Article citation info:
Feng Z. A study on fault diagnosis for marine machinery bearings based on multimodal fusion and two-stage discrimination. Diagnostyka. 2026;27(2):2026202. https://doi.org/10.29354/diag/220723.

DIAGNOSTYKA, 2026, Vol. 27, No. 2

e-ISSN 2449-5220
DOI: 10.29354/diag/220723

A STUDY ON FAULT DIAGNOSIS FOR MARINE MACHINERY BEARINGS BASED ON MULTIMODAL FUSION AND TWO-STAGE DISCRIMINATION

Zhiyuan FENG

College of Engineering, Shanghai Ocean University, Shanghai 201306, China
Corresponding author, e-mail: f_zhiyuan@outlook.com

Abstract

Marine machinery operates in harsh environments where strong background noise often masks bearing fault features. Moreover, most current models lack sufficient feature extraction and adaptive detection capabilities under complex working conditions, resulting in poor classification accuracy and robustness. To address these issues, we propose a bearing fault diagnosis framework based on multimodal fusion and two-stage discrimination. The framework consists of two stages. The first stage is the feature enhancement module, which combines Variational Mode Decomposition (VMD) and Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) to perform secondary demixing on the raw vibration signals, and screen the components based on permutation entropy and kurtosis to filter background noise. Subsequently, the selected components are adaptively weighted and reconstructed, thereby achieving feature enhancement. The second stage is the fault discrimination module, which constructs a dual-channel Convolutional Neural Network (CNN) integrated with a Convolutional Block Attention Module (CBAM). In this module, Channel 1 inputs the reconstructed 1D time-series signals to extract deep features, while Channel 2 inputs 2D time-frequency images generated via the Hilbert marginal spectrum to extract spatial features. Then CBAM is utilized to adaptively focus on critical fault regions. Finally, the dual-channel heterogeneous features are flattened and fused to complete the fault discrimination. Two modules in our model are cascaded to form a "signal demixing-heterogeneous feature fusion" processing chain. The detection accuracy on the CWRU dataset reaches 99.11%, effectively resolving the issue of feature blurring in high-noise environments while exhibiting both high accuracy and strong robustness.

Keywords: multimodal fusion; two-stage discrimination; marine machinery; bearing fault diagnosis

1. INTRODUCTION

The publication cannot be shorter than six A4 pages. The article must be written in English.

The offshore wind turbine is a kind of marine machinery which plays an important role in developing resources and generating electricity. It works under rough and complicated environmental conditions all the time, constantly facing significant challenges such as intense salt and acid corrosion, serious rusting, enormous wind force and huge wave strength, and rapid currents. Such a demanding operating environment severely compromises the reliability of mechanical components. Particularly when failures occur in critical components such as bearings, it could be chain reaction, which would result in enormous damage, sudden stop. Moreover, it may lead to personnel casualties and ruin the ocean. So there must be study and improvement of technologies of such matters such as condition monitoring and fault diagnosis of the rotating shaft, which is called the heart of machines.

Various signal processing methods exist for extracting time-domain, frequency-domain, and time-frequency domain features of bearing faults. For instance, Lv et al. [1] employed the Ensemble Empirical Mode Decomposition (EEMD) method, facilitating fault diagnosis via spectrum analysis and real-time monitoring. However, under the marine operation conditions, it is often confronted with the problem of unstructured noise interference and signal aliasing, resulting in the loss of feature information. With the development of data driven approaches, traditional machine learning approaches like SVM and RF have been used in feature classification. Deep Learning models such as the Convolutional Neural Network (CNN) model, the Long short-term memory Network (LSTM) model and the recurrent neural network (RNN) model are applied in the gradual end-to-end bearing fault identification. For example, Liu et al. [2] utilized a dual-CNN–LSTM model, in which the CNN channel, responsible for data dimension reduction and feature exploration, processes the data while the LSTM channel handles time-series data. The

Received 2026-11-11; Accepted 2026-04-16; Available online 2026-04-16

information from both channels is fused to predict the bearing degradation trend. These models significantly reduce manual intervention by automatically learning fault features from raw signals, and have made progress in fusing multimodal signals. However, traditional CNNs lack denoising structures [3], and their capability to extract features from vibration signals is constrained by noise interference. Furthermore, the model training process relies heavily on specific marine condition samples, resulting in a lack of generalization ability across different equipment and operating conditions. These issues limit the reliability and stability of bearing anomaly detection. Therefore, it is urgent to develop novel intelligent diagnostic methods with enhanced noise immunity and working-condition adaptability.

First and foremost, this paper focuses on two main technical problems:

1) To extract discriminative bearing feature signals in high-noise environments;
2) To enhance the adaptive detection capability of the model for various fault types.

As for signal acquisition of bearing fault features, scholars always select time-frequency analysis method, wavelet transform method, EMD method etc. Among these methods, time-frequency analyses such as wavelet transform and short-time Fourier transform can capture signal changes in both domains. EMD and its variants are utilized to adaptively process non-stationary signals, decomposing the original signal into a series of intrinsic mode functions (IMFs) that reflect different fault features. For instance, Ma et al. [4] proposed a denoising technique based on VMD and dynamic wavelets. Initially, VMD is utilized to decompose the signal and filter the extracted IMF components. Subsequently, dynamic wavelets are applied to denoise and reconstruct the mode-mixed IMFs, achieving joint denoising. Finally, a deep learning model is employed to extract fault features from the denoised signal to accomplish the diagnosis. The study by Bayram et al. [5] utilized wavelet transform for the noise decomposition of signals and conducted an in-depth analysis of how bearing faults influence wavelet coefficients. Furthermore, Akcan et al. [6] utilized entropy-based features combined with an Extreme Learning Machine (ELM) for diagnosis. Although these methods have achieved certain successes, they heavily rely on manually configured key parameters (such as the wavelet basis function, decomposition levels, number of modes, and entropy calculation parameters). Consequently, they exhibit poor adaptability in highly variable marine environments (e.g., non-stationary vibrations induced by salt spray and wave impacts) and still possess limited capability in extracting features from non-stationary signals. These numerous limitations restrict practical application effectiveness under marine working conditions.

In order to better adapt to detection tasks, deep learning models, CNN and LSTM included, are mainly used for bearing fault recognition. For instance, Gao et al. [7] employed a one-dimensional convolutional neural network to derive a fully automated fault diagnosis model that took in unprocessed vibration signals directly, and completely automated the transition from the process of processing the incoming signal right up until the act of classifying faults. Yang et al. [8] added attention mechanism to combine the features of different modalities effectively in order to get full complementary among signals. To capture features under complex working conditions, Kaya et al. [9] used continuous wavelet transform (CWT) to convert signals into color time-frequency images, and then applied deep transfer learning to predict fault sizes. To improve model interpretability and better capture complex features, You et al. [10] proposed a deep learning method guided by physical constraints, fusing acoustic and vibration data. By introducing nonlinear dynamic models as constraints, they made their network much more interpretable. Similarly, Chen et al. [11] used cross-layer modules and skip connections to tackle the issues of sparse features and long-term dependencies. However, these methods still fall short in marine environments. When dealing with strong background noise and non-stationary signals, standard models struggle to separate the environmental noise from actual fault features. As a result, they lack adaptive detection capabilities and tend to overfit on specific training data. This poor generalization makes it hard for them to meet the strict reliability requirements of real-world detection.

Overall, existing approaches still struggle with extracting features in highly noisy environments and achieving adaptive detection. To tackle these issues, we propose a fault diagnosis framework based on multimodal fusion and two-stage discrimination. There are two stages in the framework. Stage 1: In order to deal with the background noise and the non-stationary vibration signals, firstly the feature enhancement part is used, which applies the VMD-CEEMDAN hybrid structure on the original signals, decomposing the original data into various modal components and then extracting the fault-related sensitive components. Then, perform weighted reconstruction of the components using envelope kurtosis as an adaptive factor to enhance fault features and achieve a good SNR. Finally, this module performs the improvement of signal feature enhancement, which can present the fault characteristic information of the signal in the time and frequency domain more intuitively and make it easier for people to differentiate. The second stage is to create a fault discrimination dual-channel deep network, which serves as the fault discrimination module. Channel 1 takes time and frequency domain features as the first channel. It uses 1D convolution to extract temporal sequence structure. Channel 2 takes the Hilbert time-frequency images as input and applies 2D convolutions to extract spatial patterns.

We then incorporate a Convolutional Block Attention Module (CBAM) to highlight key features. Finally, the network fuses the information from both channels and passes it through fully connected layers to classify the fault type. Together, these two main modules form a 'signal demixing - heterogeneous fusion' processing chain. This setup effectively tackles the issues of blurred features and low recognition accuracy caused by marine noise.

The remainder of this paper is organized as follows: Section 2 introduces the related research. Section 3 presents the framework of the proposed method. Sections 4 and 5 describe the specific modules in detail. Section 6 provides the experimental analysis, and Section 7 concludes the paper.

