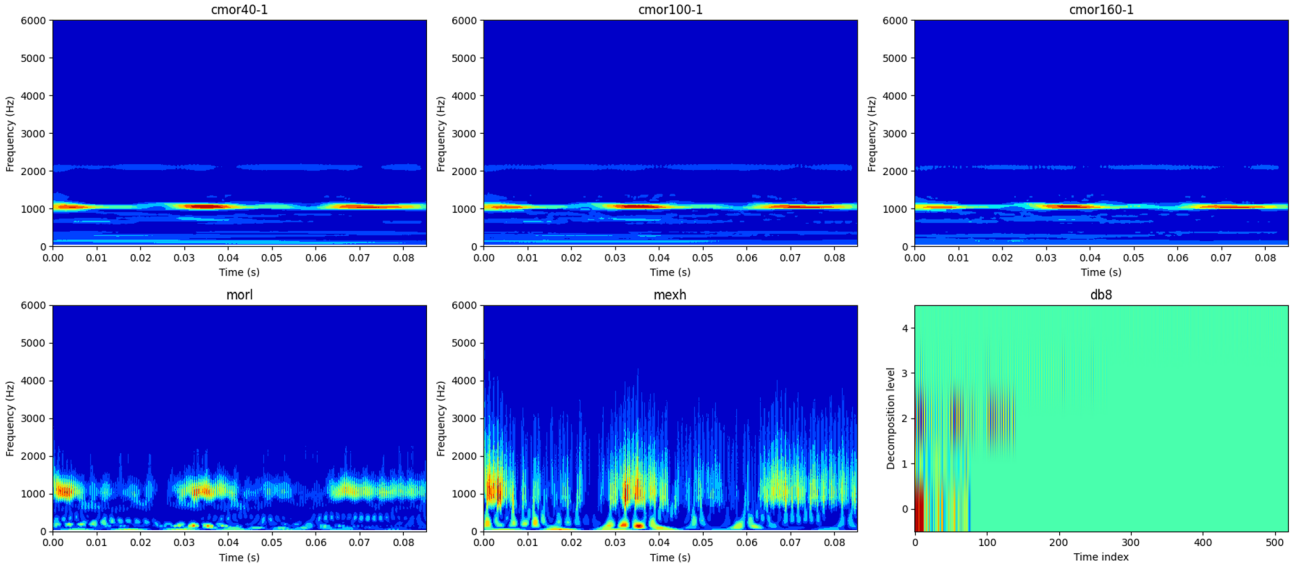


Fig. 7. Comparison of mother wavelets for time–frequency representation.

Representative time–frequency comparison images of sample signals are shown in Fig. 7, and these images were used as inputs to the Swin Transformer network. The corresponding classification performance is summarized in Table 6.

As illustrated in Fig. 7, the comparative visualization further demonstrates that the cmor100-1 wavelet provides a clearer capture of transient impulses and generates more stable and compact time–frequency features.

As summarized in Table 6, the cmor100-1 wavelet achieves the best classification performance among all considered wavelets. This superiority can be attributed to the consistency between its theoretical characteristics and the inherent frequency properties of rolling bearing fault signals.

Theoretical basis for parameter selection: Based on a 12 kHz sampling frequency and typical bearing geometric parameters, bearing fault pulse signals are primarily concentrated in the 1–10 kHz frequency range. In the cmor100-1 wavelet, the bandwidth parameter (Bandwidth = 100) was specifically selected to enhance frequency resolution, enabling precise separation of narrowband pulse components associated with local faults. Meanwhile, the center frequency (Center frequency = 1) ensures time fidelity, minimizing distortion of transient events and preserving the sharp signal features required for accurate fault diagnosis.

Time-Frequency Resolution Trade-off: The continuous wavelet transform (CWT) inherently exhibits a fundamental trade-off between temporal and spectral resolution. Mathematically, the time resolution Δt and frequency resolution Δf satisfy the inequality:

$$\Delta t \cdot \Delta f \geqslant \frac{1}{4\pi} \tag{31}$$

Mother wavelet	Accuracy (%)	Precision	Recall	F1-score
cmor40-1	98.50%	0.9857	0.9850	0.9850
cmor100-1	99.33%	0.9938	0.9933	0.9951
cmor160-1	99.00%	0.9900	0.9900	0.9900
morl	94.67%	0.9517	0.9467	0.9459
mexh	92.67%	0.9305	0.9267	0.9271
DWT	50.50%	0.5165	0.5050	0.4906

Table 6. Classification performance with different mother wavelets.

Patch size	Window size	Accuracy	Precision	Recall	F1 Score
4	5	98.50%	0.9870	0.9850	0.9850
4	7	99.33%	0.9938	0.9933	0.9951
4	9	98.50%	0.9858	0.9850	0.9850
8	5	99.17%	0.9920	0.9917	0.9917
8	7	97.83%	0.9798	0.9783	0.9781
8	9	99.00%	0.9904	0.9900	0.9900

Table 7. Patch size and window size optimization for Swin Transformer.

Increasing the bandwidth improves frequency resolution but compromises temporal localization, whereas decreasing the bandwidth enhances temporal resolution at the expense of frequency discrimination. Considering that rolling bearing fault diagnosis relies critically on capturing narrow-band impulsive events, cmor100-1 provides an optimal compromise between time and frequency resolution, balancing the need for detailed frequency information and accurate temporal representation.

To further evaluate robustness, a sensitivity study was conducted by varying the bandwidth parameter across [40, 100, 160] and the center frequency across [0.5, 1, 2]. The classification accuracy fluctuated within ± 0.9%, This confirms that the proposed diagnostic approach does not critically depend on a narrowly tuned mother wavelet, thereby ensuring generalizability.

In conclusion, the selection of cmor100-1 is strongly justified by (i) its theoretical capacity to balance time-frequency resolution in accordance with the characteristics of bearing fault signals, (ii) its superior empirical performance in classification tasks, and (iii) its demonstrated robustness under parameter variations.

CWT channel network design

To provide theoretical support for the architectural choices in the proposed dual-channel model, we analyze the use of Swin Transformer in the CWT channel from multiple perspectives, including hierarchical representation, patch and window size selection, computational efficiency, and the shifted window attention mechanism.

The continuous wavelet transform (CWT) converts vibration signals into time-frequency images that simultaneously contain local transient impulses and global spectral patterns. The hierarchical structure of Swin Transformer is particularly suitable for extracting multiscale features from such images: lower layers capture fine-grained local patterns corresponding to early-stage fault impulses, while higher layers encode more global frequency information. This multiscale modeling aligns naturally with the characteristics of CWT images and facilitates discriminative feature learning, which may be less efficiently captured by standard ViT or CNN architectures due to their fixed-scale receptive fields.

The choice of patch size and window size plays a critical role in balancing local detail extraction and computational efficiency. To systematically evaluate these parameters, we conducted experiments with varying patch sizes {4, 8} and window sizes {5, 7, 9}. The results, summarized in Table 7, indicate that a patch size of 4×4 combined with a window size of 7×7 achieves the best trade-off between fine-grained feature extraction and global context modeling, justifying its selection.

To further validate the design, we compared Swin Transformer with alternative architectures including ViT, ConViT, DeiT-S, and Mixer-B under identical dataset and training conditions. As shown in Table 8, Swin Transformer achieves the highest overall accuracy while maintaining moderate computational complexity (FLOPs) and parameter size, demonstrating a favorable balance between representation capacity and efficiency.

An important component contributing to Swin Transformer's effectiveness is the shifted window attention mechanism, which partitions the feature map into non-overlapping windows and alternates window positions across layers. This design allows cross-window information exchange while keeping computational complexity linear with respect to image size. Compared with standard multi-head self-attention, the shifted window approach enhances the model's ability to capture long-range dependencies in CWT images, which is critical for identifying bearing fault patterns distributed across both time and frequency domains.

In summary, the hierarchical multiscale representation, systematically optimized patch and window sizes, and shifted window attention mechanism collectively provide a theoretically grounded and empirically validated rationale for employing Swin Transformer in CWT-based bearing fault diagnosis.

Model	Accuracy	Precision	Recall	F1	Params	FLOPs	Overall Accuracy
Swin	99.33%	0.9938	0.9933	0.9951	86.75 M	15.19GMac	99.83%
ViT	96.00%	0.9611	0.9600	0.9592	85.81 M	12.02GMac	96.00%
ConViT	93.33%	0.9543	0.9333	0.9310	85.78 M	16.83GMac	98.67%
DeiT-S	96.83%	0.9684	0.9683	0.9681	85.82 M	12.08GMac	99.17%
Mixer-B	98.33%	0.9843	0.9833	0.9832	59.12 M	12.64GMac	99.33%

Table 8. Comparison of Swin transformer with alternative architectures.

