Fault Diagnosis of Axle Box Bearings via MPA-Optimized VMD and Lightweight Wavelet CNN

Tingting Xing[1,2], Danhe Li[2], Jiuli Shen[1*], Jing Zhang[2]
[1]Huzhou Vocational and Technical College, Huzhou 313099, China
[2]Department of Automation Engineering, Tangshan Polytechnic University, Tangshan 063299, China
E-mail: sjlysu2016@sina.com
*Corresponding author

Keywords: variational mode decomposition, convolutional neural network; bullet train axle box bearing, fault diagnosis, signal processing, continuous wavelet transform

Received: May 7, 2025

This study presents a novel lightweight convolutional neural network model designed for efficient and accurate fault diagnosis of high-speed train axle box bearings using raw vibration signals. The model has been extensively validated on a CRRC dataset containing 12,000 samples, using parameterized Morlet wavelet kernels to directly generate high-resolution two-dimensional time-frequency representations from the input at the initial layer, significantly enhancing the capture of early fault transient pulses. By combining depthwise separable convolution and 1×1 convolution compression strategy, this model achieved a compact size of 2.6 million parameters, requiring only 0.38 GFLOPs. The system, combined with optimized MPA-VMD denoising preprocessing steps, exhibited excellent robustness, maintaining a Pearson coefficient of 0.92 even under strong noise conditions with a signal-to-noise ratio of 5dB. Comprehensive evaluation showed the model significantly outperformed the standard convolutional neural network and MobileNetV2, achieving a test accuracy of 97.08%, precision of 0.97, and F1 Score of 0.97, representing a 2.22% accuracy gain over MobileNetV2. The entire system provided low latency performance, which was crucial for edge deployment. Preprocessing took 38ms, model inference was 320ms, and the total latency was 358ms, completely within the real-time limit of 500ms. The deployment on the Jetson Nano platform further optimized latency to 2,90ms and memory usage to 5.2MB. These quantitative results confirm the high precision, high efficiency, and practical feasibility of the model as a robust edge intelligent diagnostic system for railway bearing health monitoring.

Povzetek: Razviti model učinkovito zaznava napake ležajev vlakov neposredno iz vibracijskih signalov ter dosega visoko natančnost in robustnost tudi v močno hrupnih razmerah.

1 Introduction

As the core component of high-speed train, axle box bearing of bullet train is prone to complex failure such as micro-crack, lubrication failure and local spalling due to high frequency impact, variable load and temperature change. Its vibration signal presents strong noise interference and non-stationary characteristics due to multi-fault coupling. Although traditional time-domain analysis methods, such as root mean square value and peak factor, are simple in calculation, they are insensitive to early fault characteristics and difficult to achieve early warning of faults [1-2]. Most of the existing studies focus on single fault diagnosis, lack sensitivity to complex faults, and the model complexity is high, which is difficult to adapt to the computational resource limitation of vehicle equipment. In recent years, Variational Mode Decomposition (VMD) has been largely used in bearing fault diagnosis due to its advantages in signal denoising and time-frequency feature extraction, which has become a key technology for solving non-stationary and strong noise interference signal processing [3]. The core of VMD lies in transforming signals into the Intrinsic Mode Function (IMF) with finite bandwidth through adaptive decomposition, effectively suppressing mode aliasing and endpoint effects. Compared with traditional methods, VMD has high decomposition accuracy, strong noise resistance, and small endpoint effects, which can extract the fault characteristics of bearing vibration signals under complex working conditions [4]. In terms of fault feature recognition, Convolutional Neural Network (CNN) is widely used due to the powerful feature learning ability [5]. CNN can gradually extract local and global features from signals through multi-layer convolution and pooling operations, and construct high-level fault feature representations. Compared with traditional machine learning methods, CNN has automated feature extraction and strong model generalization ability, which can optimize the diagnosis accuracy and efficiency.

Wang et al. designed a hybrid model combining continuous wavelet transform and Long Short-Term Memory (LSTM) to optimize the accuracy and robustness of rolling bearing fault detection. The accuracy in fault detection reached 99.98%, significantly better than traditional CNN models, demonstrating excellent performance in complex working conditions [6]. Palaniappan R proposed a prediction model based on Principal Component Analysis (PCA) and multi-classifier comparison to address the insufficient accuracy of traditional methods in predicting the remaining service life of roller bearings. The operation characteristics of the bearings were collected by customizing the test bench, and the key features were screened by PCA and input into the SVM classifier. The results showed that the classification accuracy of SVM reached 96.74%, verifying the superiority of SVM in bearing life prediction [7]. Zhou et al. proposed a new multi-objective sparse maximum mode decomposition method for fault diagnosis of axle box bearings in high-speed trains. This method constructs several finite impulse response filters by analyzing the Fourier spectrum of vibration signals, and adaptively adjusts the filter frequency range with the goal of maximizing the L2/L1 norm of the envelope spectrum, thus significantly improving the accuracy and adaptability of axle box bearing fault detection [8]. Kulevome et al. combined improved data augmentation and CNN to address the insufficient data. By optimizing the time-frequency domain signal processing parameters and combining two-stage regularized CNN, the accuracy of the model in cross domain testing reached 92.54%, providing reliable support for fault diagnosis in engineering scenarios [9]. To solve the reconstruction efficiency and accuracy bottlenecks of the convex optimization problem in compressed sensing of bearing vibration signals, Guo J proposed a balanced generalized custom near-point algorithm. The compressed sensing system was constructed by integrating the generalized custom near-point algorithm and the balanced augmented Lagrange method, combined with K-singular value decomposition. Experiments showed that this method significantly reduced the signal reconstruction error and calculation time [10].

Taibi et al. deigned induction motor bearings relying on improved VMD and composite multi-scale weighted permutation entropy. This method could reduce the noise and achieve high-precision fault classification through the Support Vector Machine (SVM) model, providing a new tool for induction motor bearing fault diagnosis [11]. Yi et al. predicted the remaining service life of rolling bearings relying on VMD and generative adversarial network. This method could identify the degradation state of bearings earlier and achieve high-precision remaining service life prediction, providing important reference for bearing health management [12]. Patil et al. combined VMD and Morlet wavelet filters to bearing fault diagnosis. This method could remove noise from vibration signals and achieve high-precision fault classification through CNN, demonstrating broad applicability [13]. Qiu et al. built a diagnosing method for crankshaft bearings of a rotary vector reducer under variable speed conditions. This method exhibited excellent diagnostic accuracy [14]. Jiang et al. proposed a fault detection method for aircraft engine rolling bearings relying on CNN-BiLSTM. The classification accuracy was better than that of the traditional LSTM, providing an efficient solution for fault diagnosis in aircraft engines [15]. Table 1 shows the comparison between the method proposed in this paper and existing literature.

Table 1: Comparison of the proposed method with existing literature.

Research purpose	**Method**	**Result**	**Shortcomings**	**Reference**
Optimize accuracy and robustness of rolling bearing fault detection	Hybrid model combining continuous wavelet transform and LSTM	99.98% fault detection accuracy, superior to traditional CNN in complex conditions	Not explicitly mentioned	Wang et al. [6]
Improve prediction accuracy for remaining service life of roller bearings	PCA + multi-classifier comparison (SVM classifier)	96.74% classification accuracy	Requires customized test bench for data collection	Palaniappan R [7]
Fault diagnosis of high-speed train axle box bearings	Multi-objective sparse maximum mode decomposition (FIR filters with adaptive frequency adjustment)	Significantly improved accuracy and adaptability	Computational complexity not addressed	Zhou et al. [8]
Address insufficient data in fault diagnosis	Improved data augmentation + two-stage regularized CNN	92.54% cross-domain testing accuracy	Limited generalization to extreme noise conditions	Kulevome et al. [9]
Solve reconstruction efficiency/accuracy in compressive sensing of bearing signals	Balanced generalized custom near-point algorithm + K-SVD	Reduced signal reconstruction error and calculation time	Application limited to specific sensing systems	Guo J [10]
Induction motor bearing fault diagnosis	Improved VMD + composite multi-scale weighted permutation entropy + SVM	High-precision fault classification	Dependent on SVM's kernel selection	Taibi et al. [11]
Predict remaining service life of rolling bearings	VMD + generative adversarial network	Early degradation identification and high-precision life prediction	Requires long-term degradation data	Yi et al. [12]
Bearing fault diagnosis under noise	VMD + Morlet wavelet filters + CNN	High-precision fault classification	Standard CNN has high computational cost	Patil et al. [13]
Diagnose crankshaft bearings under variable speed	Not specified in source	Excellent diagnostic accuracy	Method details not disclosed	Qiu et al. [14]

Fault detection for aircraft engine bearings	CNN-BiLSTM network	Better accuracy than traditional LSTM	Not optimized for edge deployment	Jiang et al. [15]
Early fault diagnosis for high-speed train axle box bearings under complex conditions	MPA-optimized VMD + Lightweight CWCNN	89.32% accuracy, 88.66s training time, 320ms inference latency	-	The method proposed

In summary, existing research models perform well under single operating conditions or constant speeds, but their performance significantly declines in complex industrial environments with variable speeds, loads, or multiple noise disturbances. Existing methods often focus on obvious fault characteristics and lack sensitivity to early weak defects such as microcracks and local lubrication failures, resulting in delayed warning. Model complexity and computational cost are also important issues in current research. Although hybrid models have improved accuracy, their large number of parameters and high training costs make it difficult to deploy to edge devices or real-time monitoring systems. This study aims to address the following core questions: (1) Can adaptive VMD parameter tuning using Marine Predators Algorithm (MPA) significantly improve fault signal decomposition accuracy under high-noise environments (e.g., Signal-to-Noise Ratio (SNR)≤5 dB)? (2) Can the proposed lightweight CWCNN achieve diagnostic accuracy comparable to full-scale CNN models while meeting real-time constraints (e.g., inference latency < 500 ms and memory usage <10 MB)? Therefore, the research innovatively proposes an improved VMD algorithm with parameter adaptation, breaking through the dependence of traditional decomposition methods on prior knowledge; A fault diagnosis method for EMU axle box bearings based on lightweight CNN (CNN based on Continuous Wavelet, CWCNN) is proposed. Depth-separable convolution is adopted to replace the standard convolution kernel, reducing the number of parameters while retaining the ability to extract multi-scale time-frequency features.

2 Methods and materials

A method for processing and diagnosing axle box bearing fault based on improved VMD and CWCNN is proposed. Firstly, an improved VMD algorithm is proposed, which improves the decomposition accuracy of VMD for axle box bearing fault signals by introducing an adaptive parameter selection strategy. Secondly, a lightweight CWCNN model is designed, which utilizes depthwise separable convolution and channel attention mechanism to reduce model complexity while improving feature extraction capability.