2. RELATED WORKS

2.1. Marine machinery operating environment and bearing fault characteristics

Offshore wind turbines are a typical form of marine mechanical equipment, constantly facing unique operating conditions, including high humidity, high salinity, high pressure, and long-term continuous operation. In high-humidity and high-salinity environments, mechanical components are eroded by seawater, which leads to accelerated corrosion on metal surfaces and consequently shortens the equipment's service life. Under high-pressure conditions, the equipment's sealing integrity and structural strength are rigorously tested, as even minor defects can trigger significant safety hazards. Furthermore, long-term continuous operation leads to the gradual accumulation of fatigue damage, which presents additional difficulties for fault prediction and maintenance.

Bearings in this equipment are important components of the mechanical system transmission, and their status directly affects the performance and safety of the entire system. But bearings are subjected to dynamic loading and varying operating conditions as well as varying loads and corrosive media ingress, with complex coupled vibrations that work together making those subtle defects difficult to detect when they first cause trouble. When load fluctuates, it brings change of the stress when the bearing's two rings make contact, so fatigue cracks could get started. The ingress of corrosive media significantly accelerates lubricant degradation; the resulting decline in lubrication efficiency leads to increased friction and elevated temperatures. Additionally, coupled vibrations are complex, so there are nonlinear responses in the bearings, making it difficult to identify faults. Therefore, to ensure equipment reliability and safety, enhancing the robustness of fault diagnosis on the bearing of marine machinery is of great importance.

2.2. Bearing fault diagnosis methods based on signal processing

To obtain the time-domain, frequency-domain, and time-frequency domain features of the bearing fault signals, many researchers have put forward different signal processing methods. Among them, the Wavelet packet transform (WPT), adaptive multi-scale signals decomposition method, makes up for the lack of resolution of the conventional wavelet transform at high frequencies. WPT can make a fine decomposition of vibration signals over the whole frequency band. The method decomposes the raw vibration signal into evenly sized sub-bands via a repeated filtering approach; at each level of decomposition, the signal is filtered by a low-pass filter to produce the low-frequency approximations, and then by a high-pass filter to give the high-frequency details. In their research, Li et al. [12] first carried out WPT decomposition on the raw vibration signal to get time-frequency coefficient vectors. They then manually selected and kept the low-frequency sub-bands containing the fault information, and dropped the noise dominated higher frequency elements. The two dimensional matrix has been made once more with a coefficient line chosen. An act of normalization was done such that the elements in this matrix can become aligned over a normal grey scale, which normalized into a grayscale image which can then be taken as a form of input for Convolutional Neural Network. Similarly, EMD and its derivatives are employed to decompose composite signals into a set of Intrinsic Mode Functions (IMFs) and separate them. And then we do the condition identification according to the distribution of IMF components. The prior methods manage to get a little bit done in practical use, but marine operating environment is a significant challenge and the trouble caused by signals mixing and feature information loss is pretty much the same: the unstructured noise just remains a severe barrier.

2.3. Diagnosis methods based on machine learning

Today, research on machine learning for bearing fault diagnosis is both extensive and deep. While traditional models like Support Vector Machines (SVM) and Random Forests (RF) remain widely used for classification, researchers have continually introduced advanced feature extraction strategies and intelligent algorithms. For instance, Kuncan et al. [13] innovatively built a hybrid model combining 1D Local Binary Patterns (1D-LBP) and Grey Relational Analysis (GRA), classifying faults based on statistical features in the 1D-LBP plane. Similarly, Kaya et al. [14] used 1D-LBP to preprocess vibration signals, constructed co-occurrence matrices, and extracted texture features like correlation and energy to achieve highly accurate fault discrimination. Kaplan et al. [15] classified bearing defects of various sizes by extracting real-time statistical features from vibration data and feeding them into an Artificial Neural Network (ANN). Meanwhile, Huang et al. [16] proposed an Interactive Generative Feature Space Oversampling Autoencoder (IGFSO-AE) that interpolates within the latent space to generate

diverse samples, effectively boosting the model's robustness when dealing with imbalanced data.

Although these methods perform well in specific scenarios, most of them still rely heavily on hand-crafted feature extraction or specific data preprocessing steps. To streamline the diagnostic process and minimize the information loss caused by manual feature engineering, researchers are increasingly turning to deep learning models - such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Temporal Convolutional Networks (TCN - for end-to-end bearing anomaly detection. Since vibration signals are highly susceptible to noise, Guo et al. [17] proposed a CNN-based diagnostic method using motor rotational speed signals. By fusing deep speed features with frequency-domain information, they achieved highly accurate recognition under complex working conditions. Furthermore, to tackle the challenge of identifying faults in rotating equipment like Computer Numerical Control (CNC) machines amidst complex noise, Iqbal et al. [18] developed a CNN diagnostic framework that fuses vibration and acoustic signals. Their method uses the Short-Time Fourier Transform (STFT) to convert 1D signals into 2D time-frequency images, allowing the CNN to extract spatial features. This approach clearly demonstrated the effectiveness of multimodal signal fusion and significantly boosted fault classification accuracy. Wang et al. [19] adopted LSTM network for bearing life prediction. Firstly, extract feature parameter from time, frequency, and time-frequency, and select them according to feature assessment metrics, and finally create a feature set that contains time factors. The LSTM model was trained using this data set, and the prediction was performed. It is also true that the training of this kind of model needs to use some specific samples of the marine working state, has many parameters, and is very inefficient. To some extent, this leads to poor generalization capability across different equipment and operating conditions, and their robustness in marine environments needs further improvement. Nevertheless, end-to-end bearing fault diagnosis frameworks represent the future trend of automated and intelligent fault diagnosis.

2.4. Applications of Multimodality and Attention Mechanisms in Intelligent Diagnosis

To further enhance fault representation capabilities, multimodal learning and attention mechanisms have been introduced. Qin et al. [20] proposed a dual-channel TC-CNN model. One channel consists of a one-dimensional CNN (1D-CNN) that takes the frequency spectrum of the FFT-extracted vibration signals as input, while the second channel is a 2D-CNN that processes time-frequency images extracted from the Generalized S-Transform (GST). After feature extraction is done in each channel separately, the feature vectors are fused in a fusion layer then classifier will be used for classification. Saghi et al. [21] input time-series vibration data into a three-channel parallel one-dimensional Convolutional Neural Network (1D-CNN). The convolutional layers employ kernels of different sizes to obtain multi-scale local features. The output of each channel then passes through a Bidirectional GRU to capture the overall temporal dependencies. After concatenating all multi-scale spatio-temporal features, a Fully Connected layer will integrate the information and use a Softmax classifier for the final fault identification. The MSCNN - BiLSTM - AM model proposed by Zhao et al. [22] first takes a 1D vibration signal as input. Two independent convolutional layers are simultaneously employed to extract and fuse the spatiotemporal features of the signal. Next, they pass these fused multi-scale feature maps to a BiLSTM layer, which will learn the bidirectional temporal correlations from the sequence. We introduce CAM as an attention mechanism which assigns different weights to the features of BiLSTM which increases the weight of important fault feature. And finally, the diagnosis is output via a fully connected layer and a Softmax classifier.

2.5. Principles of Relevant Methods

2.5.1. Variational Mode Decomposition (VMD)

Different from EMD, VMD [23] is a complete non-recursive modal extracting algorithm, which could separate the input signal into intrinsic modes and determine the center frequency and bandwidth of each mode for separation. The heart of the algorithm is in establishing a constrained variational model, so one needs to perform the Hilbert transform on each intrinsic mode to shift their spectrum to the right baseband;

$$\left[\left(\delta(t)+\frac{j}{\pi t}\right)*u_k(t)\right]e^{-j\omega_k t} \tag{1}$$

where $\delta(t)$ is the Dirac delta function; $u_k(t)$ is the set of the intrinsic mode components, and ω_k are the set of central frequencies for each mode.

Variational model which is based on the estimate of bandwidth of respective mode is given as:

$$\min_{\{u_k\},\{\omega_k\}}\left\{\sum_{k=1}^{K}\ \partial_t\left[\left(\delta(t)+\frac{j}{\pi t}\right)*u_k(t)\right]e^{-j\omega_k t}{}_2^2\right\} \tag{2}$$

Subject to:

$$\sum_{k=1}^{K}\ u_k(t)=f(t) \tag{3}$$

where is the partial derivative of the function with respect to time t; "$*$" denotes the convolution operation.

To solve the constrained optimization problem, the variational model is transformed into an unconstrained problem by introducing a quadratic penalty term α and a Lagrange multiplier $\lambda(t)$. The augmented Lagrangian function is thus constructed as:

$$\mathcal{L}(\{u_k\},\{\omega_k\},\lambda)=\alpha\sum_{k=1}^{K}\ \partial_t\left[\left(\delta(t)+\frac{j}{\pi t}\right)*u_k(t)\right]e^{-j\omega_k t}{}_2^2+f(t)-\sum_{k=1}^{K}\ u_k(t)_2^2+\lambda(t),f(t)-\sum_{k=1}^{K}\ u_k(t) \tag{4}$$

where $\mathscr{L}$ is the augmented Lagrangian; $\lambda(t)$ is the Lagrange multiplier.