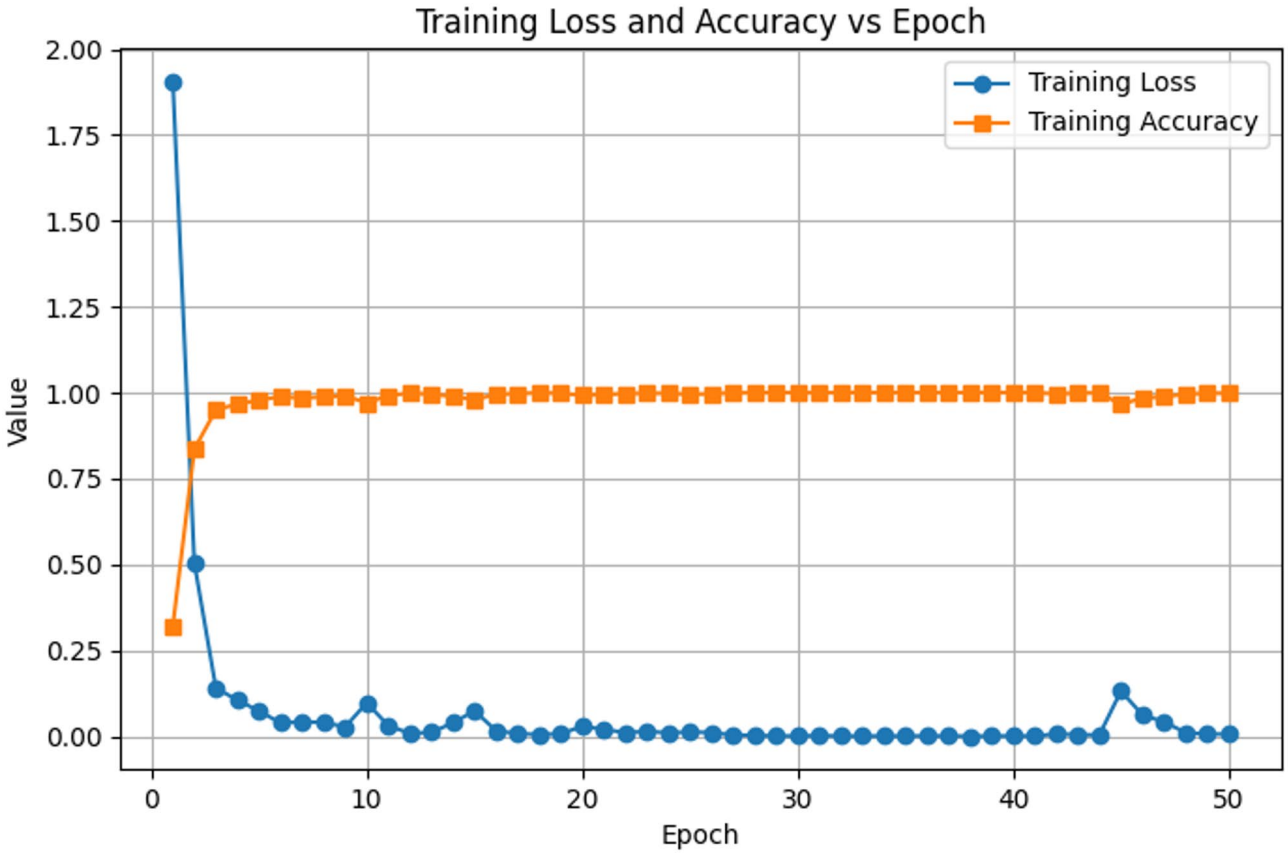


Fig. 8. Training loss and accuracy curves of the proposed model on the CWRU dataset.

Ablation study

Validation based on the CWRU dataset shows the training loss and accuracy curves as depicted Fig. 8.

The figure reveals that training loss rapidly decreases and approaches zero within the first few epochs, while training accuracy increases sharply and stabilizes near 100%. This indicates that the proposed model achieves effective convergence within a limited number of iterations, demonstrating strong fitting and convergence capabilities.

Furthermore, Under the conditions of the publicly available CWRU dataset, the proposed model ultimately achieved an accuracy of 99.83%, as illustrated in Fig. 9.

To further assess the robustness and reliability of the model, a 5-fold cross-validation was conducted, in which the dataset was randomly partitioned into training (80%) and validation (20%) subsets in each fold, ensuring no sample overlap between folds. The performance metrics of Accuracy, Precision, Recall, and F1-score were computed for each fold, and the mean and standard deviation across folds are presented in Table 9. These results demonstrate the model's stability and reproducibility.

To verify the superiority of the proposed parallel dual-channel multimodal fusion diagnostic model, the CWRU dataset was separately input into CNN, CNN-BiGRU, CNN-BiGRU-Attention, and Swin Transformer models for training. The comparative results are presented in Fig. 10.

The comparative results of the ablation study are presented in Fig. 11, which combines line and bar charts to clearly illustrate the performance improvements contributed by each module.

The results demonstrate that each module contributes significantly and positively to the overall model performance. The baseline CNN model achieves only 82.00% accuracy on the test set, indicating certain limitations

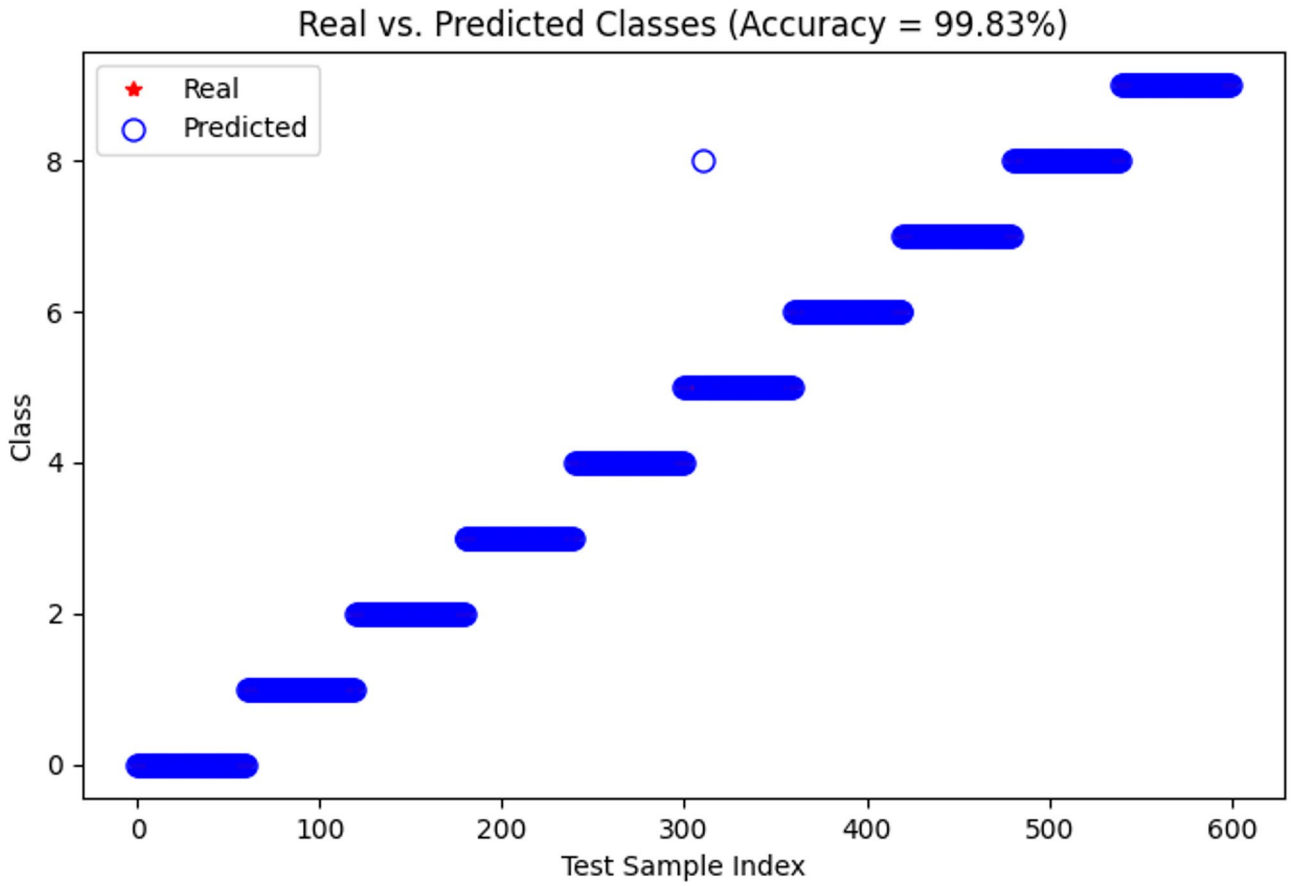


Fig. 9. Experimental results on the Public CWRU dataset.

Fold	Accuracy	Precision	Recall	F1-score
1	0.9964	0.9966	0.9964	0.9964
2	0.9929	0.9931	0.9929	0.9929
3	0.9964	0.9966	0.9964	0.9964
4	1.0000	1.0000	1.0000	1.0000
5	0.9929	0.9930	0.9929	0.9929
Mean ± Std	0.9957 ± 0.0027	0.9958 ± 0.0027	0.9957 ± 0.0027	0.9957 ± 0.0027

Table 9. Performance of the proposed model under 5-fold cross-validation on CWRU.

in extracting temporal features. Incorporating BiGRU improves the accuracy to 87.50%, reflecting enhanced capability in capturing temporal dependencies and enriching feature representation. Further introduction of the attention mechanism leads to a substantial increase in accuracy to 97.50%, validating its effectiveness in emphasizing critical features while suppressing redundant information. When used independently, the Swin Transformer module attains 99.33% accuracy, demonstrating its superior global modeling capability for time-frequency images. Ultimately, the integrated model combining Swin Transformer with CNN-BiGRU-Attention achieves an accuracy of 99.83%, realizing optimal performance through the synergistic effect of multiple submodules. These results thoroughly confirm the vital role of each key component in improving fault diagnosis accuracy and highlight the superiority of the proposed model architecture in feature extraction and fusion.