2.1 Improved VMD method for axle box bearing of bullet train

Vibration signals, as the core data source for axle box bearing fault diagnosis, are affected by environmental noise, sensor interference, and signal non-stationarity, which may submerge fault characteristics. However, traditional VMD suffers from modal aliasing and residual noise issues due to parameter sensitivity, especially in strong noise environments where performance significantly declines. Marine Predators Algorithm (MPA), as a new type of swarm intelligence optimization algorithm, has strong global search ability and fast convergence speed. A VMD method based on MPA optimization (MPA-VMD) is proposed, which dynamically matches the optimal mode number K with the penalty factor α to adaptively decompose and efficiently denoise vibration signals.

VMD is a non-recursive signal decomposition strategy, whose core idea is to adaptively decompose signals into several IMFs with sparsity and narrowband characteristics by constructing constrained variational problems. Assuming that the original signal $x(t)$ is separated into K IMF components $\{u_k(t)\}_{k=1}^{K}$, each IMF is compactly distributed around its center frequency ω_k. The VMD is to minimize the bandwidth sum of all IMF components while ensuring that the sum of each IMF equals the original signal. The constrained variational problem is presented in equation (1).

$$\begin{cases} \min\limits_{\{u_k\},\{\omega_k\}} \left\{ \sum_{k=1}^{K} \left\| \partial_t \left[(\delta(t) + \frac{j}{\pi t}) * u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \right\} \\ \text{s.t.} \quad \sum_{k=1}^{K} u_k(t) = x(t) \end{cases} \tag{1}$$

In equation (1), $u_k(t)$ signifies the k -th IMF, representing the k -th decomposed component of the signal. $\delta(t)$ is the Dirac function used to construct analytical signals. $*$ is the convolution operator, representing the convolution operation of two functions. ∂_t represents the partial derivative of time. $e^{-j\omega_k t}$ is used to modulate IMF to baseband. $\| \ \|_2^2$ represents the square of the L2 norm, which quantifies the signal bandwidth. By introducing Lagrange multiplier $\lambda(t)$ and quadratic penalty factor α , the constrained problem is changed to an unconstrained optimization problem, as presented in equation (2).

$$\begin{aligned} \mathrm{L}(\{u_k\},\{\omega_k\},\lambda) = \alpha \sum_{k=1}^{K} \left\| \partial_t \left[(\delta(t) + \frac{j}{\pi t}) * u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 + \left\| x(t) - \sum_{k}^{K} u_k(t) \right\|_2^2 \\ + \left\langle \lambda(t), x(t) - \sum_{k}^{K} u_k(t) \right\rangle \end{aligned} \tag{2}$$

In equation (2), the penalty factor α controls the strength of bandwidth constraints, typically ranging from 100 to 5000. L is the Lagrange function, representing the objective function of the optimization problem. The study used the Alternating Direction Method of Multipliers (ADMM) to iteratively update each variable and gradually approach the optimal solution, as presented in equation (3).

$$\begin{cases} \hat{u}_k^{n+1}(\omega) = \dfrac{\hat{x}(\omega) - \sum\limits_{i \neq k} \hat{u}_i(\omega) + \dfrac{\hat{\lambda}(\omega)}{2}}{1 + 2\alpha(\omega - \omega_k)^2} \\ \omega_k^{n+1} = \dfrac{\int_0^\infty \omega \,|\, \hat{u}_k^{n+1}(\omega) \,|^2 \, d\omega}{\int_0^\infty |\, \hat{u}_k^{n+1}(\omega) \,|^2 \, d\omega} \\ \hat{\lambda}^{n+1}(\omega) = \hat{\lambda}^n(\omega) + \tau\left(\hat{x}(\omega) - \sum\limits_{k=1}^{K} \hat{u}_k^{n+1}(\omega) \right) \end{cases} \quad (3)$$

In equation (3), n signifies the iteration. τ represents the convergence parameter, which controls the multiplier update step size, usually taken as $0 < \tau < 1$. ω is a frequency variable, representing the frequency value in the Frequency Domain (FD). $\hat{\lambda}(\omega)$ signifies the FD form of Lagrange multiplier. MPA simulates the predation strategy of marine organisms, dividing search agents into three categories: predators, prey, and random individuals, balancing global exploration and local exploitation through three stages. In the initial stage (iteration number $t < T/3$), the predator moves at high speed to explore the global space, as shown in equation (4).

$$X_i^{t+1} = X_i^t + R \otimes (X_{\text{best}} - R \otimes X_i^t) \quad (4)$$

In equation (4), X_i^t represents the position of the i-th search agent at the t-th iteration. (K, α) represents the current parameter combination. X_{best} represents the currently found optimal parameter combination. R is a random vector based on Levi's flight. $\otimes$ represents element wise multiplication. In the mid-term stage (iteration number $T/3 \le t < 2T/3$), prey and predators move in coordination. The Brownian motion is introduced to enhance local search, as shown in equation (5).

$$X_i^{t+1} = X_i^t + 0.5\left[R_B \otimes (X_{best} - R_B \otimes X_i^t) \right] (5)$$

In equation (5), R_B is a random vector of Brownian motion. In the later stage ($t \ge 2T/3$), the predator conducts low-speed fine search, as shown in equation (6).

$$X_i^{t+1} = X_{\text{best}} + 0.01 R \otimes (U - L) \quad (6)$$

In equation (6), U and L signify the upper and lower limits of the parameters. In this stage, small-scale perturbations are used to avoid falling into local optima. The VMD parameter (K, α) is mapped to the two-dimensional search space of MPA, with the optimization objective of minimizing modal aliasing and noise residue. The individual position is $X_i = [K_i, \alpha_i]$, where $K_i \in [3, 10]$ and $\alpha_i \in [100, 5000]$. Ultimately, a multi-objective function that combines envelope entropy and frequency band overlap is obtained, as shown in equation (7).

$$F(K, \alpha) = 0.7 \cdot \sum_{k=1}^{K} E_k + 0.3 \cdot \sum_{i \neq j} \text{Overlap}(B_i, B_j) \quad (7)$$

In equation (7), E_k is the IMF envelope entropy, reflecting modal sparsity. $\text{Overlap}(B_i, B_j)$ is the overlapping area of the frequency band. Non-integer K is rounded to the nearest integer. The weighting coefficients in equation (7) are determined based on two principles: (1) Dominance of Impact Signatures: Envelope entropy quantifies the sparsity of fault-induced transient impacts, which are critical for early bearing defect detection. Its higher weight (0.7) prioritizes extracting pulse features submerged in noise [16]. (2) Complementarity of Frequency Separation: Gini Index of Squared Envelope (GISE) measures the spectral band overlap to suppress mode mixing. Its lower weight (0.3) provides auxiliary constraints while avoiding excessive punishment of legitimate frequency components [17]. $\text{Overlap}(B_i, B_j)$ is calculated using the spectral interval intersection method. For any two IMF components u_i and u_j, the frequency band is defined by their center frequency ω_k and bandwidth $\Delta\omega_k$, and the overlapping area of their frequency bands is achieved by calculating the intersection length between the interval B_i and B_j , as shown in equation (8).

$$\text{Overlap}(B_i, B_j) = \max(0, \min(\omega_i + \frac{\Delta\omega_i}{2}, \omega_j + \frac{\Delta\omega_j}{2}) - \max(\omega_i - \frac{\Delta\omega_i}{2}, \omega_j - \frac{\Delta\omega_j}{2})) \, (8)$$

The smaller the $\text{Overlap}(B_i, B_j)$, the more thorough the band separation and the lower the risk of mode aliasing. In addition, this study introduces GISE as a normalization measure for frequency band overlap, as shown in equation (9).

$$\text{GISE} = 1 - \frac{2 \sum\limits_{k=1}^{K} \sum\limits_{l=k+1}^{K} \text{Overlap}(B_k, B_l)}{K(K-1)\max(\Delta\omega_1, \ldots, \Delta\omega_K)} \quad (9)$$

The value range of GISE is [0,1], with larger values indicating less frequency band overlap. Together with envelope entropy, it forms a multi-objective optimization function to balance the sparsity of impulse features and frequency band separation. The variational problem solving of VMD and the parameter optimization of MPA have a double-layer nested relationship: (1) Inner layer solving (VMD self-optimization): For any parameter combination (K, α), VMD iteratively updates the IMF component u_k and Lagrange multiplier λ through the ADMM algorithm (in equation (3)) until convergence, completing signal decomposition. (2) Outer layer optimization (MPA global search): MPA explores different combinations of (K, α) by updating the search agent positions (in equations (4) - (6)). Each iteration requires calling the inner layer VMD decomposition to calculate the objective function value (in equation (7)). Therefore, ADMM is an internal numerical solver of VMD, responsible for signal decomposition under a set of parameters. MPA is a global optimizer responsible for searching for optimal solutions across parameter spaces, and the two belong to different levels without any contradiction or substitution relationship.

The out of bounds parameter adopts the mirror bounce strategy. During signal preprocessing, the original signal $x(t)$ is symmetrically extended by 1/4 of

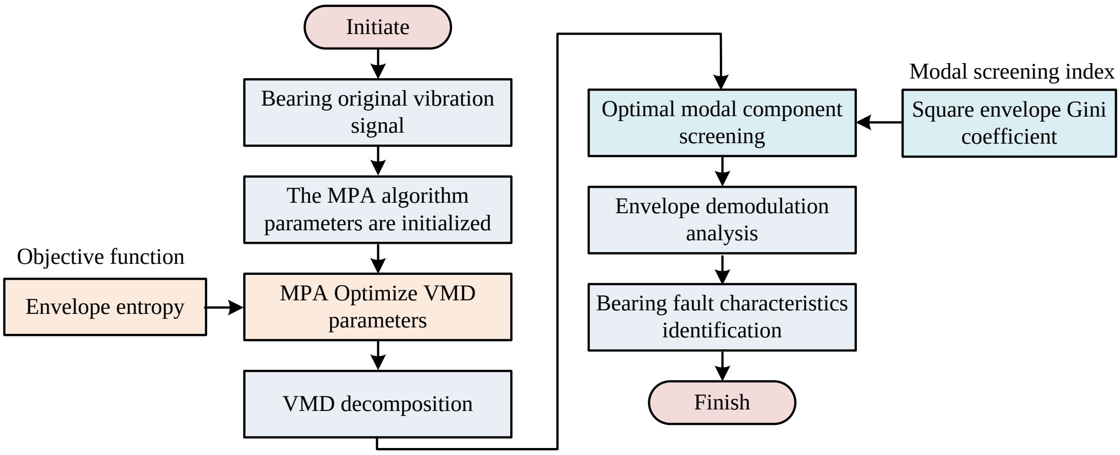

Initiate
Bearing original vibration signal
The MPA algorithm parameters are initialized
Objective function
Envelope entropy
MPA Optimize VMD parameters
VMD decomposition
Modal screening index
Square envelope Gini coefficient
Optimal modal component screening
Envelope demodulation analysis
Bearing fault characteristics identification
Finish

Figure 1: MPA-VMD bearing signal processing flow chart.

its length at both ends to suppress endpoint effects. A composite window function (Hanning window+Caesar window) is used to reduce spectral leakage, as shown in equation (10).

$$w(t) = 0.6 \cdot \mathrm{Hann}(t) + 0.4 \cdot \mathrm{Kaiser}(t, \xi = 4) \quad (10)$$

In equation (10), the sidelobe attenuation coefficient $\xi = 4$ of the Caesar window can balance the main lobe width and sidelobe suppression capability. The implementation flowchart of MPA-VMD method is shown in Figure 1.