The solution is obtained in the frequency domain by applying Parseval's theorem. The Alternate Direction Method of Multipliers (ADMM) is used to iteratively update u_k, ω_k, and λ. The update process for $k = 1, \ldots, K$ is as follows:

$$\hat{u}_k^{n+1}(\omega) = \frac{\hat{f}(\omega) - \sum_{i<k} \hat{u}_i^{n+1}(\omega) - \sum_{i>k} \hat{u}_i^{n}(\omega) + \frac{\hat{\lambda}^n(\omega)}{2}}{1 + 2\alpha(\omega - \omega_k^n)^2} \quad (5)$$

The update equation for the center frequency is:

$$\omega_k^{n+1} = \frac{\int_0^\infty \omega \left|\hat{u}_k^{n+1}(\omega)\right|^2 d\omega}{\int_0^\infty \left|\hat{u}_k^{n+1}(\omega)\right|^2 d\omega} \quad (6)$$

$$\hat{\lambda}^{n+1}(\omega) = \hat{\lambda}^n(\omega) + \tau\left[\hat{f}(\omega) - \sum_{k=1}^{K} \hat{u}_k^{n+1}(\omega)\right] \quad (7)$$

where "^" denotes the Fourier transform; i is an integer ranging from 0 to K.

If the convergence criterion is met:

$$\sum_{k=1}^{K} \frac{\left\| \hat{u}_k^{n+1} - \hat{u}_k^{n} \right\|_2^2}{\left\| \hat{u}_k^{n} \right\|_2^2} < \varepsilon \quad (8)$$

the iteration is terminated; otherwise, the process returns to the iterative steps.

2.5.2. Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN)

CEEMDAN is enhanced EEMD. After the IMF components $r(t)$ have been extracted, CEEMDAN will add back white noise to the current residual signal [24,25], and compute the next IMF components using an iterative average method. CEEMDAN compared to EMD and EEMD has better completeness and computational efficiency. It is obtained as follows:

Step 1: Add instances of white noise to the original signal and decompose the obtained signal multiple times using EMD: The first IMF component of CEEMDAN, is obtained via averaging the first IMF component from each decomposition:

$$I_1(t) = \frac{1}{n}\sum_{i=1}^{n} E_1[x(t) + \varepsilon_0 w_i(t)] \quad (9)$$

where t represents time; i is the index of the EMD trial; and $E_1[\cdot]$ is the first IMF component obtained by applying EMD to the newly formed signal $x(t) + \varepsilon_0 w_i(t)$.

Step 2: Calculate the first residual component $r_1(t)$:

$$r_1(t) = x(t) - I_1(t) \quad (10)$$

Step 3: To $r_1(t)$, add n instances of white noise $\varepsilon_1 w_i(t)$ to obtain $r_1(t) + \varepsilon_1 w_i(t)$, and perform EMD on each of the resulting signals n times. The second IMF component of CEEMDAN, $I_2(t)$, is then obtained by taking the mean of the first IMF component from each decomposition:

$$I_2(t) = \frac{1}{n}\sum_{i=1}^{n} E_1[r_1(t) + \varepsilon_1 w_i(t)] \quad (11)$$

Step 4: Calculate the k -th residual component $r_k(t)$:

$$r_k(t) = r_{k-1}(t) - I_k(t) \quad (12)$$

Step 5: To $r_k(t)$, add n instances of white noise $\varepsilon_k w_i(t)$ to obtain $r_k(t) + \varepsilon_k w_i(t)$, and perform EMD on each of the resulting signals n times. The $(k+1)$-th IMF component of CEEMDAN, $I_{k+1}(t)$, is then obtained by taking the mean of the first IMF component from each decomposition:

$$I_{k+1}(t) = \frac{1}{n}\sum_{i=1}^{n} E_1[r_k(t) + \varepsilon_k w_i(t)] \quad (13)$$

Step 6: If the residual component $r_k(t)$ contains at most one extremum point, the decomposition is considered complete; otherwise, the process returns to Step 4 to continue decomposing until the condition is met. The final residual component $r_e(t)$ is obtained as:

$$r_e(t) = x(t) - \sum_{i=1}^{K} I_i(t) \quad (14)$$

where K is the total number of IMF components obtained.

Step 7: The modal decomposition result of the original signal $x(t)$ is as follows:

$$x(t) = \sum_{i=1}^{K} I_i(t) + r_e(t) \quad (15)$$

2.5.3. Convolutional Neural Network (CNN)

CNN is a deep learning model that has become quite popular in the field of image recognitions and also in object detections recently. CNNs do reduce complexity through some techniques like weight sharing and pooling operations on networks, and the risk of overfitting is cut down along the way, thus the usage of big scale deep learning can now be possible. It consists of a convolutional layer, a pooling layer, and a fully connected layer [26,27], which are detailed as follows:

1) Convolutional Layer

The convolutional layer performs a convolution operation on the input image using multiple convolutional kernels. After a bias is added, the result is passed through an activation function to generate a feature map. The mathematical expression for the convolution process is:

$$X_j^l = f\left(\sum_{i=1}^{M} X_i^{l-1} * w_{ij}^l + b_j^l\right) \quad (16)$$

where X_j^l is the$j$ feature map of the l layer; M is the set of input feature maps; w_{ij}^l is the weight of the convolutional kernel connecting the i -th feature map of the $(l-1)$-th layer to the j -th feature map of the l -th layer; b_j^l is the bias term; "$*$" denotes the convolution operation; and $f(\cdot)$ is the activation function, commonly the ReLU function:

$$f(x) = \max(0, x) \quad (17)$$

2) Pooling Layer

The pooling layer downsamples the feature maps and maintains a degree of feature scale invariance. Common pooling methods include max pooling,

average pooling, and random pooling. This paper takes max pooling as an example:

$$P_j^l = \max_{n\times n}\left(X_j^{l-1}\right) \tag{18}$$

where $n \times n$ is the size of the pooling window.

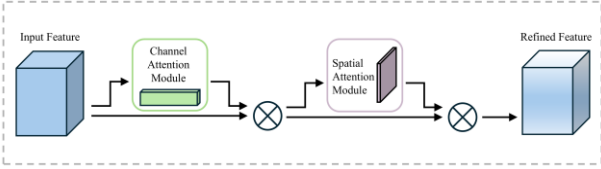

Fig. 1. Schematic diagram of CBAM

This means that while the number of output feature maps remains unchanged, both of their spatial dimensions are reduced by a factor of n.

3) Fully Connected Layer

The fully connected layer transforms the feature map matrix into a one-dimensional feature vector, which is then fed into a classifier. The model can be expressed as:

$$y^k = f(w^k x^{k-1} + b^k) \tag{19}$$

where k is the network layer index; y^k is the output of the fully connected layer; x^{k-1} is the flattened one-dimensional feature vector from the $(l-1)$-th layer; w^k is the weight coefficient; and b^k is the bias term. For multi-class classification problems, the activation function $f(\cdot)$ is the Softmax function.

For a specific classification task, the CNN adjusts the weights and biases between layers to minimize the network's loss function. This paper employs the cross-entropy loss function to optimize the network parameters:

$$E = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\ln\left(p_{i,c}\right) \tag{20}$$

where N is the number of samples for a specific fault class; C is the number of classes; $y_{i,c}$ is the true value; and $p_{i,c}$ is the predicted probability. The network parameters are updated layer by layer using the gradient descent method, with the calculation as follows:

$$\begin{cases} w' = w - \eta\dfrac{\partial E}{\partial w} \\ b' = b - \eta\dfrac{\partial E}{\partial b} \end{cases} \tag{21}$$

where w' and b' are the updated weights and biases; w and b are the original weights and biases; and η is the learning rate, which controls the step size of the weight updates. If η is too large, the process is prone to getting trapped in a local optimum, whereas if it is too small, the convergence speed will be reduced.

2.5.4. Convolutional Block Attention Module (CBAM)

CBAM is a simple, effective, lightweight, and general-purpose module that can be seamlessly integrated into any CNN architecture and trained end-to-end with the base CNN [28]. CBAM is composed of two independent sub-modules, the Channel Attention Module (CAM) and the Spatial Attention Module (SAM), with a schematic of its structure shown in Figure 1. Given an input feature map $F \in R^{C\times H\times W}$, where C, H, and W represent the number of channels, height, and width, respectively, CBAM sequentially generates a channel attention map, a spatial attention map, and the final output feature.

1) Channel Attention Module (CAM)

The CAM assigns higher weights to channels that contain critical information, enabling the model to focus on the channels that include fault features. The CAM first applies both global max pooling and average pooling to the features to generate two distinct one-dimensional vectors. These vectors are then processed through a shared Multilayer Perceptron (MLP). The output features from the MLP are activated by a Sigmoid function to produce the channel attention weight coefficients $M_C F$. This process can be expressed as:

$$M_C F = \sigma f_{mlp}\left(A(F)\right) + f_{mlp}\left(M(F)\right) \tag{22}$$

where σ denotes the Sigmoid function; f_{mlp} denotes the Multilayer Perceptron model; and A and M represent average pooling and max pooling, respectively.