To quantitatively evaluate the effectiveness of MCB, we conducted comparative experiments using four alternative fusion strategies: Feature-level Concatenation (FLC), Decision-level Fusion (DLF), Learnable Fusion Weights (LFW), and Cross-Attention Fusion (CAF). Feature-level concatenation is a first-order fusion approach, which simply stacks features without modeling inter-modal interactions. Decision-level fusion averages the output probabilities of each branch. Learnable fusion weights introduce trainable coefficients to balance contributions from each modality, while cross-attention mechanisms model pairwise dependencies but in a parametric attention space.

The performance of these fusion methods was evaluated on the CWRU bearing dataset, and the key metrics including accuracy, precision, recall, and F1-score are summarized in Fig. 12.

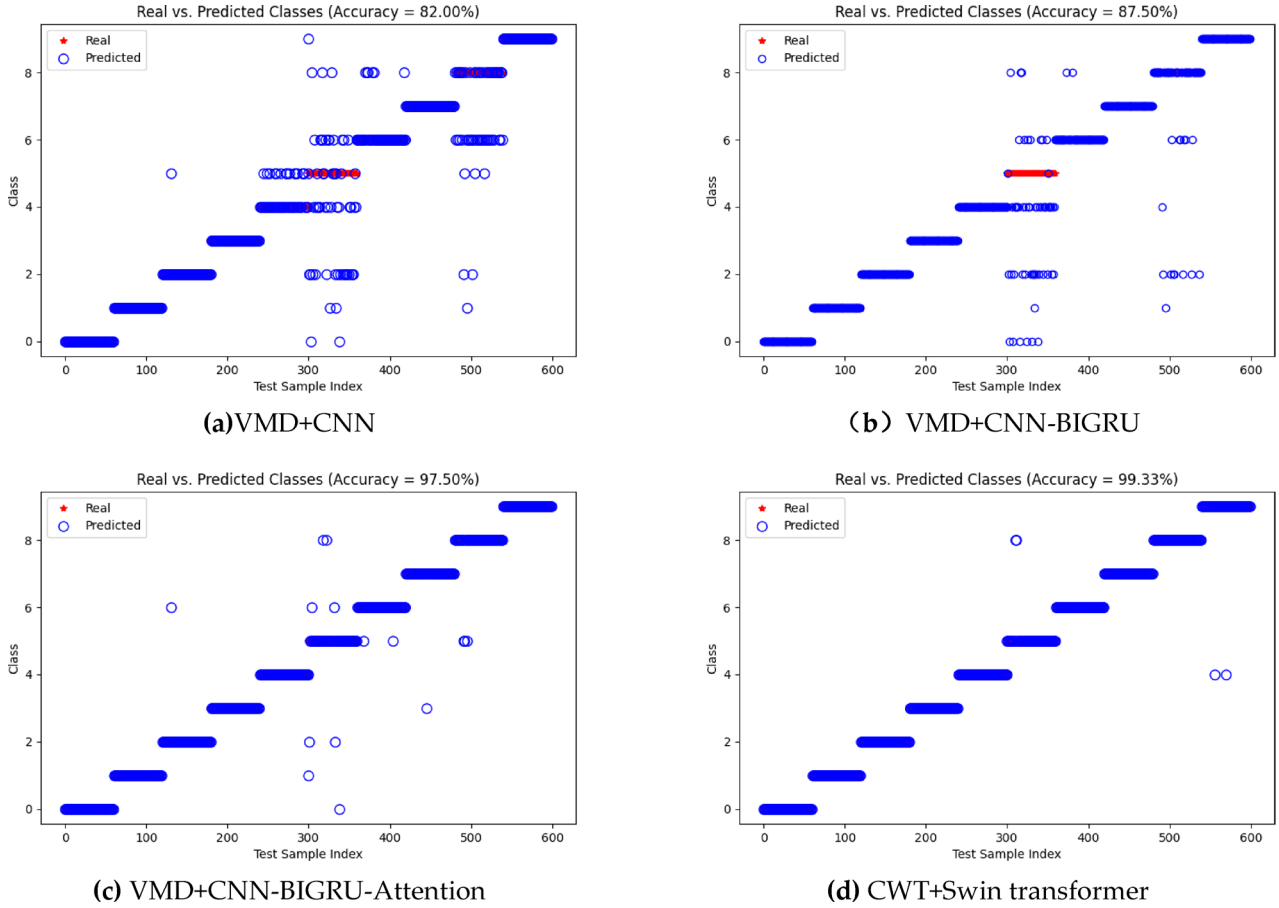


Fig. 10. Ablation study results on the CWRU dataset.

From the results, it is evident that MCB outperforms all other fusion strategies, demonstrating the benefit of capturing second-order interactions between modalities. In contrast, simpler first-order approaches (e.g., FLC) achieve slightly lower performance, while decision-level fusion and cross-attention underperform, likely due to insufficient inter-modal correlation modeling or limited training data.

Potential overfitting risks introduced by MCB's high-dimensional feature space are mitigated by several measures: (1) randomized projections via CountSketch reduce the effective dimensionality while preserving interactions; (2) batch normalization and dropout in the downstream classifier; (3) early stopping during training. No significant overfitting was observed, as evidenced by the comparable training and testing accuracies.

Overall, this study demonstrates that MCB not only preserves the discriminative information from both modalities but also enhances fault classification by modeling higher-order interactions, which is critical for accurate and robust bearing fault diagnosis.

Generalization experiment

To further evaluate the effectiveness and generalization capability of the proposed model, this study also utilizes the Southeast University bearing dataset, which includes mixed fault types. The experimental platform setup is illustrated in Fig. 13.

The Southeast University bearing dataset includes five fault categories: ball fault, inner ring fault, outer ring fault, compound fault (combined faults on both inner and outer rings), and healthy operating state. Each fault category corresponds to two working conditions: 20 Hz (1200 rpm) at 0 V load (0 Nm) and 30 Hz (1800 rpm) at 2 V load (7.32 Nm), resulting in a total of ten fault types. The dataset is divided into training and testing sets with a 70:30 ratio. Detailed dataset information is presented in Table 10.

The experimental results are presented in Fig. 14. Despite the increased complexity and difficulty in fault identification due to the inclusion of compound faults in the Southeast University dataset, the proposed model still achieved a high accuracy of 98.33%, demonstrating its excellent generalization capability and robustness.

To further validate the superiority of the proposed SCBM-Net model, the Southeast University dataset was separately fed into each individual model for training. The comparative test accuracy results of different models on the two datasets are presented in Table 11.

The comparison in Table 11 clearly shows that the proposed dual-channel model outperforms the single models and single-channel architectures in terms of accuracy, achieving an average accuracy above 99%. Therefore, the SCBM-Net dual-channel model proposed in this paper demonstrates a distinct advantage.

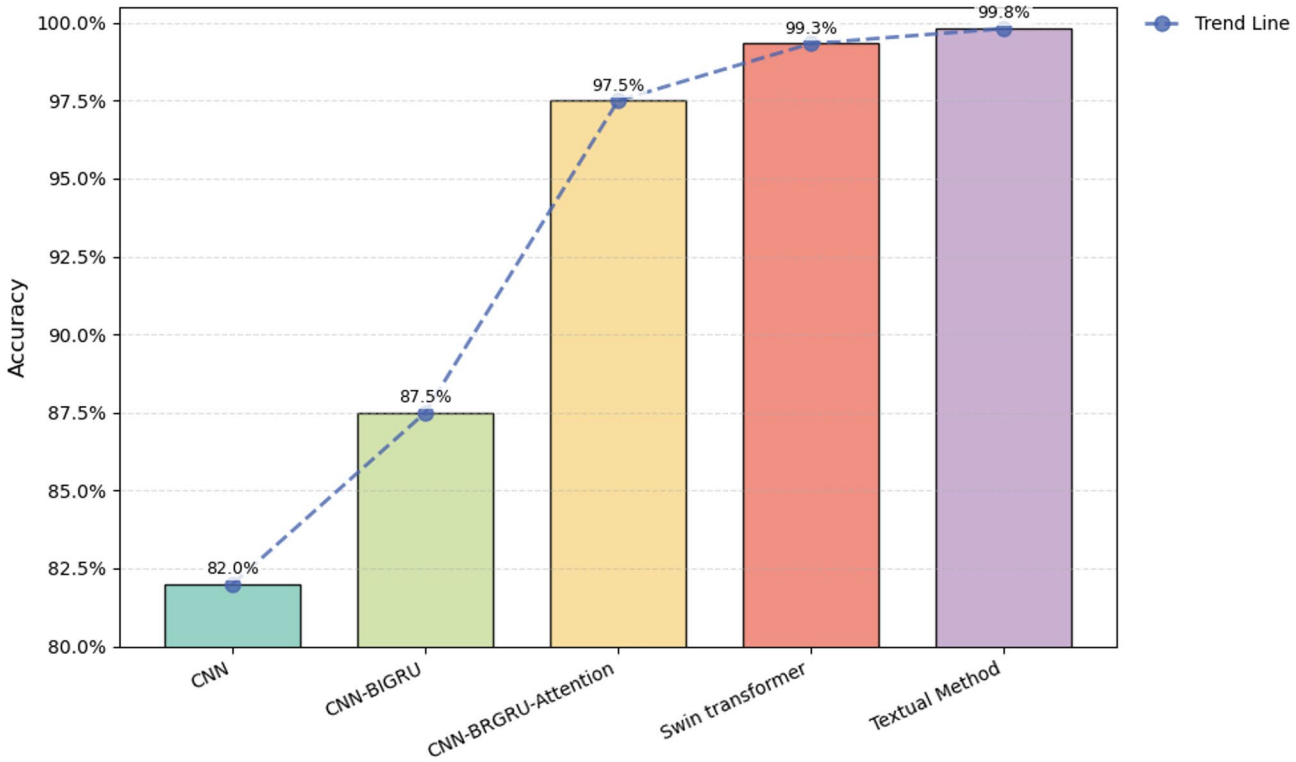


Fig. 11. Ablation study comparative results.