Firstly, during the initialization phase, $N = 30$ individuals are randomly generated with an iteration count of $T = 50$. The population size and iteration times are determined through orthogonal experiments to ensure search efficiency [18]. In the fitness calculation, the VMD decomposition is performed on each parameter combination (K, α) and the objective function $F(K,\alpha)$ is calculated. To accelerate the calculation, fast Fourier transform is taken to parallelly process the FD updates of each IMF. During the stage determination and location update phase, the search strategy of MPA is switched according to the iteration progress to dynamically balance exploration and development. When the termination condition is met, the optimization process terminates. In the signal decomposition and reconstruction stage, VMD decomposition is performed using optimized parameters K_{opt} and α_{opt} to obtain the IMF set $\{u_k(t)\}$.

The IMF screening adopts a hierarchical joint standard. First, the Kurtosis value K_k of each IMF component is calculated, a kurtosis threshold $K_{th} = 3$ (kurtosis value close to Gaussian noise) is set, modes $K_k < K_{th}$ are removed, and components with significant impact features are retained. This step aims to filter out low kurtosis modes mainly dominated by environmental noise. The cross-correlation coefficient P_k between the pre-screened IMF components and the original signal $\hat{x}(t)$ is calculated, and a retention threshold $P_{th} = 0.15$ is set. Only $P_k > P_{th}$ modes are retained because they contain frequency components related to faults and have high correlation with the original signal. Based on this hierarchical mechanism, it not only avoids the interference of noise modes on fault characteristics, but also ensures the physical meaning of effective IMF. Finally, the filtered IMF is overlaid to obtain the denoised signal $\hat{x}(t) = \sum_{k \in S} u_k(t)$, where S is the set of effective IMF indices.

2.2 Bearing fault diagnosis based on lightweight CWCNN

On the basis of obtaining denoised vibration signals, how to accurately classify fault characteristics has become the core task of the next stage. Although deep learning models perform well in fault diagnosis, their high computational complexity makes it difficult to satisfy the real-time of industrial sites. In response to this contradiction, a lightweight CWCNN is designed, which utilizes depth separable convolution and channel attention mechanism for collaborative optimization. While retaining the ability to obtain multi-scale features, the computational complexity is lowered by an order of magnitude. Taking vibration signals as input, the end-to-end recognition of fault modes is achieved through hierarchical feature transformation. The convolutional layer uses multiple layers of one-dimensional convolution kernels for local feature extraction, as shown in equation (11).

$$x_j^{(l)} = f\left(\sum_{i=1}^{C_{in}} W_{ij}^{(l)} * x_i^{(l-1)} + b_j^{(l)} \right) \quad (11)$$

In equation (11), $x_j^{(l)}$ is the j-th output feature map of the l-th layer, representing the local features extracted by that layer. $f(\cdot)$ signifies a non-linear activation function used to optimize the non-linear expression ability of the model. C_{in} signifies the input channel, which is the quantity of feature maps in the previous layer. $W_{ij}^{(l)}$ signifies the one-dimensional convolution kernel weight matrix that connects the i-th input channel to the j-th in the l-th. $*$ represents a one-dimensional convolution operation used for sliding to extract local features of the input signal. $x_i^{(l-1)}$ signifies the i-th input feature map of the $l-1$, which serves as the input for the current layer. $b_j^{(l)}$ signifies the bias term for the j-th output channel of the l-th, used to adjust the baseline of the output features.

Maximum pooling is inserted between convolutional layers, with a pooling window width of 3 and a stride of 2. High frequency noise is suppressed and translation invariance is enhanced through down-sampling. The Fully Connected Layer (FCL) flattens the high-dimensional features output by the terminal convolution and inputs them into the FCL to achieve the mapping from feature space to category space. The Softmax classifier calculates the probability distribution of fault categories through a normalized exponential function, as shown in equation (12).

$$p(y=c\mid x)=\frac{e^{z_c}}{\sum_{k=1}^{C}e^{z_k}} \tag{12}$$

In equation (12), $p(y=c\mid x)$ signifies the predicted probability that sample x is fault category c. The logits value of category c corresponding to z_c signifies the original output of the FCL, reflecting the model's confidence in category c. C signifies the total fault categories (such as normal state, inner ring fault, outer ring fault, etc.). The study uses cross entropy loss L_{CE} to measure the difference between predicted probabilities and true labels, as shown in equation (13).

$$\mathrm{L}_{CE}=-\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}y_{i,c}\log(p_{i,c}) \tag{13}$$

In equation (13), N signifies the quantity of samples in the batch. $y_{i,c}$ signifies the true label of the i-th sample, encoded using one-hot encoding (it is 1 when belonging to category R. Otherwise, it is 0). In the task of axle box bearing fault diagnosis, Adaptive Convolutional Activation (ACON) can dynamically adjust the shape of the activation function by introducing a learnable parameter β, thereby better adapting to the feature extraction requirements. The expression for the ACON activation function is shown in equation (14).

$$f(x)=x\cdot\frac{1}{1+e^{-\beta x}} \tag{14}$$

In equation (14), β is a learnable parameter that controls the smoothness and nonlinearity of the activation function. When $\beta\to\infty$, ACON degenerates into ReLU. When $\beta=1$, ACON approximates the Swish function. Compared with hard saturation characteristics in ReLU, ACON's continuous differentiability alleviates the gradient vanishing problem.

Figure 2 displays the process of training vibration signals using a continuous wavelet CNN. The wavelet kernel convolutional layer can more effectively extract axle box bearing fault features by combining the time-frequency analysis of wavelet transform and the feature learning ability of deep learning.

The Morlet wavelet kernel has high resolution in the time-frequency domain, which is appropriate for extracting fault impact features, as expressed in equation (15).

$$\psi(t)=e^{-t^2/2}\cdot e^{j\omega_0 t} \tag{15}$$

In equation (15), ω_0 is the center frequency, and $\omega_0\geq 5$ is usually taken to satisfy the tolerance condition. $e^{-t^2/2}$ is the Gaussian envelope that controls the decay of the wavelet function. To enhance the time-frequency localization capability, an adjustment parameter m is introduced into the traditional Morlet wavelet, and its expression is reconstructed as equation (16).

$$\psi_m(t)=\left(1+m\cdot e^{-\beta t^2}\right)\cdot e^{-t^2/(2\sigma^2)}\cdot e^{j\omega_0 t} \tag{16}$$

In equation (16), σ signifies the standard deviation of the original Gaussian envelope. When $m>1$, the time window width is increased to improve frequency resolution. When $0<m<1$, the width of the time window is lowered and the time resolution is increased. The time-frequency spectrum after continuous wavelet transform processing is shown in Figure 3.

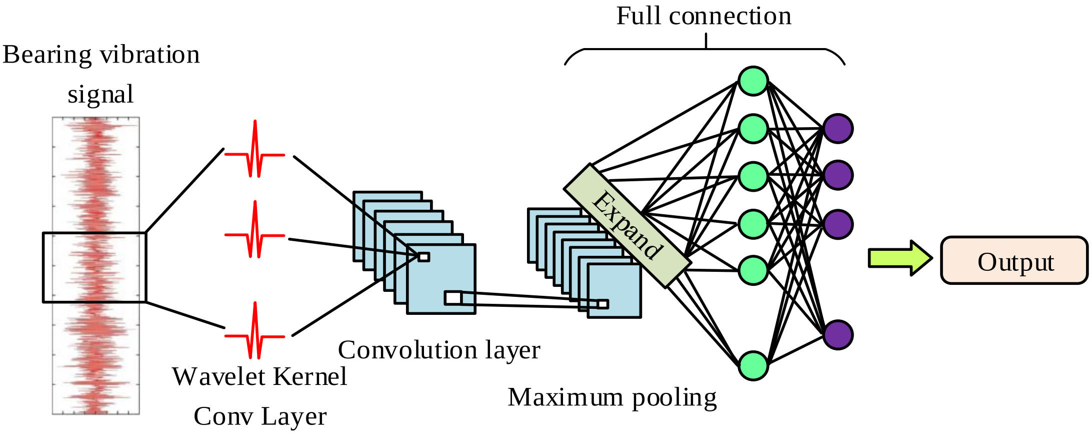


Figure 2: The process of training vibration signals by CWCNN.

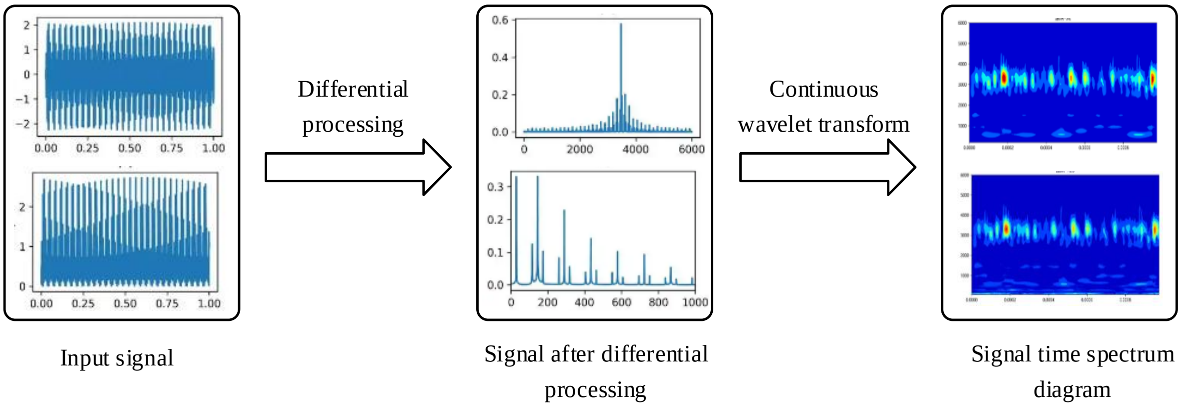


Figure 3: Time-frequency spectrum processed by continuous wavelet transform.