The channel attention weight coefficients are used to represent the importance of each channel. They are multiplied element-wise with the input features to obtain the channel attention feature map F':

$$F' = M_C F \otimes F \tag{23}$$

where "⊗" denotes element-wise multiplication.

2) Spatial Attention Module (SAM)

The feature map F' is used as the input to the SAM, which effectively extracts fault features from specific regions by learning the key information within each area of the feature map. The SAM first applies both global max pooling and average pooling to F'. The two resulting two-dimensional vectors are then concatenated along the channel dimension and fed into a convolutional layer $f_{7\times7}$ with a 7x7 kernel size. After being passed through a Sigmoid activation function, the spatial attention weights of the feature map, $M_S(F')$, are obtained. This process can be expressed as:

$$M_S(F') = \sigma f_{7\times7}[A(F'); M(F')] \tag{24}$$

The spatial attention weight coefficients are used to represent the importance of each position in the feature map. They are multiplied element-wise with the input feature F' to obtain the spatial attention feature map F'':

$$\mathrm{F''} = M_S(F') \otimes F' \tag{25}$$

2.6. Research Gaps

Although existing studies have made solid progress in bearing fault diagnosis, most methods still reveal major limitations when applied to the unique operating environments of marine machinery, primarily in the following areas:

1) Difficulty in extracting fault features in high-noise environments: In complex marine

conditions, strong noise and non-stationary signals create severe interference that often masks weak fault features. Traditional signal processing methods, like standalone VMD or wavelet transforms, lean heavily on human expertise to extract these features. Because they are highly sensitive to parameter settings, they struggle to adaptively separate valid fault components in low signal-to-noise ratio (SNR) scenarios, frequently resulting in the loss of critical information.

2) Limited adaptive detection capabilities: While deep learning has certainly boosted adaptive detection, most models still rely on a single modality—either 1D time-series signals or 2D images. This limited perceptual dimension makes it hard to fully uncover the heterogeneous features hidden within complex, coupled signals. More importantly, when faced with noise interference, conventional models lack a mechanism to adaptively focus on critical fault features. As a result, they tend to overfit on specific datasets, ultimately leading to false alarms and missed detections.

To address these deficiencies, it is urgent to develop a bearing fault diagnosis model that combines robust feature extraction with strong adaptive detection. A two-stage processing mechanism offers a practical way forward. The first stage focuses on signal demixing and feature enhancement. For data cleaning, we construct a cascaded VMD-CEEMDAN structure: VMD first extracts the principal modes from the raw signal, and CEEMDAN then performs a secondary decomposition to prevent mode mixing and preserve data integrity. Afterward, these components are adaptively weighted and reconstructed based on their envelope kurtosis. In the second stage, a CNN fuses 1D time-series features with 2D images for joint multimodal analysis. By incorporating an attention mechanism, the network automatically focuses on critical fault regions, further improving its adaptive detection. In this way, the first stage makes up for the lack of built-in denoising in standard deep learning models, while the second stage boosts overall adaptability. Together, they ensure highly accurate fault recognition, even under severe background noise.

3. PROPOSED METHOD FRAMEWORK

Aiming at the severe background noise in marine machinery bearing vibration signals and the limited adaptive detection ability of existing models, this paper studies a bearing fault diagnosis model based on multimodal fusion and two-stage discrimination. In the first part of the whole framework, this part is the feature enhancement module whose goal is to extract fault discriminative features from nonstationary and strong-noise signals. Because vibration signals are contaminated by interferences from different components and nonlinear responses, the fault patterns are easily hidden. This module first employs a VMD-CEEMDAN composite structure for signal demixing. Taking advantage of the non-recursive nature of VMD, it initially decomposes the raw vibration signal into multi-scale mode components with distinct center frequencies. This step uses VMD's frequency-domain focusing to achieve macro-level demixing and suppress strong background noise. Following this, CEEMDAN is applied to the residual signal left by VMD. By iteratively adding positive and negative adaptive white noise to the residual and averaging the results, CEEMDAN generates highly stable Intrinsic Mode Functions (IMFs). This step acts as micro-level stripping, helping to capture weak fault features that would otherwise remain hidden. Then, these components are dynamically filtered based on their permutation entropy and kurtosis. We obtain the targeted denoised signal by keeping only those valid components whose entropy is significantly lower than that of pure noise and whose kurtosis exceeds 3. Next, utilizing the fact that the amplitude of a signal's envelope varies much more in the presence of fault information, the signal's envelope is analyzed along with kurtosis [29]. The Hilbert transform is performed on the screened components to obtain their envelope signals and to calculate the envelope kurtosis. The envelope kurtosis of the denoised and screened components is used as an adaptive weighting factor for a second round of sorting and weighted fusion of the components, thus finishing the feature signal enhancement process.

In the second stage, we construct a dual-channel deep network integrated with an attention mechanism to accurately classify fault types using the enhanced features. To address the heterogeneity of bearing faults across temporal and spatial dimensions, this dual-channel design fully exploits multimodal information. Specifically, the first channel takes the reconstructed 1D enhanced time-domain signal as input. It uses a 1D-CNN to extract sequential structural features and capture the dynamic trends of the vibration signals. Meanwhile, the second channel takes the 2D Hilbert time-frequency images as input, using a 2D-CNN to extract spatial patterns and map out the frequency distribution of the faults. Once extracted, these effective features are fed into a CBAM. Within this module, CAM and SAM sequentially assign different weights to channels based on their importance [30] and evaluate the significance of various spatial locations in the images. This highly targeted approach dynamically emphasizes the most critical fault-related characteristics. Finally, the network fuses the weighted features from both channels and passes them through a fully connected layer and a Softmax classifier to pinpoint the fault type. The workflow of the proposed bearing fault diagnosis based on multimodal fusion and two-stage discrimination can be seen from Figure 2.

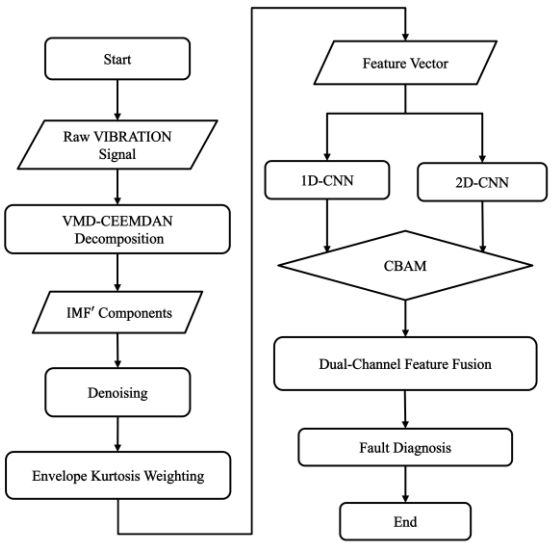

Fig. 2. Bearing fault diagnosis flowchart

4. CORE METHOD 1: FEATURE ENHANCEMENT MODULE (SIGNAL ENHANCEMENT MODULE)

4.1. Problem statement

Extracting discriminative fault features from non-stationary vibration signals contaminated with strong noise remains a core challenge in the fault diagnosis of marine machinery bearings. This challenge is attributed to the fact that vibration signals are often obscured by interference from multiple components and non-linear structural responses, leading to the weakening and masking of the fault information.

4.2. Research content and technical approach

It is developed as a 'decomposition-denoising-processing of feature enhancement', where it will mainly adopt:

Decomposition stage employs VMD to decompose the raw vibration signal into several modes with distinct center frequencies. The penalty factor and an iteration stopping threshold are set to guarantee the sparsity and frequency-domain concentration of the components. It separates the signal into multi-scale modality components each with a clearly identifiable central frequency, and the residue. And then, we carry out the CEEMDAN decomposition on the nonlinear residual term after VMD. Adding white noise back into the current residual signal and performing iterative computation, we can get more sub-sequences and a new residual since the white noise has a zero mean, its effects cancel out after multiple averages [31], but the actual signals remain. Instead of relying on a single-stage decomposition, we integrate VMD with CEEMDAN. First, VMD uses variational optimization to iteratively solve a constrained model, breaking the signal down into several intrinsic modes with specific center frequencies and limited bandwidths. This frequency-domain constraint acts as a powerful narrowband filter. It allows the algorithm to focus on the high-energy fault bands, cleanly separating the fault features from interfering modes. However, due to its inherent limitations, the residual signal left over after VMD might still hide some nonlinear features. To tackle this, we bring in CEEMDAN. By iteratively adding white noise to the residual, CEEMDAN changes the distribution of extreme points. This breaks through the non-stationary nature of the residual signal, allowing us to fully extract any remaining weak fault shocks. By combining these two methods, we solve the feature loss issue that VMD often faces when dealing with non-stationary residuals. Ultimately, this dual-filtering approach ensures that our feature extraction is truly comprehensive.

In the denoising part, calculate the permutation entropy and kurtosis of each modal element obtained through VMD-CEEMDAN decomposition. Due to the periodic impulses generated by the faulty components colliding with other elements, the impact inside the fault-related signal components is more ordered than the normal and pure noise signals. So, it has lower permutation entropy values. Consequently, the sudden changes of the permutation entropy are used to find the possible fault signal. The combined use of this along with the signal's kurtosis enables a comprehensive screening of the components which have a higher portion of fault information and retains critical fault features and completes the desired denoising.