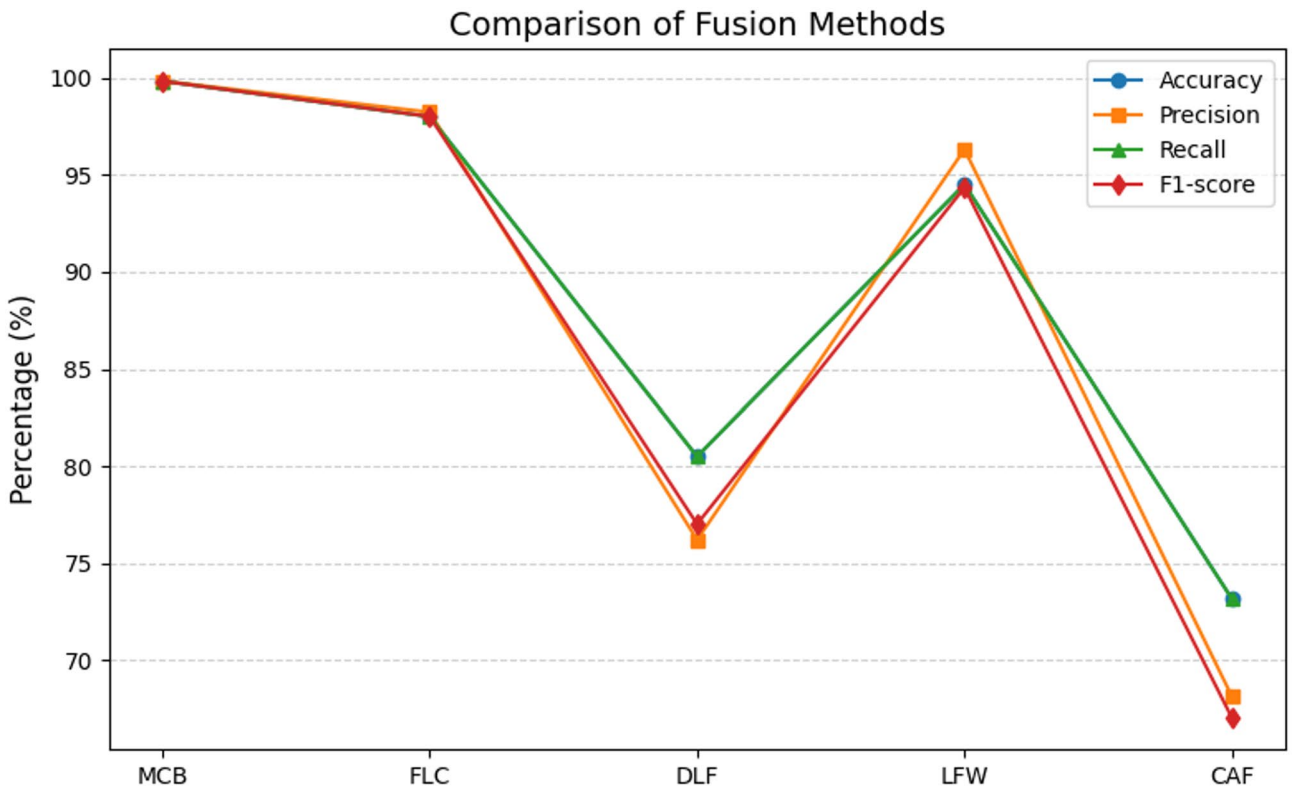


Fig. 12. Comparison of fusion methods (line chart).

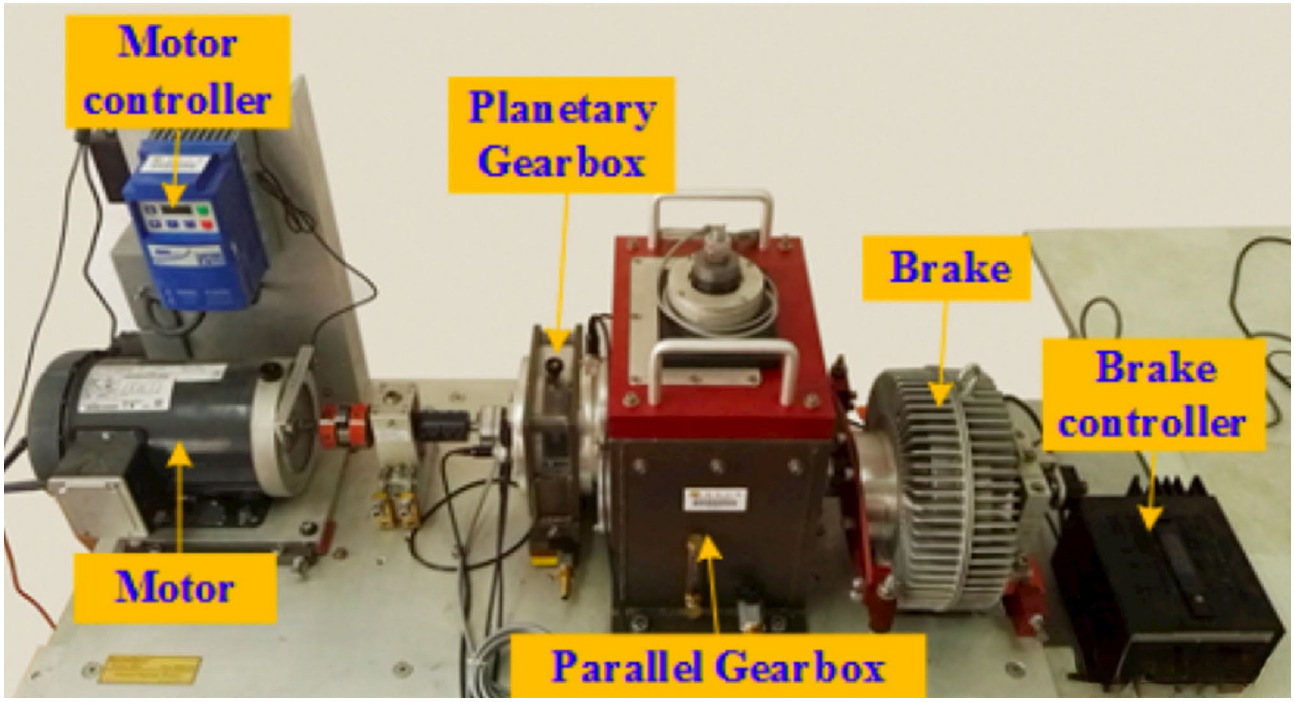


Fig. 13. Bearing test rig structure at Southeast University.

Fault type	Label	Speed/load	Training set	Test set
Ball_0	0	20/0	210	90
Ball_1	1	30/2	210	90
Comb_0	2	20/0	210	90
Comb_1	3	30/2	210	90
Health_0	4	20/0	210	90
Health_1	5	30/2	210	90
Inner_0	6	20/0	210	90
Inner_1	7	30/2	210	90
Outer_0	8	20/0	210	90
Outer_1	9	30/2	210	90

Table 10. Publicly available bearing dataset from Southeast University.

Few-shot experiment

To further evaluate the recognition capability of SCBM-Net under limited data conditions, a few-shot experiment was conducted. Specifically, the experiment was based on the CWRU bearing fault dataset. For each fault category, only 60 samples were randomly selected for training, while the remaining 140 samples were used for testing. The detailed sample distribution is shown in Table 12. This setting simulates practical industrial scenarios where fault data is scarce. All other experimental configurations were kept consistent with the full-sample setting to ensure fair comparability.

To better illustrate the classification effectiveness of the SCBM-Net model, the t-Distributed Stochastic Neighbor Embedding (t-SNE) method is employed to visualize the feature distributions on the test set. Specifically, the original data distribution, the classification results of the two single-channel branches, and the final classification results of SCBM-Net are presented separately, as shown in Fig. 15.

As shown in Fig. 15(a), the original data distribution is chaotic and disordered, lacking any clear clustering structure. After feature extraction through the VMD channel (Fig. 15(b)) and the CWT channel (Fig. 15(c)), the samples exhibit a certain degree of clustering tendency and class separation, but there is still considerable overlap between different classes, resulting in limited classification performance. In contrast, Fig. 15(d) shows that the SCBM-Net model, after fusing time-frequency features, significantly improves the class discrimination ability. The samples are distributed more compactly and with clearer boundaries in the feature space, successfully achieving effective differentiation of 10 types of bearing faults. Under small sample conditions, the model can still reach a classification accuracy of 98.64%, fully demonstrating its superior feature representation capability and fault diagnosis performance.