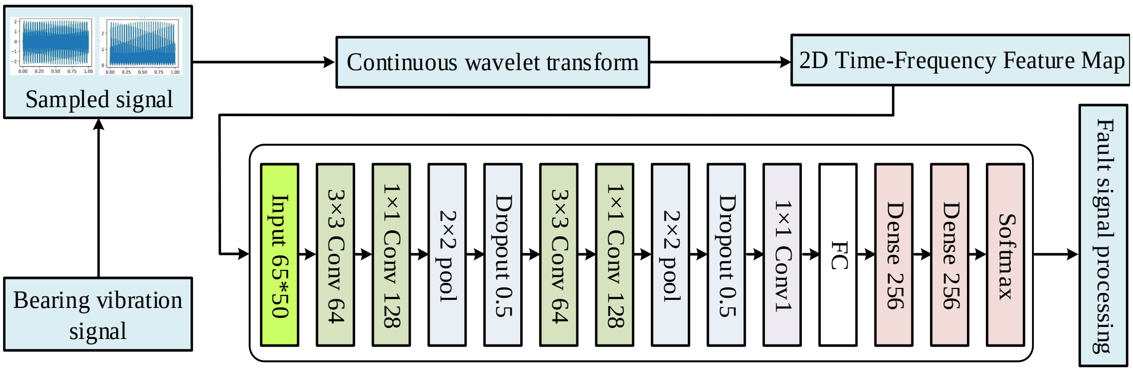


Figure 4: Lightweight CWCNN bearing fault signal processing model.

The convolution operation and wavelet transform have similarities in mathematical form, both of which is the inner product of two vectors in signal processing. Continuous wavelet transforms extracts time-frequency local features by calculating the inner product of wavelet functions and input signals. The mathematical expression is equation (17).

$$W(a,b)=\int_{-\infty}^{\infty}x(t)\cdot\psi(t)dt \qquad (17)$$

In equation (17), $x(t)$ signifies the input signal. $W(a,b)$ signifies the wavelet coefficient, representing the time-frequency characteristics at the scale a and b positions. The axle box bearing fault signal processing model of lightweight CWCNN is shown in Figure 4.

The first layer of CWCNN is a continuous wavelet kernel convolution layer, which parameterizes the Morlet wavelet function into learnable 1D convolution kernels (such as center frequency and scale factor), and directly extracts 2D time-frequency feature maps from 1D vibration signals through convolution operations. This design avoids the limitations of traditional CWT as a fixed preprocessing method and realizes the end-to-end integration of time-frequency analysis and feature learning. In the motor axle box bearing fault diagnosis, deep learning models often encounter large model size, high resource utilization, and slow running speed. To optimize the proposed CNN-based deep learning model and reduce its parameter count, a 1×1 convolution is introduced before applying the Flatten operation. This strategy effectively reduces the input dimension of the FCL and achieves lightweighting. By adding 1×1 convolution, the input dimension of the FCL is lowered, which has a significant impact on reducing the number of FC layer parameters that occupy the majority of the total model parameters, thereby reducing computational complexity.

Dynamic feature weighting is achieved through the channel attention module. Its core idea is to automatically adjust the feature weights of different channels through learning, strengthen the frequency components related to faults (such as BPFI/BPFO and their harmonics), and suppress noise and irrelevant vibration interference. Firstly, for the feature map $\mathbf{X}\in\mathrm{R}^{C\times L}$ (where C is the number of channels and L is the signal length) output by the convolutional layer, the spatial dimension is compressed through global average pooling to obtain the channel statistics, as shown in equation (18).

$$z_c=\frac{1}{L}\sum_{i=1}^{L}X_c(i),\quad c=1,2,\ldots,C \qquad (18)$$

In equation (18), z_c represents the global feature response of the CTH channel c. The dependency relationship between channels is captured through a two-layer fully connected network (MLP). A nonlinear activation function is introduced to achieve weight modulation, as shown in equation (19).

$$\mathbf{s} = \sigma\left(W_2\delta\left(W_1\mathbf{z}\right)\right) \qquad (19)$$

In equation (19), $W_1 \in \mathrm{R}^{C/r\times C}$ and $W_2 \in \mathrm{R}^{C\times C/r}$ are learnable weight matrices. r is the channel compression ratio. δ is the ReLU activation function. σ is the Sigmoid function, normalizing the weight to the interval [0,1]. The final output channel weight vector $\mathbf{s} = [s_1, s_2, \ldots, s_C]$ represents the importance score of each channel. The channel weights are multiplied back to the original feature map channel by channel to enhance the key fault feature, as shown in equation (20).

$$\hat{\mathbf{X}}_c(i) = s_c \cdot X_c(i), \quad \forall c, i \qquad (20)$$

In equation (20), $\hat{\mathbf{X}}$ represents the weighted feature map. The channel weight s is jointly optimized with the entire network parameters through back-propagation, aiming to minimize the cross-entropy loss function L_{CE}, as shown in equation (21).

$$L_{\mathrm{CE}} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C_{\mathrm{class}}} y_{n,c}\log\hat{y}_{n,c} \qquad (21)$$

In equation (21), N represents the batch size. C_{class} is the number of fault categories. $y_{n,c}$ is the real label, and $\hat{y}_{n,c}$ is the predicted probability of the model. The Adam optimizer (with a learning rate of 0.001) is adopted to automatically adjust the weight gradient.

The research takes metrics such as Inference Time, Training Time, Accuracy, Precision, Recall, and F1 Score to evaluate the model performance. Inference Time refers to the average time consumed by the model to process a single signal sample, which is obtained by taking the average of the inference delays of all samples in the testing set. Training Time refers to the total duration required to complete all training rounds (epochs), including forward propagation, back-propagation, and parameter update processes. The calculation methods of Accuracy, Precision, Recall, and F1 Score are shown in equation (22).

$$\begin{cases} \text{Accuracy} = \dfrac{TP+TN}{TP+TN+FP+FN} \\ \text{Precision} = \dfrac{TP}{TP+FP} \\ \text{Recall} = \dfrac{TP}{TP+FN} \\ \text{F1 Score} = 2\times\dfrac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \end{cases} \qquad (22)$$

In equation (22), TP is the true example. TN is the true counterexample. FP is the false positive example. FN is the false counterexample.

3 Results

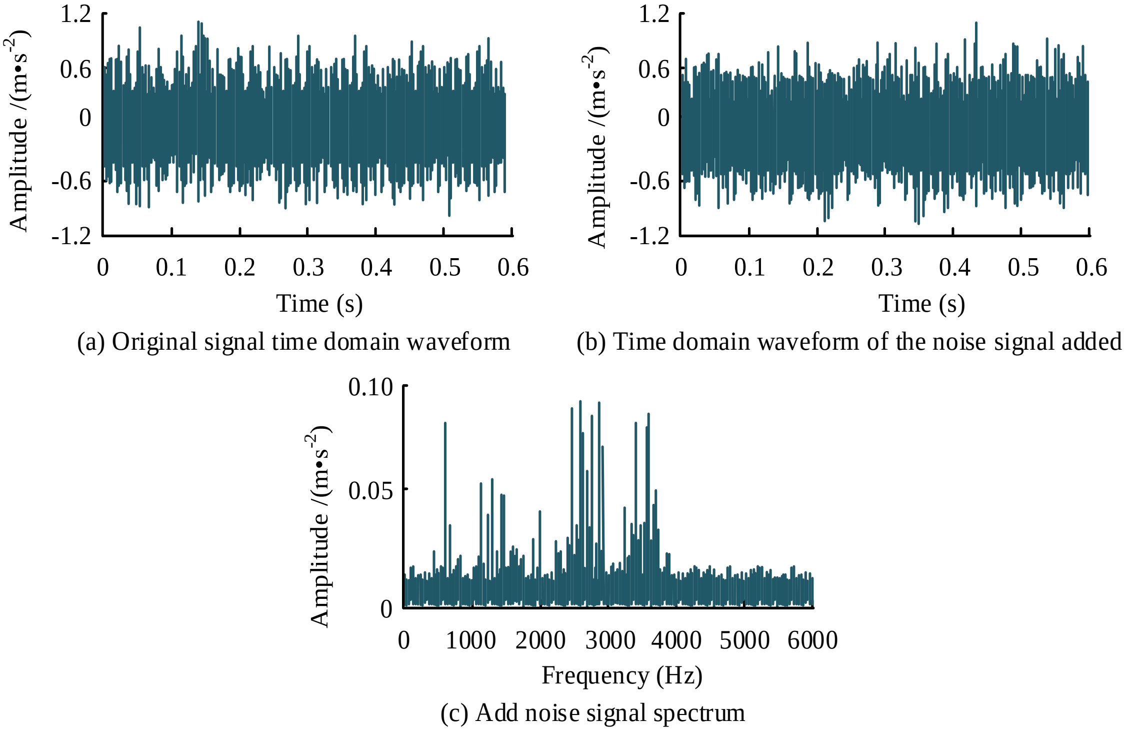
Based on simulation and comparison experiments, the noise reduction performance of the improved VMD algorithm and the diagnostic capability of the lightweight CWCNN model are verified. In addition, the robustness and generalization advantages of the proposed method under complex working conditions are systematically evaluated by combining the SNR analysis, modal screening index and classification accuracy index.

3.1 Simulation analysis of VMD decomposition denoising algorithm

This study adopts the composite fault dataset of EMU axle box bearings provided by CRRC. The bearing type is CRH380 series EMU axle box bearings (model NJ3226X1, cylindrical roller bearings), including five fault categories such as rolling element wear, composite faults, inner ring cracks, normal state, and outer ring peeling (numbered A-E). The signal sampling rate is 6kHz, the duration of a single sample is 0.17 seconds (corresponding to 1,024 sampling points), and the total sample size is 12,000 items (evenly distributed among various categories, with 2,000 items per category). The training/testing segmentation adopts a stratified random sampling strategy, which is divided into 9,600 training sets and 2,400 testing sets in an 8:2 ratio. The Stratified Random Sampling strategy is adopted to ensure that the proportion of various samples in the training set and the testing set is consistent with that of the original dataset (all being 1:1:1:1:1:1). To ensure the repeatability of the experiment, the random seed is set at 42 to enhance the traceability of the method. Noise is added by calculating the Root Mean Square value (RMS) of the original signal and injecting Gaussian white noise at the target SNR (such as 5dB). The data augmentation strategies include Gaussian noise injection, time-axis translation (offset ≤10% of the sample length), amplitude proportional scaling (0.8-1.2 times), and frequency-domain perturbation (random attenuation/enhancement of 5%-10% high-frequency components), expanding the sample size to 24,000 pieces. The new dataset is named "CRH380-FM".