Although a denoise is performed in the last step, there will still be a mixture of fault signatures among the modal components that have been screened out, that is, it can appear in different resonant frequency bands. Therefore, a quantitative metric is required to fuse these components well and highlight the critical fault information.

Envelope kurtosis is better at representing periodic influence, the larger the envelope kurtosis of a component, the stronger the periodic fault information is contained in it. Then during this period, taking the envelope kurtosis as an adaptive factor, a weighted reconstruction of the screened modal components is performed to obtain a result signal with maximally enhanced fault features.

4.3. Innovations

Introducing a cascaded VMD-CEEMDAN structure to dual-filter offshore wind turbine bearing signals. First, VMD macro-decomposes the raw signal to isolate heavy background noise. Then, CEEMDAN micro-strips the residual to capture weak shocks without mode mixing. By combining VMD's frequency focus with CEEMDAN's sensitivity to weak features, this structure deeply purifies the signal from global bands down to local residuals, solving the feature loss issue typical of single methods in noisy marine environments. The signal is denoised based on permutation entropy and kurtosis. Also overcomes the problem of loss features compared to the hard threshold method, increases noise reduction specificity.

We utilize permutation entropy and kurtosis to denoise the signal. Permutation entropy measures the signal's regularity to pinpoint fault components, while kurtosis serves as a validation screen. Unlike traditional hard thresholding, this approach prevents feature loss and achieves highly targeted noise suppression.

We use envelope kurtosis to quantify fault features and derive adaptive weight factors for non-linear signal fusion. Instead of simply adding the components together—which often yields a poor signal-to-noise ratio, this weighted approach actively highlights critical fault impulses and ensures complete feature preservation. We use envelope kurtosis to quantify the fault features and create adaptive weights. By doing this, we can fuse the feature signals nonlinearly instead of just adding them up. This avoids the poor SNR that usually comes with simple reconstruction and makes the key fault pulses much clearer, keeping the important data intact.

5. CORE METHOD 2: FAULT DISCRIMINATION MODULE (FAULT CLASSIFICATION MODULE)

5.1. Problem statement

Beyond feature enhancement, achieving adaptive discrimination between various fault types and operating conditions still remains a significant challenge. Most current methods lack sufficient adaptability in complex environments, resulting in poor recognition accuracy and inadequate robustness.

5.2. Research content and technical approach

The first path which contains the time and frequency domain features from above enhancement module is concatenated together along a new dimension with the time series feature in order to form an augmented time-series feature vector, which serves as the input for the 1D-CNN. This channel consists of multiple layers of 1D convolutional kernel for the local time feature. Then, normalization and ReLU activation are applied, followed by max-pooling to compress the data while retaining crucial temporal features. In this step we use max-pooling layer to extract maximum values from features maps and to compress them. And then this is composed of some of the top inputs of this max-pooled output of a certain feature map.

The second channel extracts the Hilbert time-frequency spectrum and converts it into a 128×128 image for input into the 2D-CNN. And it will do a 2D convolution many times with different 2D layers of convoluted kernel to get the space texture spatial pattern feature map, and then reduce the dimension with a pooling layer.

For each channel, downsampling is performed, and the pooled features are fed into a CBAM. Due to the difference in feature dimensions, Channel 1 uses a 1D-CBAM, while Channel 2 employs a standard 2D-CBAM. Essentially, CBAM leverages attention mechanisms to dynamically reweight feature maps, forcing the model to focus on key features and boosting representation. For a feature map with dimensions $C \times H \times W$, the CAM first compresses the spatial dimensions by performing global average pooling and global maximum pooling. These two resulting 1D vectors are sent to an MLP, and their outputs are summed and passed through a Sigmoid activation function to generate the weights M_c. These weights are then element-wise multiplied with the original input to produce the feature map F', which serves as the input for the SAM. Similarly, the SAM complements the CAM by applying global average and maximum pooling along the channel dimension of F'. The two resulting maps are concatenated and convolved using a single kernel, then passed through a Sigmoid activation to obtain the spatial weights M_s. Finally, these weights are multiplied with the SAM's input to obtain the final result of our "modified" and therefore "improved" " feature map".

Finally, all the features from all the channels are combined and passed into the fully connected layer which performs a non-linear transformation of the features and then performs the fault classification through the softmax classifier. The structure of the fault discrimination module is shown in Figure 3.

5.3. Innovations

Build a dual-channel model that integrates heterogeneous time-frequency information. Channel 1 uses a 1D-CNN to sense local dynamics, while Channel 2 employs a 2D-CNN to capture global spatial patterns in the time-frequency images. This multimodal complementarity enhances the model's precision in representing complex, coupled signals.

The CBAM module is used to sharpen the model's focus on key fault features. By cascading CAM and SAM, the network adaptively weights the feature maps to lock onto fault-sensitive regions while suppressing any tiny noise remaining after dual filtering, thereby enhancing overall adaptability and robustness.

6. EXPERIMENTAL VERIFICATION

6.1. Data Source

To verify whether the above models work well, the public rolling bearing dataset from Case Western Reserve University (CWRU) is employed for experimentation. The dataset contains vibration signals collected under various working conditions, including inner race, outer race, and rolling element faults. The data sets have already been used for validating the effect of bearing fault diagnosis method. Our dataset comes from the CWRU Bearing Data Center's standard test rig. This setup features a 2 HP motor connected via a coupling to a dynamometer, which applies varying loads to simulate real conditions. Single-point faults were

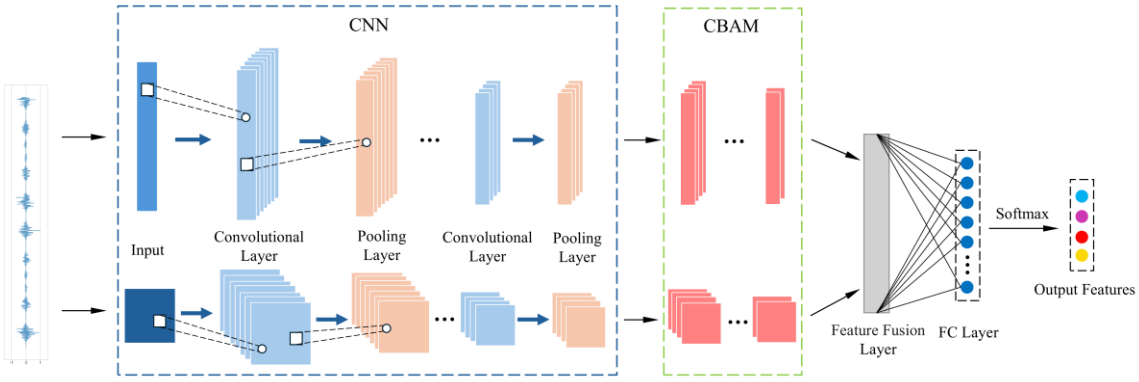

Fig. 3. Fault discrimination module architecture

introduced to the bearings using electro-discharge machining. To capture high-frequency vibrations, an accelerometer with a magnetic base was mounted vertically at the 12 o'clock position on the Drive End of the motor housing. All signals were captured through a 16-channel DAT recorder and post-processed in Matlab.

To conduct the study, an SKF6205-2RS deep groove ball bearing was chosen as the sample. The structural resonance frequency typically excited by impact-induced faults in this bearing model ranges from 2 kHz to 5 kHz. Per the Nyquist theorem, the sampling rate must exceed twice the highest frequency of interest to fully capture the resonance bands. While higher rates increase data density, they also introduce high-frequency environmental noise and unnecessary computational load without meaningfully improving accuracy. Therefore, we selected a 12 kHz sampling frequency, which adequately supports envelope analysis and feature extraction. Using 12 kHz also aligns with common practice for the CWRU dataset, ensuring a fair comparison between our model and existing methods under identical data conditions. Moreover, the experiment was conducted under no-load conditions (1797 r/min). Vibration signals were collected across four conditions, including normal, inner race fault, outer race fault and rolling element fault. With three fault diameters (0.007, 0.014, and 0.021 inches) for each fault type, the experiment involves nine fault categories and 1 normal category, totaling ten classes. To ensure sufficient feature extraction, we expanded the dataset using overlapping sampling, collecting 450 samples per category with a segment length of 2048 points.

6.2. Data processing

Our study involves 10 data categories. We collected 450 samples per category, each with a length of 2048 data points. From these, samples were randomly partitioned into training, validation, and test sets according to a 7:2:1 ratio. Figure 4 shows the time-domain waveform diagrams of the raw vibration signals for four different condition types, where significant differences between the four signals can be observed. Taking the inner race fault with a diameter of 0.014 inches as an example, the fault features in the waveform are obscured by noise, which is especially prominent in the mid- to high-frequency bands. Conversely, the characteristic fault frequency (162.18 Hz) is located in the low-frequency band, where its low amplitude and overlapping spectral lines make it indistinct. Consequently, the vibration signal is subjected to VMD-CEEMDAN decomposition. Prior to VMD, the number of decomposition levels, K, must be determined. An excessively large K value would cause over-decomposition [32]. By feeding the raw data into the model, the optimal number of decomposition levels was calculated to be $K = 4$, and the penalty parameter α was set to its default value of 2000. Applying VMD to the vibration signal yielded four IMF components and one residual term. CEEMDAN decomposition was then applied to this residual, and the time-domain waveforms of the resulting seven IMF' components and the final residual are shown in Figure 5.