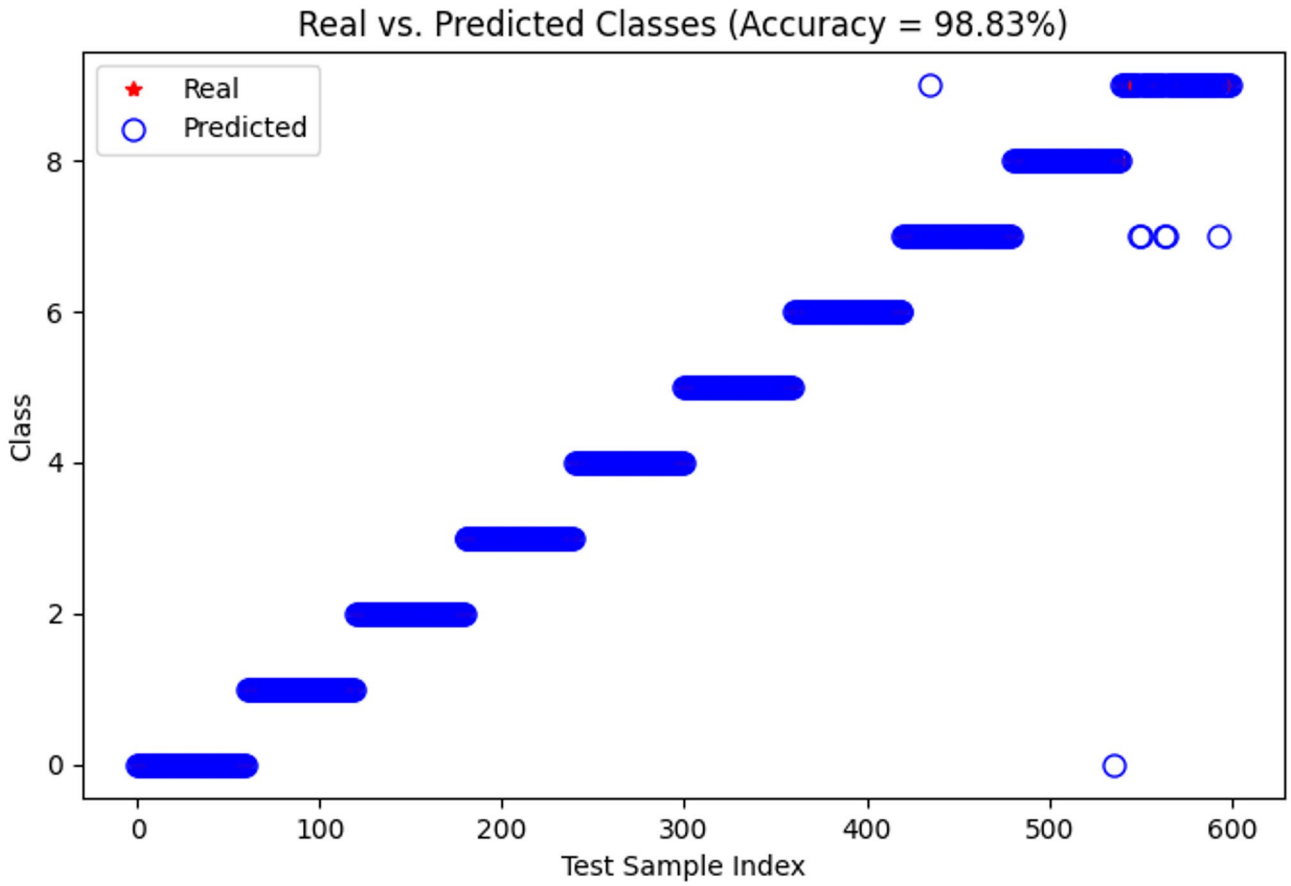


Fig. 14. Experimental results based on the Southeast University bearing dataset.

Model	Dataset		Accuracy (%)	Average accuracy (%)
	CWRU	SEU		
VMD + CNN	√		82.00%	56.78%
		√	31.56%	
VMD + CNN-BiGRU	√		87.50%	66.31%
		√	45.11%	
VMD + CNN-BiGRU-Attention	√		97.50%	76.42%
		√	55.33%	
CWT + Swin transformer	√		99.33%	98.50%
		√	97.67%	
SCBM-Net	√		99.83%	99.08%
		√	98.33%	

Table 11. Experimental accuracy comparison.

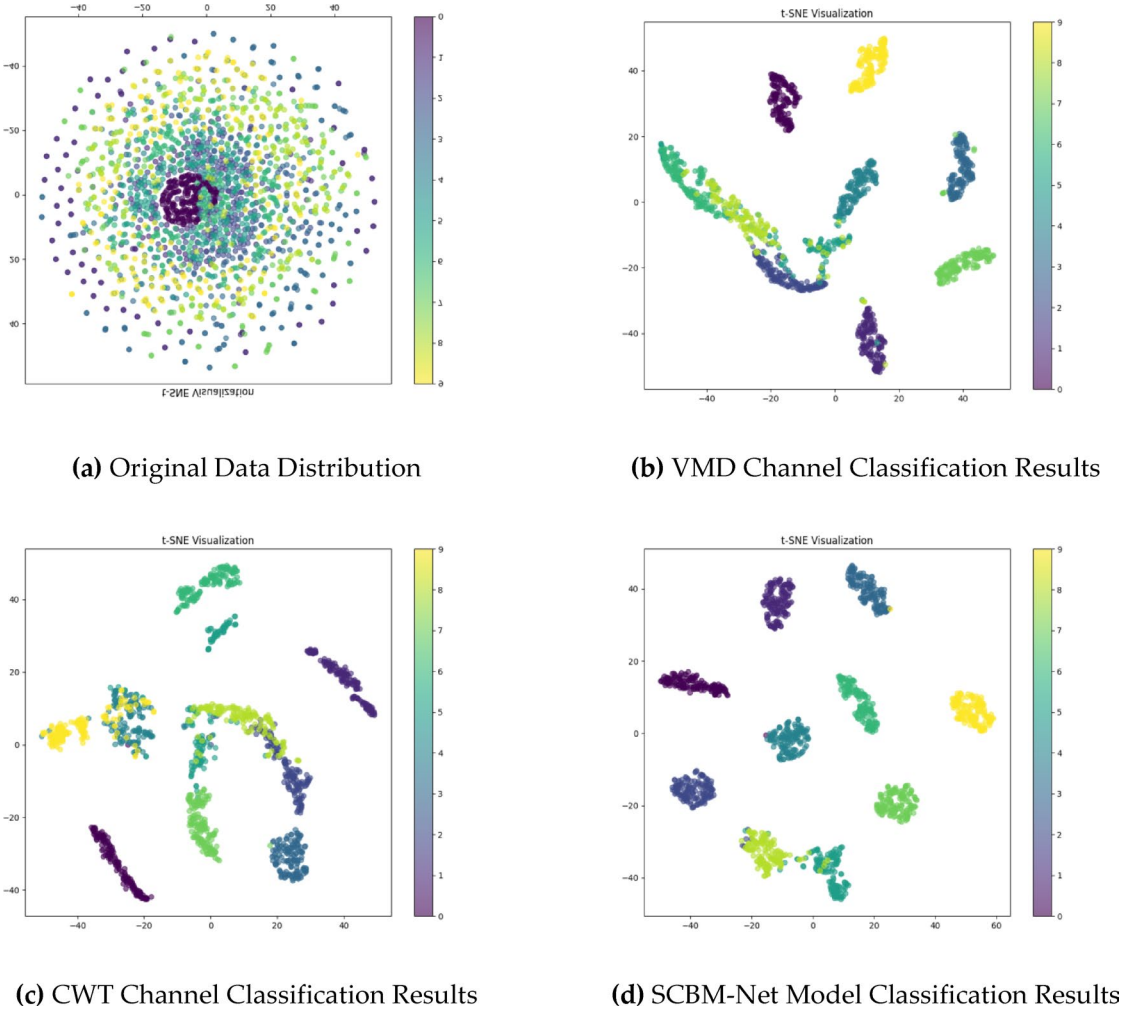
Unbalanced dataset experiment

To further verify the robustness and generalization capability of the proposed SCBM-Net model under practical complex working conditions, a comparative experiment based on an unbalanced dataset was designed and conducted. In industrial equipment operation, the distribution of fault categories typically exhibits significant imbalance, where some fault types occur less frequently and thus constitute minority classes. It is necessary to evaluate the model's ability to recognize various fault types under such imbalanced data distributions.

The experimental data is constructed based on the CWRU bearing fault dataset. To simulate an imbalanced scenario, four major fault categories (Normal, IR007, IR014, IR021) each retain 200 samples, with 140 samples for training and 60 for testing; the remaining six minor categories (B007, OR007, B014, OR014, B021, OR021) are each limited to 45 samples, divided into 30 for training and 15 for testing. The detailed composition of this unbalanced dataset is shown in Table 13.