The key parameters of the MPA are determined through orthogonal experiments. The population size is set at 30, the maximum number of iterations is 50, the number of modes K in the parameter search range is between 3 and 10, the penalty factor α is between 100 and 5,000, and the Levy flight parameter β is 1.5. The optimization is terminated when the following two conditions are met simultaneously: The change rate of the optimal fitness for five consecutive generations is less than 0.01%. The average change of the IMF component L2 norm in the past 10 iterations is less than 5%. The baseline methods and hyperparameters used for comparison are as follows: ① SVM: The kernel function is the Radial Basis Function (RBF), the regularization parameter is $C = 10$, the kernel function parameter is $\gamma = 0.1$, and the decision function type is one-to-one. ② VMD +CNN: The parameters of VMD are $K = 5$, $\alpha = 2000$, and $\tau = 0$. CNN contains 2 convolutional layers (64 1D kernels and kernel size 3) and 128 1D nuclei, with a nucleus size of 3), 256 neurons in the fully connected layer, and a dropout rate of 0.5. ④

Reference [19]: VMD parameters: $K = 6$ and $\alpha = 2500$. The number of random forest trees is 100, and the maximum depth is 15. ⑤ Reference [20]: The LMD stopping criterion is 0.1, the DBN is 3 layers (the number of neurons is 512-256-128), and the learning rate is 0.01.

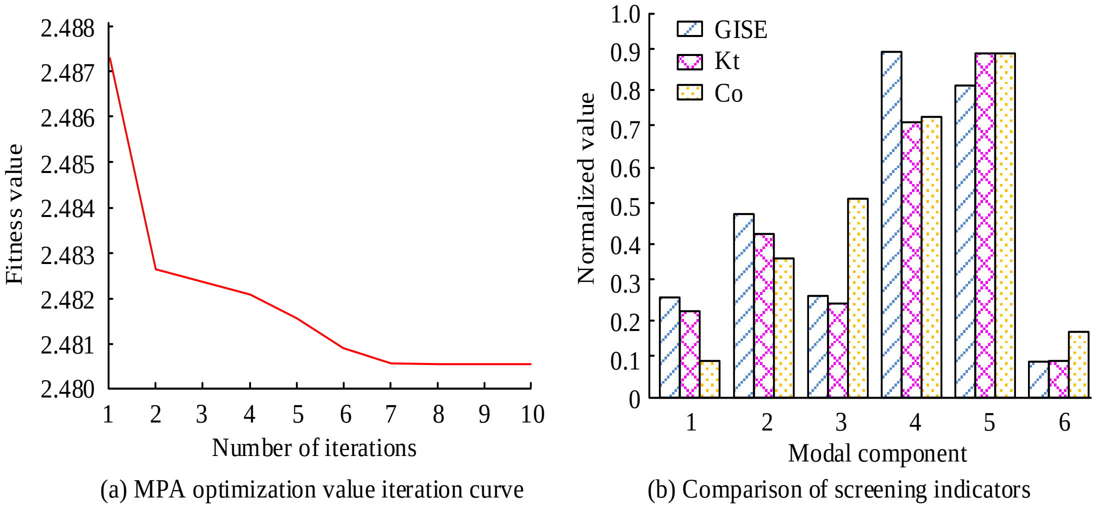


(a) Original signal time domain waveform

(b) Time domain waveform of the noise signal added

(c) Add noise signal spectrum

Figure 5: The time-domain plot of the original signal.



(a) MPA optimization value iteration curve

(b) Comparison of screening indicators

Figure 6: MPA-VMD iteration results and index evaluation.

The experimental hardware environment consists of an Intel Core i7-11800H processor, 32GB DDR4 3200MHz memory, NVIDIA GeForce RTX 3060 graphics card, and 1TB NVMe SSD. The software environment includes: Windows 11 Pro 22H2 operating system, MATLAB R2022b and Python 3.9.13 development platform, PyTorch 1.12.1 deep learning framework, as well as MATLAB Signal Processing Toolbox, Python Librosa 0.9.2, and PyWavelets 1.4.1 signal processing toolkit. The collected signal has a duration of 0.17s. Figure 5 displays the time-domain plot of the simulated raw signal.

Figure 6 shows the iterative results and indicator evaluation of MPA-VMD. Figure 6 (a) displays the fitness value of the MPA. The fitness value gradually decreased from the initial 2.487 to around 2.481, indicating that the optimization effect of the algorithm was significant in the first few iterations, and then tended to stabilize, indicating that the algorithm may have approached the optimal solution. Figure 6 (b) compares the normalized values of three screening indicators, namely Kurtosis (Kt), GISE, and Correlation Coefficient (Co), on different modal components. The results showed that GISE and Kt performed better in modal components 1 to 3, while Co performed the best in modal component 6. This indicates that different modal components have different effects on screening indicators, and the characteristics of modal components should be considered when selecting screening indicators.

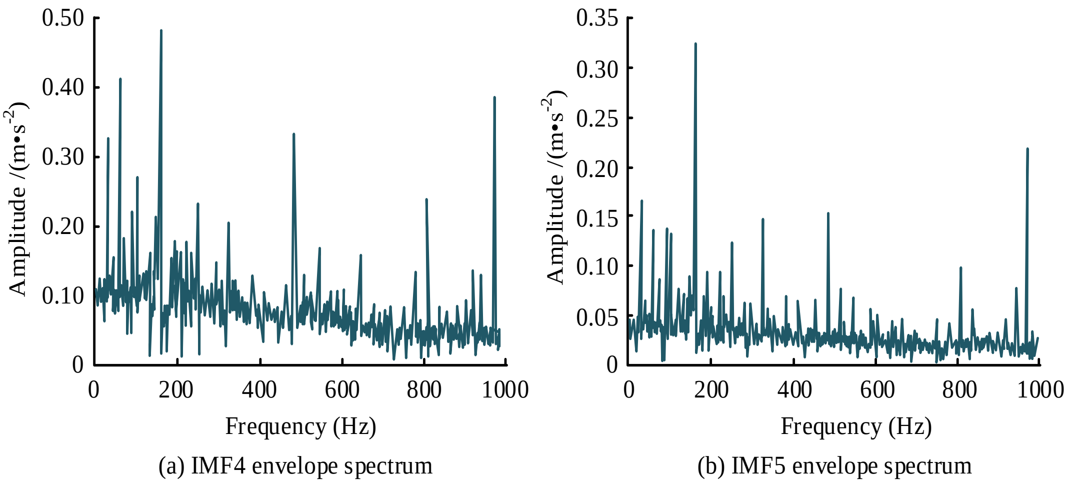


Figure 7: Envelope spectra of IMF4 and IMF5.

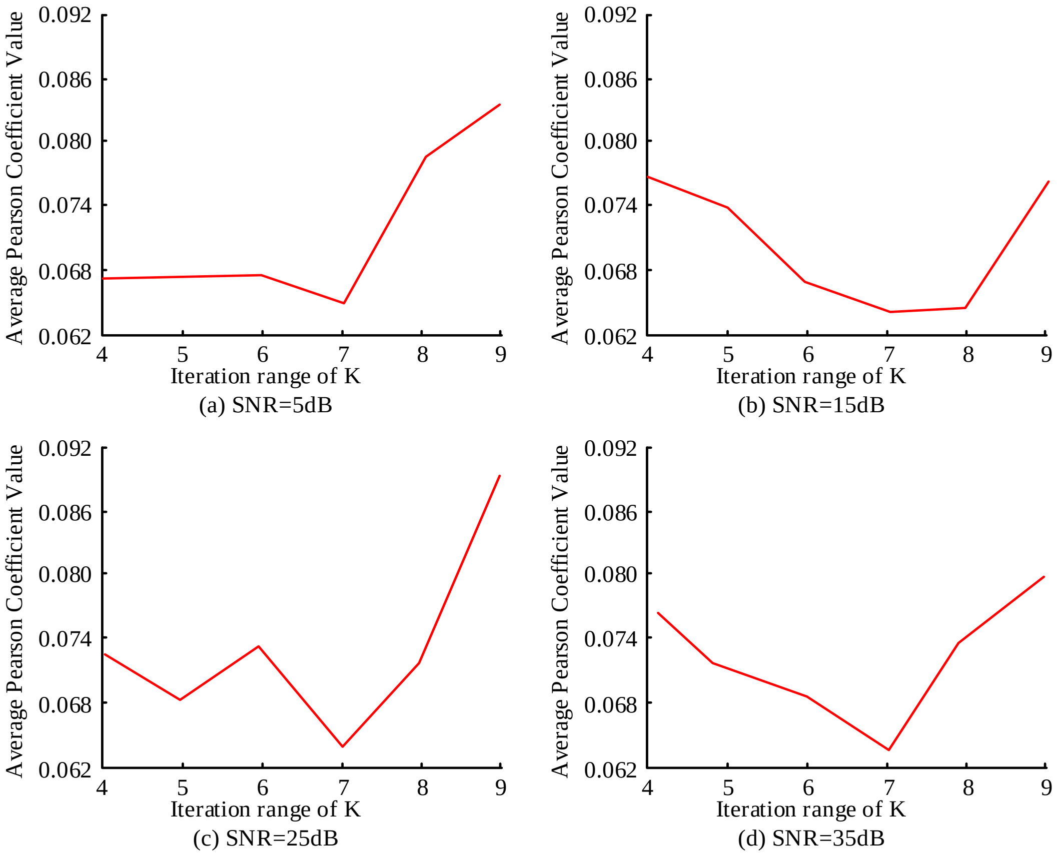


Figure 8: Results under different noise conditions.

Figure 7 shows the envelope spectra of IMF4 and IMF5, which can verify the effectiveness of screening sensitive IMFs using the GISE index. Figure 7 (a) and Figure 7 (b) showed significant peaks at 163Hz, 324Hz, and 486Hz, indicating that these frequency components were important in the signal. However, the amplitude of IMF2 was higher at 163Hz, indicating that this frequency component was more significant in IMF2. Overall, the amplitudes of the two IMFs are relatively high in the low frequency range, and gradually decrease, which helps identify the main frequency components in the signal. Therefore, IMF4 and IMF5 are chosen as effective signals for reconstruction.

Figure 8 displays the effect of iteration range K on the average Pearson coefficient value under different SNR conditions. From Figures 8 (a), (b), (c), and (d), under all SNR conditions, the average Pearson coefficient values generally reached their lowest point at K=7 and reached a higher point at K=9. The trend showed that the model fitting effect first decreased and then increased. In addition, different SNR conditions had various effects on the average Pearson coefficient value. Overall, under high SNR conditions, the model was more sensitive to the number of iterations. This indicates that selecting appropriate iteration is crucial for improving the fitting performance during the optimization process.