According to the definition of permutation entropy, the permutation entropy values were calculated for each IMF'component obtained from the VMD-CEEMDAN decomposition of the inner race fault signal, with a set time delay of $\tau = 2$, an embedding dimension of $m = 3$, and a signal length of 2048. The kurtosis of all components was then calculated using the relevant formula, with the specific data presented in Table 1. As shown in Table 1, the first component has a high kurtosis, but its other parameters do not meet the selection criteria. IMF_2' has the highest kurtosis, and its other indicators are within the expected range. The kurtosis values for components $IMF_4'$and $IMF_5'$are both greater than 3; their permutation entropy values show minor abrupt changes, and their joint coefficients are large, indicating that they contain more complete and rich fault feature information. Therefore, IMF_2',IMF_4',and $IMF_5'$are selected as the effective components. Using envelope kurtosis as an adaptive weighting factor, a weighted reconstruction of IMF_2',IMF_4', and $IMF_5'$is performed. The time-domain and frequency-domain plots of this

reconstructed signal serve as the feature inputs to the adaptive CNN-CBAM model.

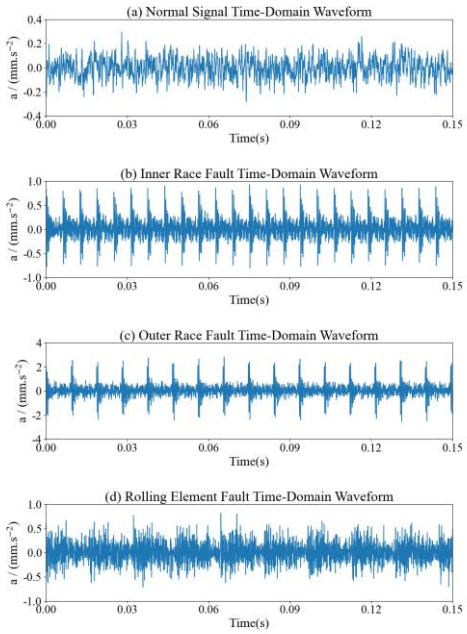

Fig. 4. Raw vibration signal waveforms

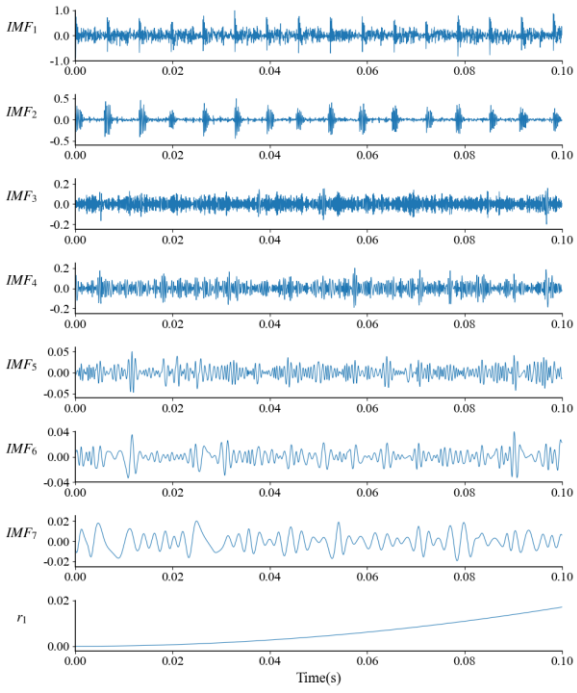

Fig. 5. CEEMDAN decomposition of the inner race fault signal

Table 2 lists the structural parameters for the dual-channel CNN-CBAM fault discrimination module. Channel 1 (1D-CNN) takes a 1 × 2048 input, using a wide 64-size kernel in the first layer for rapid dimensionality reduction, followed by stacked 3-size kernels to extract deep temporal features. Channel 2 (2D-CNN) consistently employs 3×3 kernels. For both channels, the number of kernels increases from 16 to 128 (16-32-64-128). This setup expands the channel depth while gradually reducing the data length and spatial size through pooling, effectively extracting deep-level features. In the attention module, the reduction ratio is $r = 16$, and the spatial attention uses 1D and 2D kernels of size 7. For fusion and classification, the concatenated features are flattened and fed into a fully connected layer. To prevent overfitting, we introduced a Dropout mechanism with a rate of $p = 0.5$, and the final classification is output via a Softmax layer.

Table 1. Indicators of IMF' components for inner race fault signal

Modal component	Permutation Entropy	Correlation Coefficient	Joint Coefficient	Kurtosis
IMF_1'	0.83	0.26	0.31	5.03
IMF_2'	0.64	0.23	0.42	10.01
IMF_3'	0.53	0.18	0.29	2.78
IMF_4'	0.36	0.69	0.78	3.46
IMF_5'	0.31	0.57	0.62	3.31
IMF_6'	0.22	0.16	0.31	2.74
IMF_7'	0.13	0.21	0.14	2.81

Table 2. Parameters of the dual-channel CNN-CBAM

Layer Name	Parameter	Output Size
$C_{1,1}$	Conv1d(64 × 16), Stride=2	1024 × 16
$C_{2,1}$	Conv2d(3 × 3 × 16)	128 × 128 × 16
$p_{1,1}$	MaxPool1d(2), Stride=2	512 × 16
$p_{2,1}$	MaxPool2d(2 × 2), Stride=2	64 × 64 × 16
$C_{1,2}$	Conv1d(3 × 32)	512 × 32
$C_{2,2}$	Conv2d(3 × 3 × 32)	64 × 64 × 32
$p_{1,2}$	MaxPool1d(2), Stride=2	256 × 32
$p_{2,2}$	MaxPool2d(2 × 2), Stride=2	32 × 32 × 32
$C_{1,3}$	Conv1d(3 × 64)	256 × 64
$C_{2,3}$	Conv2d(3 × 3 × 64)	32 × 32 × 64
$p_{1,3}$	MaxPool1d(2), Stride=2	128 × 64
$p_{2,3}$	MaxPool2d(2 × 2), Stride=2	16 × 16 × 64
$C_{1,4}$	Conv1d(3 × 128)	128 × 128
$C_{2,4}$	Conv2d(3 × 3 × 128)	16 × 16 × 128
$p_{1,4}$	MaxPool1d(2), Stride=2	64 × 128
$p_{2,4}$	MaxPool2d(2 × 2), Stride=2	8 × 8 × 128
CAM_1	$r = 16$	64 × 128
CAM_2	$r = 16$	8 × 8 × 128
SAM_1	Conv1d(7 × 1)	64 × 128
SAM_2	Conv2d(7 × 7 × 1)	8 × 8 × 128
C&F	-	16384
FC_1	16384 × 256	256
D	$p = 0.5$	256
FC_2	256 × 10	10

For each fault category, 315 samples were used for training the CNN-CBAM model, 90 for validation, and the remaining 45 for testing. The prediction results, shown in Figure 6, achieved an overall accuracy of 99.11%. To further evaluate the model's macro-positioning capability, we

aggregated different damage levels at the same location to generate a confusion matrix based on fault position, as illustrated in Figure 7, where the vertical axis represents the true fault types and the horizontal axis represents the model-diagnosed fault types. Figure 7 shows that the classification accuracy for the inner race fault reached 100%, while the accuracies for the normal condition, outer race fault, and rolling element fault were 97.8%, 98.5%, and 99.3%, respectively. Across multiple experiments, the diagnostic accuracy for inner race faults consistently exceeded that of outer ring and rolling element faults, aligning with the findings of Smith et al. [33]. This is likely because the inner ring rotates with the main shaft, generating intense amplitude-modulated impacts that the VMD-CEEMDAN structure can precisely separate. In contrast, the outer ring damage points were distributed at the 3, 6, and 12 o'clock; when impacts at the 6 o'clock position reach the sensor, structural energy attenuation occurs, leading to lower accuracy. Furthermore, although the complex transmission path of rolling element faults makes them harder to diagnose than inner ring faults, their accuracy still reached 99.3%. Overall, our method demonstrates high precision in classifying single fault types within the CWRU dataset, confirming the model's feasibility.