Fault type	Label	Damage diameter (mil)	Training set	Test set
N	0	–	60	140
IR007	1	7	60	140
B007	2	7	60	140
OR007	3	7	60	140
IR014	4	14	60	140
B014	5	14	60	140
OR014	6	14	60	140
IR021	7	21	60	140
B021	8	21	60	140
0R021	9	21	60	140

Table 12. Sample distribution of the few-shot dataset based on CWRU.



(a) Original Data Distribution

(b) VMD Channel Classification Results

(c) CWT Channel Classification Results

(d) SCBM-Net Model Classification Results

Fig. 15. T-SNE visualization of classification results.

The SCBM-Net model was trained and tested on the aforementioned imbalanced dataset while retaining the same training parameter settings. The confusion matrix on the test set is shown in Fig. 16.

Based on Fig. 16, the proposed model achieves 100% recognition accuracy on majority classes such as Normal, IR007, IR014, and IR021. It also correctly classifies minority classes including OR007, B014, OR014, B021, and OR021 without errors, demonstrating stable recognition capability for these less frequent fault types. For class B007, the model correctly classifies 11 samples but misclassifies 4 samples as B021, resulting in an accuracy of 73.3%, which is slightly lower than other categories. This indicates some confusion between feature-similar classes when sample size is relatively small.

Fault type	Label	Damage diameter (mil)	Training set	Test set
N	0	–	140	60
IR007	1	7	140	60
B007	2	7	30	15
OR007	3	7	30	15
IR014	4	14	140	60
B014	5	14	30	15
OR014	6	14	30	15
IR021	7	21	140	60
B021	8	21	30	15
0R021	9	21	30	15

Table 13. Based on the imbalanced CWRU dataset.

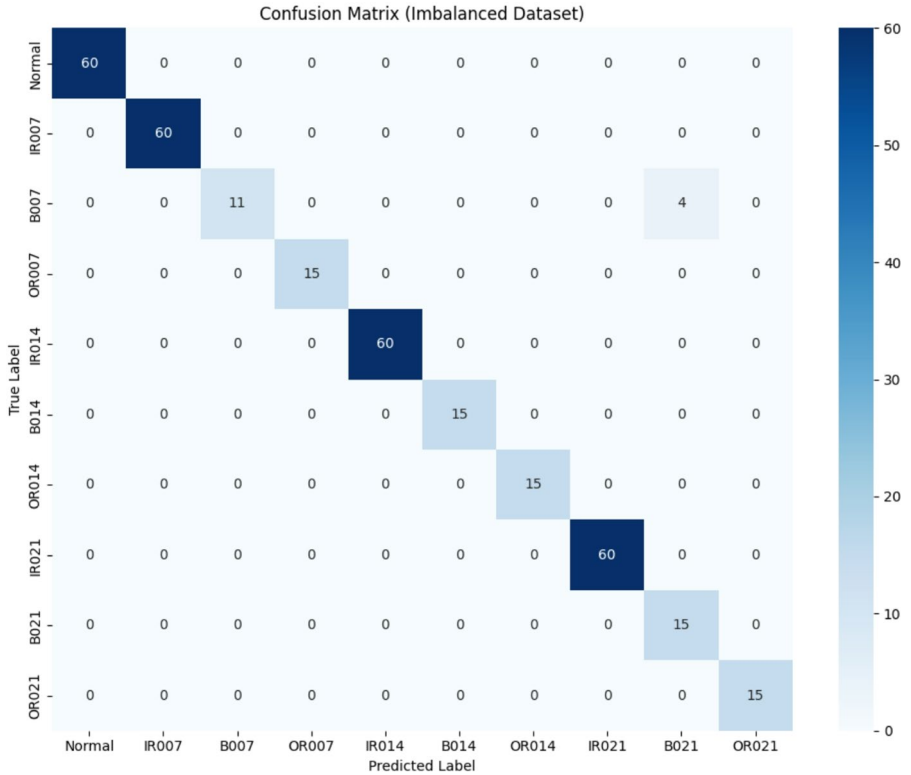


Fig. 16. Confusion matrix of test results based on the imbalanced dataset.

To further comprehensively evaluate the diagnostic performance of the proposed method, we incorporated additional metrics beyond accuracy. The proposed model achieves a precision of 0.9715, a recall of 0.9667, an F1-score of 0.9662. Furthermore, Fig. 17 illustrates the ROC curves of the proposed model across different fault categories. The macro-average AUC of 0.9992 highlights the strong discriminative capability and stability of SCBM-Net in distinguishing among various fault types.

Overall, SCBM-Net maintains strong discriminative ability under severely imbalanced class distributions, achieving an average accuracy of 97.33% across all 10 fault categories. The model effectively extracts complementary multimodal features and accurately classifies faults despite class imbalance, highlighting its enhanced capability to capture both time-frequency and temporal characteristics. This further verifies the model's robustness and applicability in complex real-world operating conditions.

Noise experiment

In practical engineering scenarios, the acquisition of rolling bearing vibration signals is inevitably affected by multi-source interference noise, including mechanical structure resonance, electromagnetic coupling effects, and other complex environmental factors. Adding noise to the original signals can create industrial condition datasets with high realism, thereby enhancing the generalization ability and engineering applicability of fault diagnosis models.

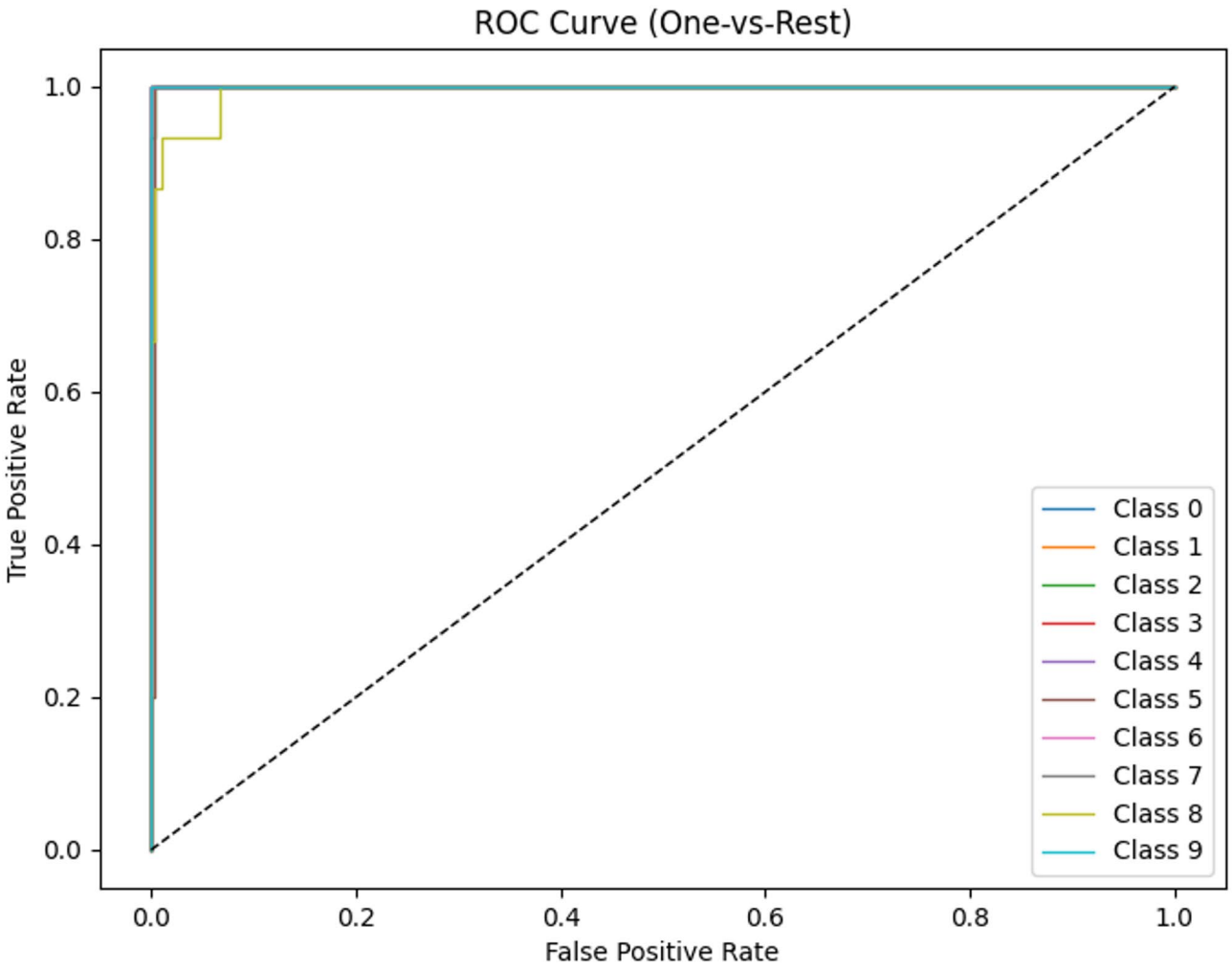


Fig. 17. ROC curves of the proposed model for different fault classes.