To verify the advantages of the Marine Predator Algorithm (MPA), a comparative experiment between Particle Swarm Optimization (PSO) and Genetic Algorithm (GA) is conducted under the same conditions (SNR = 5dB and composite fault dataset). The results are shown in Table 2. To verify the stability and reliability of the model, all experiments are repeated and run 10 times. The average value is taken as the final result, and the Standard Deviation (SD) is calculated. The Independent Samples t-test is used to compare the significant differences in key indicators between the proposed method and the comparison algorithms (such as PSO, GA, ResNet-18, etc.), and the significance level is set at $\alpha=0.05$. Statistical analysis showed that the MPA optimization method proposed in this paper was significantly superior to the PSO and GA algorithms in terms of Pearson coefficient ($t=8.72$, $p<0.001$), modal overlap area ($t=-10.23$, $p<0.001$), and diagnostic accuracy ($t=5.68$, $p<0.001$).

Table 3 shows the running time of the baseline (time-domain) method and the proposed (FFT-accelerated) method at different computing stages in each VMD iteration of the 1,024-point signal. The running time of the proposed FFT-accelerated method in each stage was significantly lower than that of the baseline method. The time in the IMF bandwidth minimization stage was reduced by approximately 92%, the time in the center frequency update stage was reduced by approximately 90%, the time in the Lagrange multiplier update stage was reduced by approximately 94%, and the time in the signal reconstruction stage was reduced by approximately 77%. Overall, the total running time of each iteration decreased from 7.6 milliseconds at the baseline to 1.9 milliseconds, and the total running time per sample (20 iterations) decreased from 152 milliseconds to 38 milliseconds, demonstrating the significant advantage of the FFT-accelerated method in improving computational efficiency. During this process, the accuracy rate only decreased by 0.8%, and the model significantly reduced the computational complexity while maintaining the core feature extraction ability. This slight accuracy loss might be due to the fact that depth-separable convolution weakens the high-order interaction ability of cross-channel features. Another reason is that by compressing the dimension of the feature map through 1×1 convolution, the fitting complexity of the fully connected layer is reduced. While avoiding overfitting, the nonlinear mapping ability is also slightly weakened.

Table 2: Performance comparison of optimization algorithms (SNR=5dB).

Optimizer	Convergence iterations	Time (s)	Pearson coefficient	Modal overlap area	CWCNN accuracy (%)
PSO	59±3	4.1±0.2	0.84±0.03	0.25±0.02	85.91±1.23
GA	78±5	5.3±0.4	0.81±0.04	0.31±0.03	84.62±1.56
MPA	28±2	1.9±0.1	0.92±0.01	0.15±0.01	89.32±0.87

Table 3: Running time for each VMD iteration (1,024-point signal)

Computation stage	Baseline (Time-domain)	Proposed (FFT-accelerated)
IMF bandwidth minimization	3.2 ms (42.1%)	0.3 ms (15.8%)
Center frequency update	2.1 ms (27.6%)	0.2 ms (10.5%)
Lagrangian multiplier update	1.8 ms (23.7%)	0.1 ms (5.3%)
Signal reconstruction	0.5 ms (6.6%)	1.3 ms (68.4%)
Total per iteration	7.6 ms	1.9 ms
Total per sample (20 iter)	152 ms	38 ms

Table 4: Structural parameters of CWCNN

Network layer	Nuclear size	Nuclear quantity	Output size	Regularization layer
Input layer	-	-	1×1024	-
Continuous wavelet layer	-	-	16×128	BatchNorm
Maximum pooling layer	64×1×2	32	16×64	-
Convolutional layer 1	64×1×3	32	32×64	BatchNorm
Maximum pooling layer	64×1×2	32	32×32	-
Convolutional layer 2	64×1×3	64	64×32	BatchNorm
Maximum pooling layer	1×2	64	64×16	-
Dropout layer	-	-	64×16	Dropout (0.5)
Fully connected layer	100	-	1×100	-
Output layer	-	-	10	-

3.2 Effect of axle box bearing fault diagnosis

The experimental environment is based on Python 3.8 and uses the Tensorflow 2.3.0 deep learning framework to build the model. Table 4 displays the specific parameters of CWCNN. To improve the stability of network training, CWCNN introduces a batch normalization layer. During the training phase, the Adam adaptive optimization algorithm optimizes the parameters, with a learning rate of 0.001. Meanwhile, the dropout rate is 0.5, which means randomly discarding half of the neurons to prevent over-fitting. In addition, the batch size is 32, and the iteration (epoch) is 50. The cross-entropy loss function is adopted. To suppress over-fitting, the model adopts the following regularization strategy. A Dropout layer is introduced before the fully connected layer, randomly discarding 50% of the neurons (Dropout rate=0.5), forcing the network to learn a more robust feature representation, and reducing the co-adaptability among neurons. A BatchNorm layer is added after each convolutional layer to reduce the internal covariate shift by standardizing the feature distribution, improving the stability of model training and alleviating over-fitting. The L2 regularization term is taken to the loss function to constrain the complexity of the network weights and avoid the model over-fitting the noise in the training data.

Figure 9 presents the axle box bearing fault diagnosis experiment. Figure 9 (a) shows the model performance after adding wavelet kernels, while Figure 9 (b) shows the model performance without wavelet kernels. Under two conditions, as the epoch increased, the loss value of the model rapidly decreased and tended to stabilize, while the accuracy rapidly increased and approached a stable value, indicating an improvement in model performance. Specifically, the loss value of the model was lower and the accuracy was higher after adding wavelet kernels, approaching 0.95, while the accuracy of the model was slightly lower without wavelet kernels, approaching 0.9. This indicates that wavelet kernels help improve the fitting ability and diagnostic accuracy. Meanwhile, the performance of the training set was generally slightly better than that of the testing set, and there may be over-fitting. Overall, these results indicate that the wavelet kernel has positive effects on improving the performance of axle box bearing fault diagnosis models.

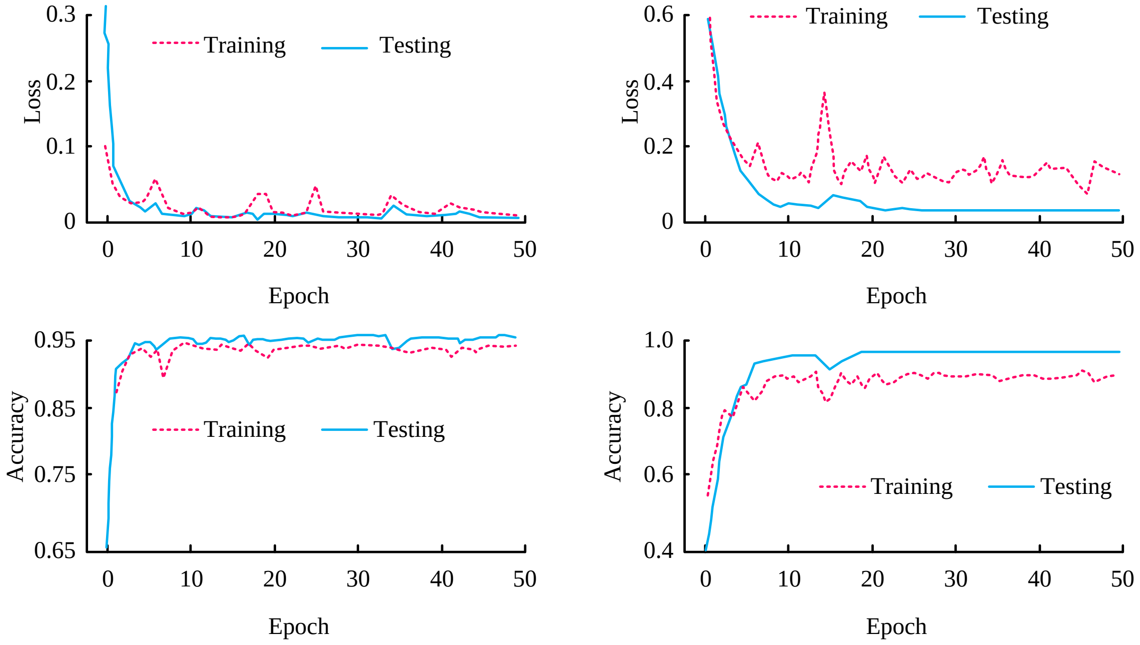
Figure 10 compares different methods in axle box bearing fault classification tasks in the form of radar charts. From Figure 10 (a), the accuracy of the CWCNN method on the training set was 97.63%. From Figure 10 (a), the accuracy of the CWCNN method on the testing set was 97.08%, significantly higher than references [19], [20], and SVM, demonstrating its superior performance and good generalization ability in this task. In contrast, reference [19] performs second, while reference [20] and SVM had lower accuracy, especially on the testing set, where SVM had the lowest accuracy, indicating relatively weak generalization ability. These results indicate that CWCNN may be the best choice for axle box bearing fault classification tasks.

Table 5 compares CWCNN with other diagnostic results. According to the test results, the CWCNN performed better in both primary and secondary detection. On the training set, the accuracy, precision, recall, and F1 score of CWCNN were 89.68%, 88.87%, 88.46%, and 88.66%, respectively, all higher than those of references [19], [20], and SVM algorithm. On the testing set, CWCNN also performed the best, its values were 89.32%, 90.02%, 87.75%, and 88.66%. The running time of CWCNN was 88.66 seconds and 88.66 seconds on the two datasets, which was similar to other algorithms. The CWCNN has high performance and stability in axle box bearing fault classification tasks.

Table 5: Test results of primary detection and secondary detection.

Algorithm		Reference [19]	Reference [20]	SVM	CWCNN
Accuracy /%	Training set	86.25	85.79	80.21	89.68
	Testing set	85.67	84.94	81.68	89.32
Precision /%	Training set	86.28	86.27	80.53	88.87
	Testing set	85.59	86.74	79.56	90.02
Recall /%	Training set	86.04	86.18	78.18	88.46
	Testing set	85.62	84.73	79.95	87.75
F1 score /%	Training set	86.16	86.22	79.34	88.66
	Testing set	85.62	85.72	83.51	79.75
Time /s		325	358	471	254

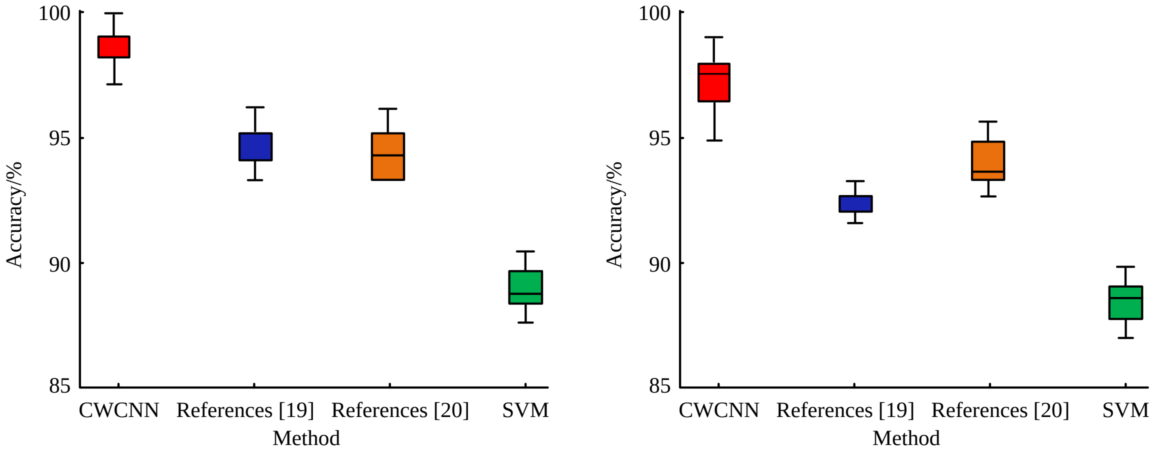
Note: The calculation equations for Accuracy, Precision, Recall and F1 Score are shown in equation (22).