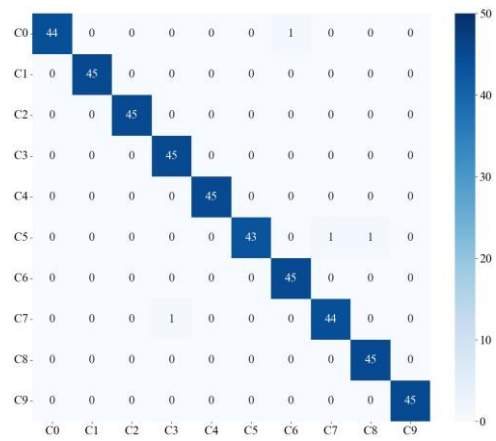

Fig. 6. Fault diagnosis confusion matrix

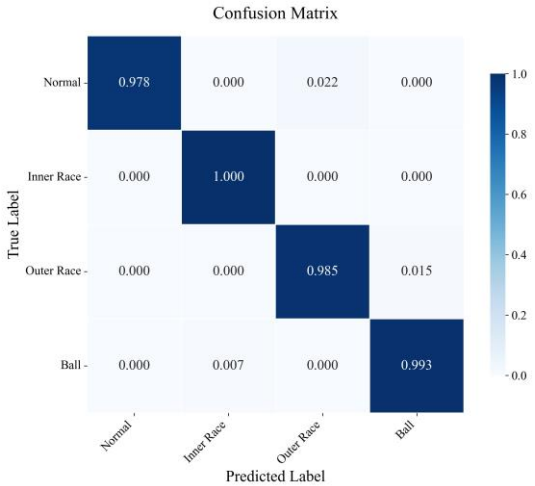

Fig. 7. Confusion matrix for the test set

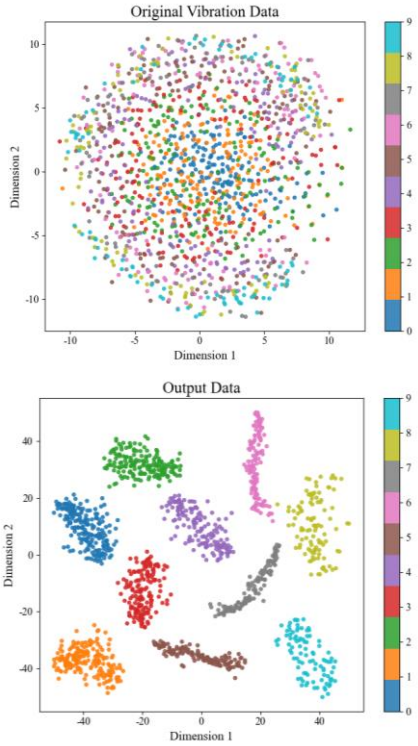

Fig. 8. Input layer and output layer visualization

To visually examine the classification performance, we performed a visualization analysis on the output layer. Figure 8 compares the visualization results of the input and output layers. The model effectively separates the 10 classic fault types. The data points of different types of fault have no obvious overlap, which indicates that the classification performance is very good.

The model was trained and then tested on the same dataset for 100 epochs. The loss curves and the accuracy of the training and test sets are shown in Figure 9 and Figure 10, respectively. The model converges quickly within the first 20 epochs, with the loss dropping sharply. This demonstrates that the high-SNR features from VMD-CEEMDAN reduce the complexity of feature extraction, thereby accelerating the optimization path for gradient descent. After 50 epochs, the train/test set accuracy all stabilize at over 98.0%. More importantly, a model can have a pretty nice capability already in identification even though it is trained for less epochs.

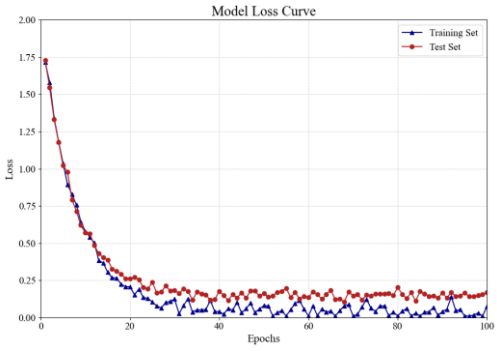

Fig. 9. Loss function curve

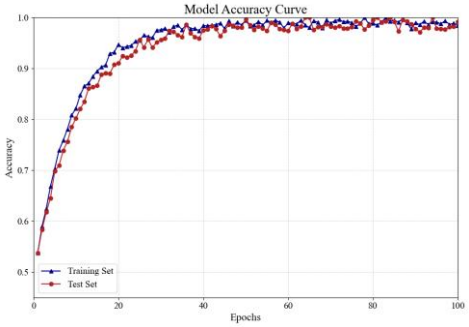

Fig. 10. Accuracy curve

6.3. Generalization Validation

To verify the model's generalization ability, we conducted cross-condition experiments using CWRU samples under different loads. While the original model was trained on data from the no-load condition (0 HP, 1797 r/min), it was then tested against three different conditions: 1 HP (1772 r/min), 2 HP (1750 r/min), and 3 HP (1730 r/min).

Once the model converged on the 0 HP training set, the network weights were fixed, and the untrained 1 HP, 2 HP, and 3 HP samples were directly input for prediction. As the load gap between the source and target conditions increased, the accuracy declines, yielding results of 95.1%, 92.9%, and 90.7%, respectively. Although the accuracy dropped under the largest load gap, it remained at a high level overall. This suggests that, thanks to the adaptive modules, the model possesses robust cross-condition generalization.

6.4. Model Comparison

To further evaluate the proposed multimodal fusion and two-stage discrimination model, we compared it with several recent state-of-the-art methods, as listed in Table 3. While some existing models achieve exceptional accuracy on ideal, noise-free datasets, our model maintains high precision while effectively handling adaptivity and high environmental noise. Consequently, when faced with the complex, variable conditions and strong background noise typical of offshore machinery, the proposed method offers better engineering applicability and comprehensive advantages over most existing models.

7. CONCLUSION AND FUTURE WORKS

To address the difficulties of bearing fault diagnosis in high-noise environment and the limited generalization of existing models for offshore machinery, this study proposes a bearing fault diagnosis framework based on multimodal fusion and two-stage discrimination. Unlike traditional methods that rely on single-signal features or direct end-to-end deep learning networks, the originality of this work lies in the " signal demixing-heterogeneous feature fusion " cascaded processing chain. Through the synergy of feature enhancement and multi-channel attention mechanisms, the model's robustness under harsh operating conditions is significantly improved. The main conclusions are as follows:

1) In the feature enhancement module, a composite VMD-CEEMDAN de-mixing method with adaptive envelope kurtosis weighting is proposed. This approach effectively overcomes mode mixing and residual noise issues inherent in traditional decomposition methods. It can precisely separate and enhance weak fault impact components within strong background noise, providing high-quality feature inputs for subsequent diagnosis stage.

2) In the fault discrimination module, a dual-channel CNN model integrated with the CBAM attention module is constructed. Compared to single-channel networks, this model leverages the multimodal complementarity of time-domain and time-frequency images. By combining attention mechanisms to dynamically focus on key feature regions, it achieves high-accuracy identification across various fault categories.

3) Although the experimental dataset validates the effectiveness of this model, several limitations still remain. First, the two-stage architecture involves complex signal decomposition and dual-channel computations; compared to lightweight models, its higher computational cost poses challenges for real-time deployment on embedded edge devices. Second, model training currently relies on a fairly well-labeled dataset, whereas early-stage weak fault samples are often scarce in actual operating conditions. Future research will focus on the lightweight implementation of the model for real-time edge deployment, while also exploring adaptive fault detection in small-sample learning scenarios to further enhance generalization performance.

Table 3. Reported studies on the bearing fault

Author(s)	Model	Dataset	Fault Type	Accuracy
Kaplan et al. [34]	Signal2Image+LBP	Experimental setup of authors	NS, ORF, IRF, BF +Size	100%
Yoo et al. [35]	Lite CNN	CWRU	NS, ORF, IRF, BF	99.97%
Zhou et al. [36]	TimesBlock and Multi-scale CNN	CWRU	NS, ORF, IRF, BF	99.0%
Hoang & Kang [37]	CNN	CWRU	NS, ORF, IRF, BF	97.74%
Author of this article	VMD-CEEMDAN+CNN-CBAM	CWRU	NS, ORF, IRF, BF	99.11%

Acknowledgments: *The original vibration signals come from the Case Western University Bearing Data Center website.*

Source of funding: *This research received no external funding.*

Data availability: *The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.*

Declaration of competing interest: *The author declares no conflict of interest.*