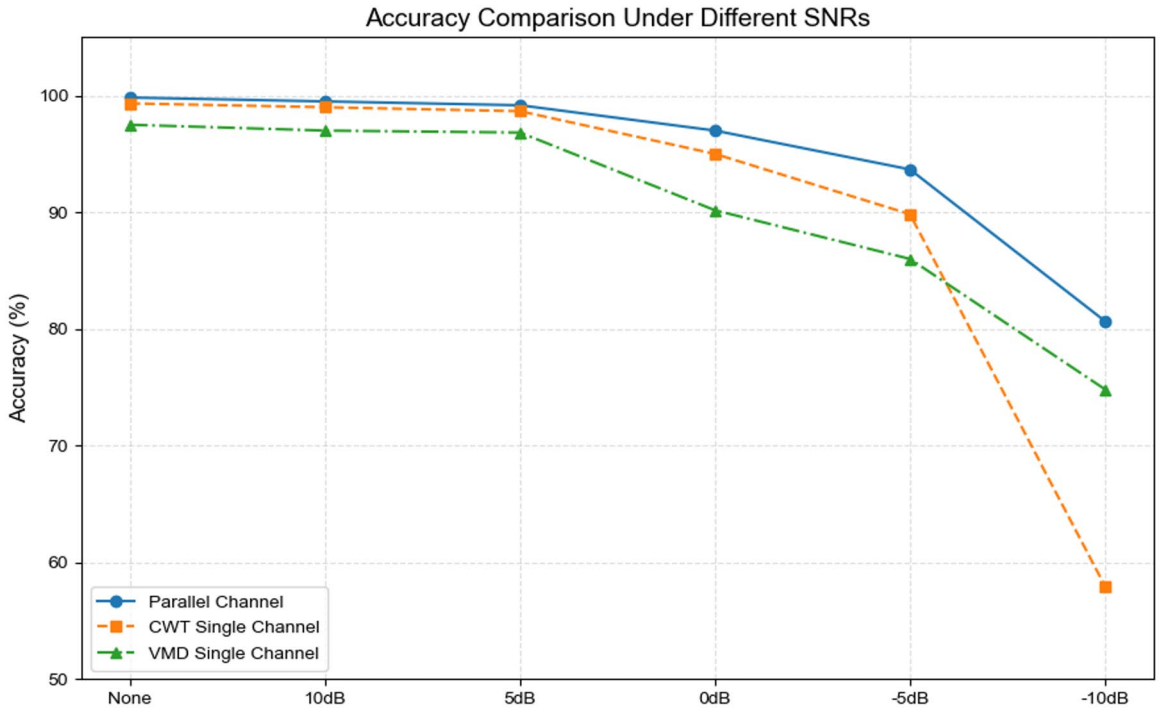
Gaussian noise, as a typical additive noise model, follows a zero-mean normal distribution and can effectively characterize random fluctuations caused by typical interference sources such as electronic device thermal noise and atmospheric background radiation. The introduction of Gaussian noise can realistically simulate the complex noise environment encountered in engineering practice, helping to evaluate and optimize the model's performance.

In this experiment, samples of length 1024 were selected from the CWRU dataset, and Gaussian noise with signal-to-noise ratios (SNR) ranging from − 10 dB to 10 dB was added. The results are shown in Fig. 18.

As shown in Fig. 18, the performance of all three methods degrades to varying degrees with increasing noise; however, the parallel dual-channel model consistently maintains the highest accuracy, demonstrating the advantage of multimodal fusion in noise robustness. Under noise-free conditions, all methods achieve over 95% accuracy. When the signal-to-noise ratio (SNR) drops to 0 dB, the parallel model still achieves approximately 97% accuracy, while the single-channel models for CWT and VMD channels decline to about 95% and 90.17%, respectively. At − 5 dB SNR, the fusion model's accuracy remains above 93%, whereas the CWT and VMD channels fall below 90%. Under extreme noise conditions (− 10 dB), the fusion model achieves about 80.67% accuracy, significantly outperforming the CWT channel (58%) and VMD channel (74.83%). This indicates that multimodal fusion not only enhances the overall classification performance but also improves robustness against severe noise interference. These results validate the effectiveness of the proposed model's parallel extraction and fusion of time–frequency image features and time-series signal features.

To comprehensively evaluate the proposed dual-channel fault diagnosis method, we compare it with several baseline and state-of-the-art approaches under various noise conditions.

The ACNN-LFSwinT is a dual-channel framework that utilizes both one-dimensional vibration signals and two-dimensional time-frequency images. In one channel, intrinsic mode functions (IMFs) extracted via CEEMDAN are processed by an attention-based CNN (ACNN) for feature extraction. In the other channel, time-frequency images from STFT are input into a Local Feature Swin Transformer (LFSwin Transformer) to capture spatial features. Features from both channels are concatenated for classification, enabling robust multimodal fault diagnosis.The CNN-BiLSTM combines convolutional neural networks (CNNs) and bidirectional long short-



	None	10dB	5dB	0dB	-5dB	-10dB
Parallel Channel	99.83%	99.50%	99.17%	97.00%	93.67%	80.67%
CWT Single Channel	99.33%	99.00%	98.66%	95.00%	89.83%	58.00%
VMD Single Channel	97.50%	97.00%	96.83%	90.17%	86.00%	74.83%

Fig. 18. Diagnostic performance under various noise conditions.

term memory networks (BiLSTM) to capture both local spatial features and long-term temporal dependencies from vibration signals. This hybrid structure enhances the recognition of fault patterns across both spatial and temporal domains. The Multi-Scale Dilated CNN + Self-Attention (MSDCSA) employs dilated convolutions at multiple scales to enlarge the receptive field, extracting features from both short-term and long-term signal patterns, while a self-attention mechanism emphasizes the most informative features for classification. The TST (Time Series Transformer) leverages the transformer architecture to process raw vibration sequences, capturing global temporal dependencies via self-attention, providing an effective framework for sequential modeling of vibration signals. Finally, the feature-based machine learning (Feature-based ML) approach extracts classical time-domain and frequency-domain features—including mean, RMS, skewness, kurtosis, spectral RMS, and spectral kurtosis—which are then fed into an MLP classifier as a baseline for comparison with deep learning and multimodal approaches.

Additionally, we include two recently published multimodal fault diagnosis methods from 2022 to 2024 to provide a state-of-the-art comparison. MLF-MSI[28] (A Multi-Level Fusion Framework for Bearing Fault Diagnosis Using Multi-Source Information) employs a multi-level fusion strategy to integrate features from multiple information sources for robust fault classification. MS-ResidualNet[29] (Multi-Source Information-Based Bearing Fault Diagnosis Using Multi-Branch Selective Fusion Deep Residual Network) utilizes multi-branch residual networks to combine various signal modalities, enabling effective and accurate fault diagnosis under different operating conditions.

The models were evaluated under different Gaussian noise levels, with signal-to-noise ratios (SNR) ranging from −10 dB to 10 dB. The comparison results, summarized in Table 14 and illustrated in Fig. 19, demonstrate the robustness of the proposed method relative to other approaches. As shown, the proposed dual-channel method achieves the highest accuracy across all SNR levels, outperforming ACNN-LFSwinT, CNN-BiLSTM, MSDCSA, TST, Feature-based ML, MLF-MSI, and MS-ResidualNet under both high- and low-SNR conditions. To ensure a fair comparison, all baseline methods were re-implemented using the same dataset, preprocessing procedures, and consistent hyperparameter settings.

Conclusion

This study presents SCBM-Net, a novel dual-channel multimodal fusion model designed for robust bearing fault diagnosis under complex operating conditions. By leveraging the complementary characteristics of time–frequency representations and temporal sequences, SCBM-Net integrates Continuous Wavelet Transform

Method	None	10 dB	5 dB	0 dB	− 5 dB	− 10 dB
Textual Method	99.83%	99.50%	99.17%	97.00%	93.67%	80.67%
ACNN-LFSwinT	97.10%	93.30%	91.50%	88.90%	85.20%	79.70%
CNN-BILSTM	96.46%	92.84%	90.23%	77.34%	73.33%	70.77%
MLF-MSI	99.33%	99.00%	95.50%	93.30%	87.23%	75.17%
MS-Residual	98.67%	97.50%	96.67%	94.83%	82.17%	69.50%
MSDCSA	94.50%	88.50%	85.00%	79.17%	76.67%	70.83%
TST	90.00%	88.17%	85.33%	82.67%	75.83%	75.00%
Feat-ML	95.67%	93.67%	92.83%	89.00%	81.33%	75.33%

Table 14. Accuracy of fault diagnosis methods under different SNR levels (%).