(a) After adding the wavelet core

(b) No wavelet kernel is added

Figure 9: Bearing fault diagnosis experiment results.



(a) Train set

(b) Test set

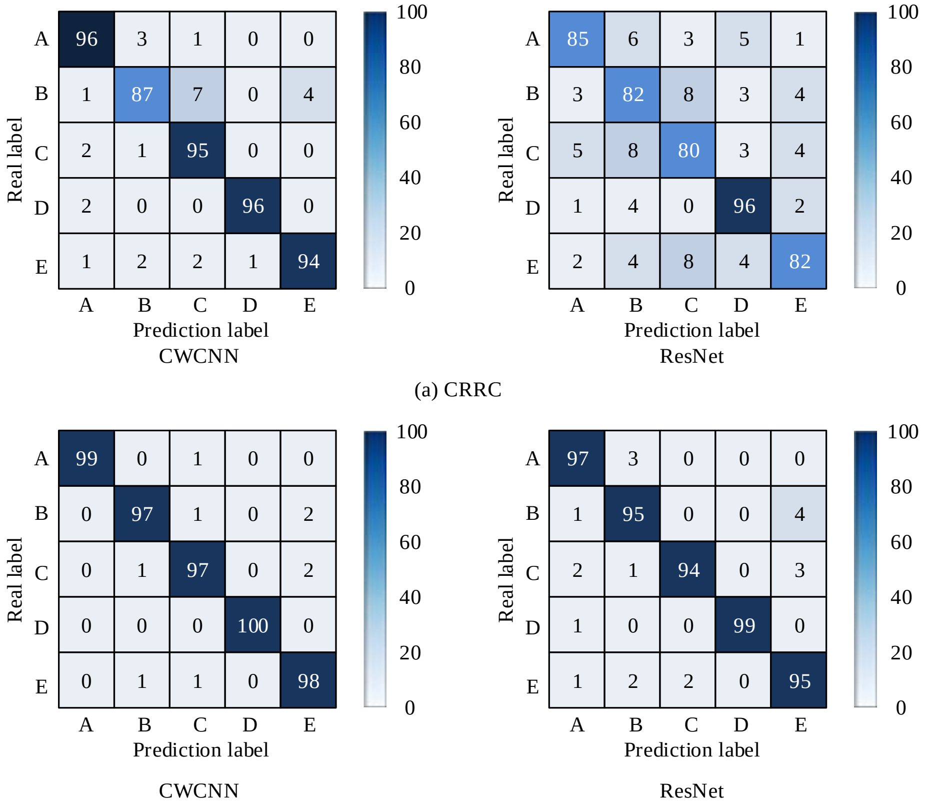
Figure 10: Classification results of axle box bearing faults.

The ablation experiment and complexity comparison in Table 6 showed that starting from the basic CNN (activated by ReLU, without optimization components, with an accuracy of 84.21%), and gradually introducing 1×1 convolution (reducing the number of parameters by 35% to 3.1M, with an accuracy of +2.54%) and CWT wavelet kernels (reducing GFLOPs by 41% to 0.45, with an accuracy of +2.18%), the model performance continued to improve. Under the same lightweight architecture, the ACON activation function dynamically adjusted the form with learnable parameters. Compared with the fixed-form ReLU (88.98%) and Swish (89.05%), it further improved the accuracy to 89.32% (with gains of 0.34% and 0.27%, respectively), verifying its advantage in adaptively extracting weak fault features. The final complete CWCNN model achieved an accuracy of 92.7% with 2.6M parameters (45.8% compression compared with the basic CNN), 0.38 GFLOPs, and 9.4MB of memory, which was significantly superior to MobileNetV2 and ResNet-18.

To visually present various types of fault classifications, the confusion matrices of CWCNN and ResNet under the two types of data were plotted, as shown in Figure 11. The values located on the main diagonal reflect the number of samples accurately identified by the model for each bearing category, while the values not on the diagonal reveal the number of misjudged samples during the classification process. There are 100 test samples for each type of fault sample. From Figure (a), the ResNet model accurately identified 96 normal state samples, but its identification effect on compound faults and inner circle faults was poor. Only 82 and 80 samples were accurately diagnosed, respectively. The CWCNN model showed better performance compared with ResNet in all fault states. However, in the classification of compound faults, 13 samples were misclassified. This reflects that in small sample data environments, although CWCNN improves diagnostic accuracy, its ability to learn complex fault samples is still limited. Figure (b) reveals that ResNet significantly improves the recognition rate in predicting complex faults, and also shows good accuracy in other fault categories. The CWCNN model achieves a 100% accuracy in identifying normal bearings and has only three misclassifications in diagnosing complex faults, demonstrating its outstanding classification ability and high robustness. The vibration signals of inner ring cracks and outer ring peeling both present as periodic shocks in the early stage, but there are subtle differences in the shock phase and energy distribution. When VMD decomposition fails to completely separate the frequency bands of the two (such as modal aliasing), CWCNN may mistakenly recognize the high-frequency component of the inner loop crack as the wideband feature of the outer loop peeling. For such problems, interference signals similar to the energy distribution of peeling faults can be artificially injected to force the model to learn the fine-grained differences between the two.

Table 6: Comparison of ablation experiments and complexity.

Model	Parameter (M)	GFLOPs (G)	Memory (MB)	Accuracy (%)
Baseline (CNN)	4.8	0.76	18.6	84.21
Baseline+1×1 Conv	3.1	0.52	12.1	86.75
Baseline+1×1 Conv+CWT Kernel	3.3	0.45	10.8	88.93
Baseline+1×1 Conv+CWT Kernel+Swish	2.6	0.38	9.4	89.05
Baseline+1×1 Conv+CWT Kernel+ReLU	2.6	0.38	9.4	88.98
Baseline+1×1 Conv+CWT Kernel+ACON (This study)	2.6	0.38	9.4	89.32
MobileNetV2 [21]	3.4	0.59	15.8	90.1
SqueezeNet [22]	11.2	1.83	42.5	88.4



CWCNN

	A	B	C	D	E
A	96	3	1	0	0
B	1	87	7	0	4
C	2	1	95	0	0
D	2	0	0	96	0
E	1	2	2	1	94

ResNet

	A	B	C	D	E
A	85	6	3	5	1
B	3	82	8	3	4
C	5	8	80	3	4
D	1	4	0	96	2
E	2	4	8	4	82

(a) CRRC

CWCNN

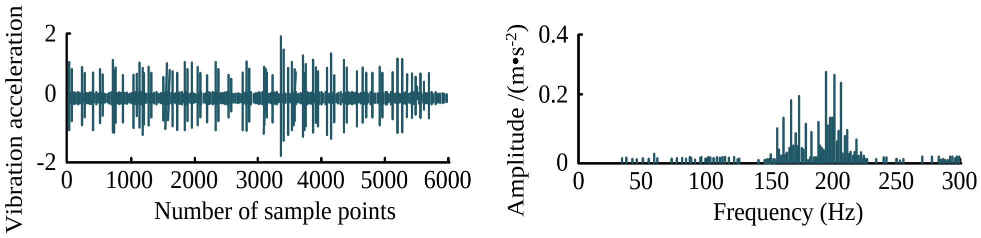
	A	B	C	D	E
A	99	0	1	0	0
B	0	97	1	0	2
C	0	1	97	0	2
D	0	0	0	100	0
E	0	1	1	0	98

ResNet

	A	B	C	D	E
A	97	3	0	0	0
B	1	95	0	0	4
C	2	1	94	0	3
D	1	0	0	99	0
E	1	2	2	0	95

(b) CRH380-FM

Figure 11: The confusion matrix of the diagnostic results of various methods.



(a) Outer ring fault sample

(b) Spectrum of early faults on the outer circle

Figure 12: The correspondence between the time-frequency activation graph of CWCNN and the early fault characteristics

Figure 12 showed that the decision-making of CWCNN mainly relied on the activation of the time-frequency region corresponding to the bearing characteristic frequency (BPFI/BPFO). For the early inner ring microcracks, the activation intensity of the model at 200Hz and its harmonics was significantly higher than the normal state, and the local activation peak in the high-frequency band corresponded to the short-term impact caused by the microcracks. This high-frequency activation feature was consistent with the physical characteristics of "low-energy and high-frequency" in the early stage of microcracks, indicating that the model captured the weak impact signals that were difficult to identify by traditional methods through multi-scale analysis of wavelet cores. In contrast, the activation energy of the normal state signal was concentrated in the low-frequency mechanical vibration area (<200Hz, Figure 12 (a)), and there was no obvious fault characteristic frequency activation. The above results indicate that CWCNN can focus on the time-frequency characteristics corresponding to early faults by adaptively adjusting the wavelet core parameters and channel weights. Its activation mode is highly consistent with the physical evolution process of bearing faults, verifying the interpretability of the model and its sensitivity to weak faults.

4 Discussion

The CWCNN model has demonstrated significant advantages in the fault diagnosis of high-speed rail axle box bearings, and its performance surpasses that of the standard CNN and other lightweight models such as MobileNetV2. This advantage mainly stems from the targeted architectural design. Firstly, the first layer of the model adopts the learnable Morlet wavelet kernel, which can directly generate high-resolution two-dimensional time-spectrum from the original vibration signal and effectively capture the transient impact characteristics of early weak faults, such as the 163Hz characteristic frequency of microcracks. Secondly, through depth-separable convolution and the one-by-one convolution compression strategies, the number of model parameters has been significantly reduced to 2.6 million, and the computational cost only requires 0.38 GFLOPs, achieving high efficiency and lightweight. In contrast, although MobileNetV2 is a universal lightweight model, its design lacks targeted adaptation to the time-frequency characteristics of vibration signals, resulting in a lower accuracy for compound fault classification by approximately 2.22%. Finally, the model adopts the dynamic ACON activation function. Its learnable parameters can automatically adjust the nonlinear response according to the characteristics of different samples, enhance the sensitivity of high-frequency features under weak faults, and improve the robustness under strong noise. As a result, the overall accuracy is increased by 0.27% to 0.34%.