REFERENCES

1. Lv Z, Liu Y, Han X, Liu M. Study on rolling element bearing fault diagnosis methods based on ensemble empirical mode decomposition. Applied Mechanics and Materials. 2014;457:602-607. https://doi.org/10.4028/www.scientific.net/AMM.457-458.602.
2. Liu G, Zhao J, Zhang X. Bearing degradation trend prediction under different operational conditions based on CNN-LSTM. In: IOP Conference Series: Materials Science and Engineering. 2019;612(3):032042. https://doi.org/10.1088/1757-899x/612/3/032042.
3. Jia L, Chow T, Yuan Y. GTFE-Net: A gramian time frequency enhancement CNN for bearing fault diagnosis. Engineering Applications of Artificial Intelligence. 2023;119:105794. https://doi.org/10.1016/j.engappai.2022.105794.
4. Ma J, Li H, Chen Y, Wang J, Zou Z. Application of VMD and dynamic wavelet noise reduction techniques in rolling bearing fault diagnosis. In: Journal of Physics: Conference Series. 2023;2528(1):012048. https://doi.org/10.1088/1742-6596/2528/1/012048.
5. Bayram S, Kaplan K, Kuncan M, Ertunç HM. The effect of bearings faults to coefficients obtaned by using wavelet transform. In: 2014 22nd Signal Processing and Communications Applications Conference (SIU). 2014:991-994. https://doi.org/10.1109/siu.2014.6830398.
6. Akcan E, Kuncan M, Kaplan K, Kaya Y. Diagnosing bearing fault location, size, and rotational speed with entropy variables using extreme learning machine. Journal of the Brazilian Society of Mechanical Sciences and Engineering. 2024;46(1):4. https://doi.org/10.1007/s40430-023-04567-2
7. Gao Z, Wei Z, Chen Y, Ying T, Gao H. Bearing fault diagnosis using one-dimensional convolutional neural network. In: 2022 22nd International Conference on Control, Automation and Systems (ICCAS). 2022:158-162. https://doi.org/10.23919/iccas55662.2022.10003748.
8. Yang J, Han H, Dong X, Wang G, Zhang S. Bearing fault diagnosis grounded in the multi-modal fusion and attention mechanism. Applied Sciences. 2025; 15(3):1531. https://doi.org/10.3390/app15031531.
9. Kaya Y, Kuncan F, Ertunç HM. A new automatic bearing fault size diagnosis using time-frequency images of CWT and deep transfer learning methods. Turkish Journal of Electrical Engineering and Computer Sciences. 2022;30(5):1851-1867. https://doi.org/10.55730/1300-0632.3909.
10. You K, Wang P, Huang P, Gu Y. A sound-vibration physical-information fusion constraint-guided deep learning method for rolling bearing fault diagnosis. Reliability Engineering & System Safety. 2025; 253:110556. https://doi.org/10.1016/j.ress.2024.110556.
11. Chen Z, He P. Rolling bearing fault diagnosis considering long-term dependence and time-frequency feature fusion. Int J Automot Manuf Mater. 2025. https://doi.org/10.53941/ijamm.2025.100016.
12. Li G, Deng C, Wu J, Chen Z, Xu X. Rolling bearing fault diagnosis based on wavelet packet transform and convolutional neural network. Applied Sciences. 2020;10(3):770. https://doi.org/10.3390/app10030770.
13. Kuncan M. An intelligent approach for bearing fault diagnosis: combination of 1D-LBP and GRA. IEEE Access. 2020;8:137517-137529. https://doi.org/10.1109/access.2020.3011980.
14. Kaya Y, Kuncan M, Kaplan K, Minaz MR, Ertunç HM. A new feature extraction approach based on one dimensional gray level co-occurrence matrices for bearing fault classification. Journal of Experimental & Theoretical Artificial Intelligence. 2021;33(1):161-178. https://doi.org/10.1080/0952813X.2020.1735530.
15. Kaplan K, Bayram S, Kuncan M, Ertunç HM. Feature extraction of ball bearings in time-space and estimation of fault size with method of ANN. In: Proceedings of the 16th Mechatronika 2014. 2014 Dec 3-5; Brno, Czech Republic.
16. Huang F, Zhang K, Li Z, Zheng Q, Ding G, Zhao M, Zhang Y. A rolling bearing fault diagnosis method based on interactive generative feature space oversampling-based autoencoder under imbalanced data. Structural Health Monitoring. 2025;24(2):979-997. https://doi.org/10.1177/14759217241248209.
17. Guo Z, Yang M, Huang X. Bearing fault diagnosis based on speed signal and CNN model. Energy Reports. 2022;8:904-913. https://doi.org/10.1016/j.egyr.2022.08.041.
18. Iqbal M, Madan AK, Ahmad N. Vibration and acoustic signal-based bearing fault diagnosis in CNC machine using an improved deep learning. Iran Journal of Computer Science. 2024;7(4):723-733. https://doi.org/10.1007/s42044-024-00205-9
19. Wang F, Liu X, Deng G, Yu X, Li H, Han Q. Remaining life prediction method for rolling bearing based on the long short-term memory network. Neural Processing Letters. 2019;50(3):2437-2454. https://doi.org/10.1007/s11063-019-10016-w
20. Qin Y, Shi X. Fault diagnosis method for rolling bearings based on two-channel CNN under unbalanced datasets. Applied Sciences. 2022; 12(17):8474. https://doi.org/10.3390/app12178474.
21. Saghi T, Bustan D, Aphale SS. Bearing fault diagnosis based on multi-scale CNN and bidirectional GRU. Vibration. 2022;6(1):11-28. https://doi.org/10.3390/vibration6010002.
22. Zhao D, Tian C, Fu Z, Zhong Y, Hou J, He W. Multi scale convolutional neural network combining BiLSTM and attention mechanism for bearing fault diagnosis under multiple working conditions. Scientific Reports. 2025;15(1):13035. https://doi.org/10.1038/s41598-025-96137-w.
23. Zarour D, Meziani S, Kedadouche M, Thomas M. Faulty bearing features by variational mode decomposition. Vibroengineering Procedia.

2017;16:29-34. https://doi.org/10.21595/vp.2017.19336.
24. Elouaham S, Dliou A, Latif R, Laaboubi M, Zougagh H, Elkhadiri K. Analysis electroencephalogram signals using denoising and time-frequency techniques. International Journal of Advanced Trends in Computer Science and Engineering. 2021; 10(1). https://doi.org/10.30534/ijatcse/2021/101012021.
25. Xiao M, Zhang C, Wen K, Xiong L, Geng G, Wu D. Bearing fault feature extraction method based on complete ensemble empirical mode decomposition with adaptive noise. Journal of Vibroengineering. 2018;20(7):2622-2631. https://doi.org/10.21595/jve.2018.19562.
26. Iqbal M, Madan AK. CNC machine-bearing fault detection based on convolutional neural network using vibration and acoustic signal. Journal of Vibration Engineering & Technologies. 2022;10(5):1613-1621. https://doi.org/10.1007/s42417-022-00468-1.
27. Jiang H, Wang F, Shao H, Zhang H. Rolling bearing fault identification using multilayer deep learning convolutional neural network. Journal of Vibroengineering. 2017;19(1):138-149. https://doi.org/10.21595/jve.2016.16939.
28. Sun B, Hu W, Wang H, Wang L, Deng C. Remaining useful life prediction of rolling bearings based on CBAM-CNN-LSTM. Sensors. 2025;25(2):554. https://doi.org/10.3390/s25020554.
29. Leite VCMN, Borges da Silva JG, Borges da Silva LE, Veloso GFC, Lambert-Torres G, Bonaldi EL, de Oliveira LEL. Experimental bearing fault detection, identification, and prognosis through spectral kurtosis and envelope spectral analysis. Electric Power Components and Systems. 2016;44(18):2121-2132. https://doi.org/10.1080/15325008.2016.1209705.
30. Jayasurya S, Geetha S, Abdullah AS, Mishra U. UWE-Net: A deep learning framework for underwater image enhancement integrating CBAM and Charbonnier loss. Procedia Computer Science. 2025;258:689-698. https://doi.org/10.1016/j.procs.2025.04.302.
31. Sarno Filho EP, Santos AD, Sinezio HM, Simas Filho EF, Fernandes Jr ACL, de Seixas JM. Empirical mode decomposition: Theory and applications in underwater acoustics. Journal of Communication and Information Systems. 2022;37(1):145-167. https://doi.org/10.14209/jcis.2022.16.
32. Fu L, Ma Z, Zhang Y, Wang S, Zhang L. An improved bearing fault diagnosis method based on variational mode decomposition and adaptive iterative filtering (VMD-AIF). Journal of Mechanical Science and Technology. 2023;37(4):1601-1612. https://doi.org/10.1007/s12206-023-0303-2.
33. Smith WA, Randall RB. Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study. Mechanical Systems and Signal Processing. 2015;64:100-131. https://doi.org/10.1016/j.ymssp.2015.04.021.
34. Kaplan K, Kaya Y, Kuncan M, Minaz MR, Ertunç HM. An improved feature extraction method using texture analysis with LBP for bearing fault diagnosis. Applied Soft Computing. 2020;87:106019. https://doi.org/10.1016/j.asoc.2019.106019.
35. Yoo Y, Jo H, Ban SW. Lite and efficient deep learning model for bearing fault diagnosis using the CWRU dataset. Sensors. 2023;23(6):3157. https://doi.org/10.3390/s23063157.
36. Zhou C, Li Y, Ding Z, Li S. Bearing fault diagnosis based on TimesBlock and multi-scale CNN. In: 2024 36th Chinese Control and Decision Conference (CCDC). 2024:5875-5880. https://doi.org/10.1109/ccdc62350.2024.10588076.
37. Hoang DT, Kang HJ. Rolling element bearing fault diagnosis using convolutional neural network and vibration image. Cognitive Systems Research. 2019;53:42-50. https://doi.org/10.1016/j.cogsys.2018.03.002.

Zhiyuan FENG was born on March 16, 2005, in Gansu, China. He is currently majoring in mechanical design, manufacturing, and automation at the College of Engineering, Shanghai Ocean University.
e-mail: f_zhiyuan@outlook.com