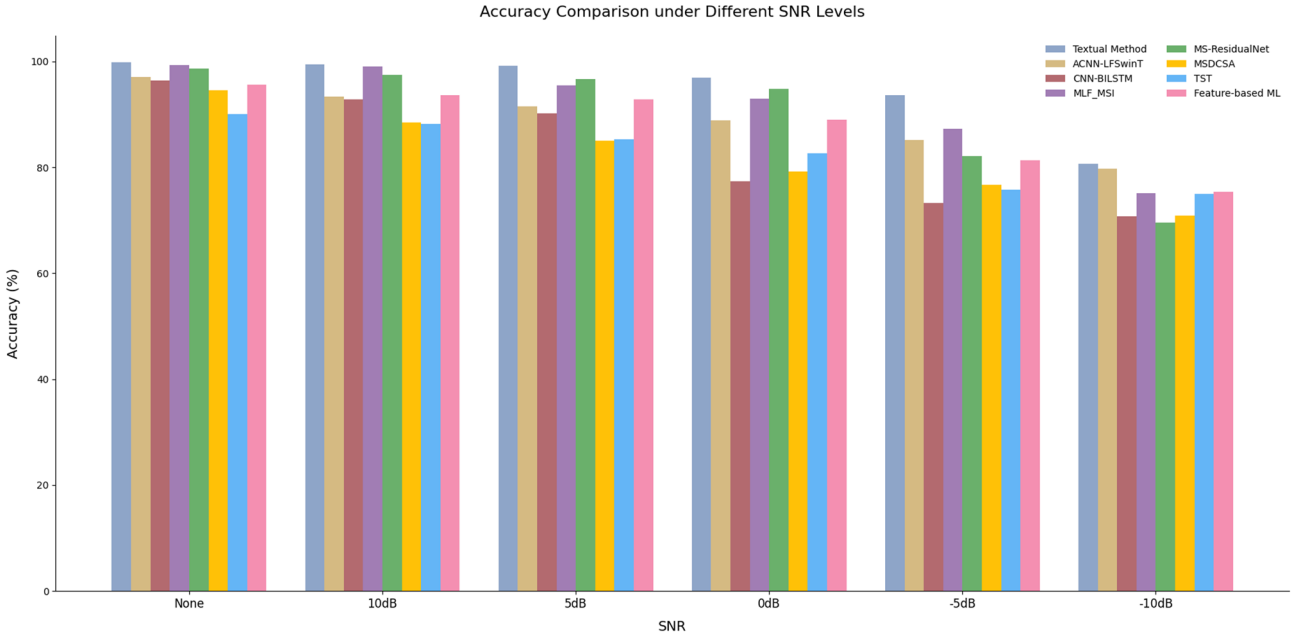


Fig. 19. Comparison of the proposed model with other fault diagnosis models.

(CWT) and Variational Mode Decomposition (VMD) into a unified architecture. Specifically, a Swin Transformer is employed in the image channel to extract both local and global features from CWT-based time–frequency images, while a CNN-BiGRU-Attention network captures temporal dependencies from the intrinsic mode functions generated by VMD in the sequential channel. These heterogeneous features are effectively fused via a Multimodal Compact Bilinear (MCB) pooling module, substantially enhancing the model's feature representation capability.

Extensive experiments conducted on the CWRU and Southeast University datasets validate the proposed model's superior diagnostic accuracy, generalization performance, and robustness to noise. SCBM-Net achieves 99.83% accuracy under noiseless conditions, 98.33% in cross-domain evaluations, and maintains 80.67% accuracy at a low signal-to-noise ratio of − 10 dB. Additionally, it demonstrates strong performance on imbalanced datasets, accurately identifying both majority and minority fault classes. Ablation studies further confirm the contribution of each core module to the overall performance gains. Furthermore, under small-sample conditions, SCBM-Net still achieves a high classification accuracy of 98.64%, highlighting its excellent feature representation capability and adaptability in data-scarce scenarios.

Data availability

The data used to support the findings of this study are available from the corresponding author upon request.

Received: 2 June 2025; Accepted: 23 September 2025
Published online: 29 October 2025

References

1. Xu, T., Ji, J., Kong, X., Zou, F. & Wang, W. Bearing Fault Diagnosis in the Mixed Domain Based on Crossover-Mutation Chaotic Particle Swarm. *Complexity* 6632187 (2021). (2021).

2. Yao, J., Zhao, J., Deng, Y. & Langari, R. Weak fault feature extraction of rotating machinery based on Double-Window spectrum fusion enhancement. *IEEE Trans. Instrum. Meas.* **69**, 1029–1040 (2020).
3. Fault Diagnosis of Rolling Bearing Based on Fractional Fourier Instantaneous Spectrum | Experimental Techniques. https://link.springer.com/article/10.1007/s40799-021-00478-w?utm_source=xmol&utm_medium=affiliate&utm_content=meta&utm_campaign=DDCN_1_GL01_metadata.
4. Boudiaf, R., Abdelkarim, B. & Issam, H. Bearing fault diagnosis in induction motor using continuous wavelet transform and convolutional neural networks. *Int. J. Power Electron. Drive Syst. (IJPEDS)*. **15**, 591–602 (2024).
5. Research on an early warning method for bearing health diagnosis based on EEMD-PCA-ANFIS | Electrical Engineering. https://link.springer.com/article/10.1007/s00202-023-01821-7?utm_source=xmol&utm_medium=affiliate&utm_content=meta&utm_campaign=DDCN_1_GL01_metadata.
6. Guo, L., Gu, X., Yu, Y., Duan, A. & Gao, H. An analysis method for interpretability of convolutional neural network in bearing fault diagnosis. *IEEE Trans. Instrum. Meas.* **73**, 1–12 (2024).
7. Ince, T., Kiranyaz, S., Eren, L., Askar, M. & Gabbouj, M. Real-Time motor fault detection by 1-D convolutional neural networks. *IEEE Trans. Industr. Electron.* **63**, 7067–7075 (2016).
8. Convolutional Neural Network Based Fault Detection for Rotating Machinery. - ScienceDirect. https://www.sciencedirect.com/science/article/abs/pii/S0022460X16301638?via%3Dihub
9. Jia, L., Chow, T. W. S., Wang, Y. & Yuan, Y. Multiscale residual attention convolutional neural network for bearing fault diagnosis. *IEEE Trans. Instrum. Meas.* **71**, 1–13 (2022).
10. Chen, X., Yang, Y., Cui, Z. & Shen, J. Wavelet denoising for the vibration signals of wind turbines based on variational mode decomposition and multiscale permutation entropy. *IEEE Access.* **8**, 40347–40356 (2020).
11. Research on denoising of. Second harmonic signal in photoacoustic spectroscopy based on SSA-VMD-WTD method. *Infrared Phys. Technol.* **138**, 105204 (2024).
12. Ma, Z. & Zhang, Y. A study on rolling bearing fault diagnosis using RIME-VMD. *Scientific Reports* **15**, (2025).
13. Wang, X. et al. Fault diagnosis method of rolling bearing based on SSA-VMD and RCMDE. *Scientific Reports* **14**, (2024).
14. Rotating machinery fault diagnosis based on one-dimensional convolutional neural network and modified multi-scale graph convolutional network under limited labeled data. Engineering Applications of Artificial Intelligence 137, 109129. (2024).
15. Cross-domain fault. Diagnosis of bearing using improved semi-supervised meta-learning towards interference of out-of-distribution samples. *Knowl. Based Syst.* **252**, 109493 (2022).
16. Liu, Z. et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. Preprint at (2021). https://doi.org/10.48550/arXiv.2103.14030
17. Tang, X., Xu, Z. & Wang, Z. A. Novel fault diagnosis method of rolling bearing based on integrated vision transformer model. *Sensors* **22**, 3878 (2022).
18. Ji, M. & Zhao, G. D. E. V. T. Deformable Convolution-Based vision transformer for bearing fault diagnosis. *IEEE Trans. Instrum. Meas.* **73**, 1–13 (2024).
19. MARNet, Wang, G., Zhao, C. & Zhang, Y. Multi-head attention residual network for rolling bearing fault diagnosis under noisy condition - Linfeng Deng, (2024). https://journals.sagepub.com/doi/10.1177/09544062241259614
20. Fault diagnosis for. Small samples based on attention mechanism. *Measurement* **187**, 110242 (2022).
21. A Bearing Fault Diagnosis Method Based on Dilated Convolution and Multi. -Head Self-Attention Mechanism. https://www.mdpi.com/2076-3417/13/23/12770
22. Jawadekar, A., Paraskar, S., Jadhav, S. & Dhole, G. Artificial neural network-based induction motor fault classifier using continuous wavelet transform. *Syst. Sci. Control Eng.* **2**, 684–690 (2014).
23. Deep Learning Aided Data-Driven Fault Diagnosis of Rotatory Machine. A Comprehensive Review. https://www.mdpi.com/1996-1073/14/16/5150
24. Dragomiretskiy, K. & Zosso, D. Variational mode decomposition. *IEEE Trans. Signal Process.* **62**, 531–544 (2014).
25. Chung, J., Gulcehre, C., Cho, K. & Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *Preprint at.* https://doi.org/10.48550/arXiv.1412.3555 (2014).
26. Lipton, Z. C., Berkowitz, J. & Elkan, C. A. Critical Review of Recurrent Neural Networks for Sequence Learning. Preprint at (2015). https://doi.org/10.48550/arXiv.1506.00019
27. Fukui, A. et al. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. Preprint at (2016). https://doi.org/10.48550/arXiv.1606.01847
28. Deng, X., Sun, Y., Li, L. & Peng, X. A. Multi-Level fusion framework for bearing fault diagnosis using Multi-Source information. *Processes* **13**, 2657 (2025).
29. Xiong, S., Zhang, L., Yang, Y., Zhou, H. & Zhang, L. Multi-Source Information-Based bearing fault diagnosis using Multi-Branch selective fusion deep residual network. *Sensors* **24**, 6581 (2024).

Author contributions

Qiang Liu conceived and designed the study, developed the methodology, implemented the model, conducted the experiments, and drafted the manuscript; Weiyan Tong supervised the research, provided critical revisions, and was responsible for overall project coordination and funding acquisition; Hongwei Bai assisted with data preprocessing, experimental setup, and visualization; Shien Dong contributed to the analysis of the results and helped revise and refine the manuscript.

Funding

This study was supported by Liaoning Provincial Science and Technology Program Joint Project (2024-BSLH-210), China.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to W.T.