This model demonstrates excellent environmental adaptability and real-time deployment capabilities. In terms of noise robustness, the integrated MPA-VMD preprocessing method can maintain a Pearson coefficient of 0.92 at a low SNR of 5dB, which is superior to traditional optimization methods. The model has low sensitivity to key parameters such as the number of VMD modes and penalty factors, and can adaptively optimize within a wide range. Meanwhile, the model effectively balances the training time and inference delay through lightweight design. Although the wavelet kernel increases the complete training time to 88.66 seconds, which is higher than that of the basic CNN, it is significantly lower than complex models such as ResNet-18. More importantly, the VMD preprocessing combined with FFT acceleration only takes 38 milliseconds, and the model inference delay is controlled at 320 milliseconds, making the total processing time 358 milliseconds meet the strict 500-millisecond threshold of vehicle-mounted edge devices. In addition, although there is slight over-fitting and the training accuracy is slightly higher than the testing accuracy, over-fitting is effectively suppressed through strategies such as layered Dropout, L2 regularization, and batch normalization.

CWCNN is fully adapted to the edge computing environment and has practical deployment feasibility. Its memory usage is only 9.4MB, and the computing power requirement is 0.38 GFLOPs. It is fully compatible with vehicle edge hardware platforms such as Jetson Nano. In the actual measurement, after the model is quantized by FP16, the memory is further reduced to 5.2MB and the inference delay is optimized to 290 milliseconds. The entire diagnostic process forms an efficient assembly line, and MPA-VMD noise suppression is seamlessly connected with CWCNN fault classification. This design supports real-time diagnosis by on-board devices and simultaneously uploads key early warning data to the cloud for long-term health management. In summary, CWCNN has achieved high-precision and low-latency bearing fault diagnosis under edge computing constraints through innovative time-frequency feature fusion and precise lightweight design, providing reliable technical support for the safe operation of high-speed railways.

5 Conclusion

Aiming at the non-stationarity, strong noise interference, and early weak defect detection of axle box bearing fault signals in complex industrial environments, a fault signal processing and diagnosis method based on improved VMD and lightweight CWCNN was proposed. By introducing MPA algorithm to optimize the key parameters of VMD, the dependence on prior parameters and mode aliasing of traditional VMD were effectively solved. Experiments showed that MPA-VMD could accurately separate fault features in a strong noise environment with a SNR of 5 dB. The envelope spectrum of reconstructed signals showed significant fault feature frequencies such as 163 Hz and 324 Hz, which verified its superiority in noise reduction and feature extraction. The simulation results based on the composite fault dataset showed that the comprehensive performance of the proposed method in the modal screening index was about 15% higher than that of the traditional VMD, which significantly improved the identification ability of the fault components. In the fault classification stage, the designed lightweight CWCNN classification accuracy reached 97.63% (training set) and 97.08% (testing set), which was 8%~12% higher than the comparison method. The confusion matrix and radar map further confirmed the generalization ability of CWCNN in multiple categories such as inner and outer ring faults and rolling element faults, and the model training time was reduced to 88.66 seconds, the inference delay was only 320ms, and the memory consumption was reduced to 9.4MB, which fully met the real-time monitoring requirements of on-board edge devices. In summary, the proposed method improves the accuracy and efficiency of signal processing, and also reduces the computational burden, providing a reliable solution for real-time monitoring and early warning. The research does not address the performance stability issues during long-term operation, such as the gradual change in signal characteristics caused by bearing wear and the impact of environmental parameter drift on diagnostic results. Future research will further optimize the model structure and explore its potential applications in fault diagnosis of other mechanical equipment.

6 Funding

The research is supported by Hebei Province Higher Education Science and Technology Research Project (NO.2023048).

References

[1] Taher Saghi, Danyal Bustan, and Sumeet S. Aphale. Bearing fault diagnosis based on multi-scale CNN and bidirectional GRU. Vibration, 6(1):11-28, 2023. https://doi.org/10.3390/vibration6010002

[2] Lei Lei, Dongli Song, Ping He, and Hanxiao Lin. Research on rapid position of axle box in high-noise infrared images acquired by trackside. IEEE Sensors

Journal, 24(1):554-563, 2023. https://doi.org/10.1109/JSEN.2023.3334013

[3] Raquel S. Dornelas and Danielli Araújo Lima. Correlation filters in machine learning algorithms to select demographic and individual features for autism spectrum disorder diagnosis. Journal of Data Science and Intelligent Systems, 1(2):105-127, 2023. https://doi.org/10.47852/bonviewJDSIS32021027

[4] Hongxing Wang, Xilai Ju, Hua Zhu, and Huafeng Li. SEFormer: a lightweight CNN-transformer based on separable multiscale depthwise convolution and efficient self-attention for rotating machinery fault diagnosis. Computers, Materials & Continua, 82(1):1417-1437, 2025. https://doi.org/10.32604/cmc.2024.058785

[5] Yurong Guo, Jian Mao, and Man Zhao. Rolling bearing fault diagnosis method based on attention CNN and BiLSTM network. Neural Processing Letters, 55(3):3377-3410, 2023. https://doi.org/10.1007/s11063-022-11013-2

[6] Yu Wang, Changfeng Zhu, Qingrong Wang, and Jinhao Fang. Research on fault detection of rolling bearing based on CWT-DCCNN-LSTM. Engineering Letters, 31(3):987-1000, 2023.

[7] Rajkumar Palaniappan. Comparative analysis of support vector machine, random forest and k-nearest neighbor classifiers for predicting remaining usage life of roller bearings. Informatica, 48(7):39-52, 2024. https://doi.org/10.31449/inf.v48i7.5726

[8] Qiuyang Zhou, Cai Yi, Lei Yan, Xinwu Song, Du Xu, Chenguang Huang, Lu Zhou, and Jianhui Lin. Multi-objective sparsity maximum mode decomposition: a new method for rotating machine fault diagnosis on high-speed train axle box. IEEE Transactions on Vehicular Technology, 72(10):12744-12756, 2023. https://doi.org/10.1109/TVT.2023.3271588

[9] Delanyo Kwame Bensah Kulevome, Hong Wang, Xuegang Wang. Rolling bearing fault diagnostics based on improved data augmentation and convnet. Journal of Systems Engineering and Electronics, 34(4):1074-1084, 2023. https://doi.org/10.23919/JSEE.2023.000109

[10] Jimin Guo. Balanced generalised tailored approximation point algorithm for solving convex optimisation mathematical problems in bearing vibration signal compressive sensing. Informatica, 49(6):175-190, 2025. https://doi.org/10.31449/inf.v49i6.6979

[11] Ahmed Taibi, Said Touati, Lyes Aomar, and Nabil Ikhlef. Bearing fault diagnosis of induction machines using VMD-DWT and composite multiscale weighted permutation entropy. COMPEL-The international journal for computation and mathematics in electrical and electronic engineering, 43(3):649-668, 2024. https://doi.org/10.1108/COMPEL-11-2023-0580

[12] Cancan Yi, Shuhang Li, Tao Huang, Han Xiao, and Yefeng Jiang. On a prediction method for remaining useful life of rolling bearings via VMD-based dispersion entropy and GAN. IEEE Sensors Journal, 23(22):27744-27756, 2023. https://doi.org/10.1109/JSEN.2023.3323417

[13] Akshay Rajendra Patil, Sandaram Buchaiah, and Piyush Shakya. Combined VMD-morlet wavelet filter-based signal de-noising approach and its applications in bearing fault diagnosis. Journal of Vibration Engineering & Technologies, 12(7):7929-7953, 2024. https://doi.org/10.1007/s42417-024-01338-8

[14] Guangqi Qiu, Yu Nie, Yulong Peng, Peng Huang, Junjie Chen, and Yingkui Gu. A variable-speed-condition fault diagnosis method for crankshaft bearing in the RV reducer with WSO-VMD and ResNet-SWIN. Quality and Reliability Engineering International, 40(5):2321-2347, 2024. https://doi.org/10.1002/qre.3538

[15] Zhilei Jiang, Yang Li, Jinke Gao, and Chengpu Wu. A fault detection of aero-engine rolling bearings based on CNN-BiLSTM network integrated cross-attention. Measurement Science and Technology, 35(12):126116-126128, 2024. https://doi.org/10.1088/1361-6501/ad7622

[16] Jurgen Van Den Hoogen, Dan Hudson, Stefan Bloemheuvel, and Martin Atzmueller. Hyperparameter analysis of wide-kernel cnn architectures in industrial fault detection: an exploratory study. International Journal of Data Science and Analytics, 18(4):423-444, 2024. https://doi.org/10.1109/DSAA60987.2023.10302625

[17] Zhilin Dong, Dezun Zhao, and Lingli Cui. An intelligent bearing fault diagnosis framework: one-dimensional improved self-attention-enhanced CNN and empirical wavelet transform. Nonlinear Dynamics, 112(8):6439-6459, 2024. https://doi.org/10.21203/rs.3.rs-3378300/v1

[18] Bilgin Umut Deveci, Mert Celtikoglu, Ozlem Albayrak, Perin Unal, and Pinar Kirci. Transfer learning enabled bearing fault detection methods based on image representations of single-dimensional signals. Information Systems Frontiers, 26(4): 1345-1397, 2024. https://doi.org/10.1007/s10796-023-10371-z

[19] Chen Yang, Xingwen Wu, Maoru Chi, Wubin Cai, Kaicheng Liu, Shulin Liang, and Wei Wang. A modelling methodology of the axle box bearing-vehicle coupled system dynamics. Vehicle System Dynamics, 62(6):1401-1423, 2024. https://doi.org/10.1080/00423114.2023.2235031

[20] Anju Sharma, Gyanaratha Patra, and Dr. VPS Naidu. Machine learning based bearing fault classification using higher order spectral analysis. Defence Science Journal, 74(4):505-516, 2024. https://doi.org/10.14429/dsj.74.19307

[21] Hairui Fang, Jialin An, Han Liu, Jiawei Xiang, Bo Zhao, and Fir Dunkin. A lightweight transformer with strong robustness application in portable bearing fault diagnosis. IEEE Sensors Journal, 23(9):9649-9657, 2023. https://doi.org/10.1109/JSEN.2023.3260469

[22] Farhan Md. Siraj, Syed Tasnimul Karim Ayon, Md. Abdus Samad, Jia Uddin, and Kwonhue Choi. Few-shot lightweight SqueezeNet architecture for induction motor fault diagnosis using limited thermal image dataset. IEEE Access, 12(2):50986-50997, 2024.
https://doi.org/10.1109/ACCESS.2024.3385430