A rolling bearing fault classification method based on feature optimization and transformer-SVM

Chunxue Wei
Henan Vocational college of Light Industry, Zhengzhou, Henan, 450002, P. R. China
E-mail: *weichx2025@163.com*

Received 17 November 2025; accepted 15 February 2026; published online 27 February 2026
DOI https://doi.org/10.21595/jve.2026.25842

Copyright © 2026 Chunxue Wei. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract. Deep learning-based intelligent fault diagnosis methods have been widely applied in industrial production. However, in practical scenarios, the non-stationary characteristics and strong noise interference of bearing vibration signals significantly constrain the improvement of diagnostic accuracy. To address this issue, this paper proposes an intelligent fault diagnosis framework based on Variational Mode Decomposition (VMD) and Transformer-SVM. This method first employs the Osprey-Cauchy-Sparrow Search Algorithm (OCSSA), with minimum envelope entropy as the optimization objective, to adaptively determine and optimize VMD's mode number K and penalty factor α, thereby obtaining the optimal signal decomposition result. Multi-dimensional indicators are then extracted from the reconstructed signal to construct feature vectors. Subsequently, leveraging the transformer's powerful capability for modeling global dependencies, it mines the deep nonlinear relationships among features. Combined with the Support Vector Machine's strong generalization performance in classification tasks, it achieves accurate classification of bearing faults under complex operating conditions. Comparative experiments on two public datasets show that the proposed method outperforms several existing methods in terms of both classification accuracy and robustness, verifying its effectiveness and advancement.

Keywords: variational mode decomposition, Osprey-Cauchy-Sparrow search algorithm, transformer, support vector machine, fault classification.

1. Introduction

As crucial supporting components [1, 2], rolling bearings are widely used in industrial fields, and their operational status directly affects the reliability and stability of the entire machinery equipment [3]. Moreover, since these key components often operate continuously under extreme working conditions – such as hostile environments, impact loads, and complex dynamic stresses – such demanding circumstances make rolling bearing failures one of the most prevalent types of equipment malfunctions [4-6]. Therefore, achieving efficient fault diagnosis for rolling bearings is of paramount importance for ensuring equipment safety, enhancing utilization efficiency, and preventing significant losses.

At present, the bearing fault diagnosis methods can mainly be divided into three categories: based on signal processing, machine learning and deep learning, respectively. The diagnosis methods based on signal processing mainly use various signal processing techniques such as time-domain statistical parameters (such as kurtosis, skewness, etc.), frequency-domain analysis (envelope spectrum, spectral kurtosis, etc.) and time-frequency transformation (short-time Fourier transform, wavelet analysis, empirical mode decomposition, etc.) to extract fault features. However, this kind of method requires corresponding knowledge reserves and is identified by humans independently.

Machine learning-based diagnostic methods can autonomously learn fault features, reducing the degree of human intervention. Common approaches utilize machine learning algorithms such as Support Vector Machines (SVM) and BP neural networks for fault pattern recognition. However, these methods also exhibit significant limitations. Firstly, the performance of the diagnostic model is highly dependent on the accuracy of manual feature extraction [7]. Secondly, the diagnostic accuracy of the model has certain limitations [8], as its capability to handle nonlinear problems is limited, and machine learning-based methods still require combining signal processing techniques to manually extract and select features [9]. In complex industrial environments, strong background noise and interference signals often submerge key fault characteristics, placing extremely high demands on feature extraction algorithms [10]. Simultaneously, the non-stationary and nonlinear nature of bearing vibration signals can lead to classification difficulties and misjudgment risks [11]. Therefore, researchers often preprocess signals using various signal processing methods, such as Wavelet Packet Transform (WPT) [12, 13], Local Mean Decomposition (LMD) [14], and Empirical Mode Decomposition (EMD) [15, 16]. As one of the most commonly used methods, EMD is suitable for analyzing and processing nonlinear and non-stationary signals but suffers from mode mixing and end effects. Huang proposed the Ensemble Empirical Mode Decomposition (EEMD) method [17] to address the mode mixing issue in EMD, but it introduces noise and still fails to resolve the end effects. Dragomiretskiy [18] proposed the Variational Mode Decomposition (VMD) method, which can adaptively determine relevant frequencies while estimating corresponding modes, successfully handling nonlinear signals. The performance of mode decomposition methods is often significantly affected by parameter settings. To overcome this limitation, various parameter optimization strategies have been proposed in the literature. Li et al. [19] introduced dispersion entropy as an evaluation metric for mode reconstruction and combined it with a dual optimization mechanism to achieve adaptive parameter selection, thereby improving decomposition accuracy and robustness. Zhou et al. employed particle swarm optimization (PSO) to select VMD parameters [20], effectively reducing the influence of manual parameter tuning. Anil Kumar et al. [21] proposed a method based on mutual information (MI) to screen effective IMFs, thereby circumventing the dependency on fixed parameter settings.

With the development of deep learning, deep learning-based fault diagnosis methods have been widely applied in bearing diagnostics. Wang et al. [22] proposed a multimodal diagnostic method that integrates vibration and acoustic signals, utilizing a one-dimensional convolutional neural network (1D-CNN) for feature fusion and fault identification, effectively improving diagnostic accuracy. Yan et al. [23] extracted health information using frequency-domain features and combined it with an enhanced long short-term memory (LSTM) network to further extract features, enabling effective prediction of the remaining useful life of motor bearings. Among the numerous deep learning algorithms, the Transformer, with its self-attention mechanism and multi-head attention structure, can effectively capture global feature dependencies and possesses excellent parallel computing capabilities, significantly enhancing model training and inference efficiency. Dosovitskiy et al. [24] introduced the Vision Transformer (ViT) model, applying Transformer networks in computer vision. Li et al. proposed a novel CNN-Transformer network capable of extracting both local and global discriminative features from vibration signals, demonstrating strong noise resistance [25]. While deep learning-based methods have achieved significant results in the field of fault diagnosis, their classifiers often fall short of expectations when handling nonlinear problems and may reduce the model's generalization ability. In contrast, SVM exhibits strong generalization capability and compared to common classification methods such as decision trees, Softmax, and logistic regression, it is less prone to overfitting and delivers excellent classification performance. Therefore, some scholars have adopted methods that combine deep learning with machine learning to achieve better classification results. Zhang Xunjie et al. [26] proposed an algorithm based on a modified convolutional neural network (CNN) and support vector machine, fully leveraging the powerful feature learning capability of CNN and the superior classification performance of SVM on small samples, thereby improving classification accuracy. Li Zhijun et al. [27] proposed a CNN-SVM hybrid model that effectively enhanced the accuracy of lithology identification.

In summary, to address the issues of manual parameter dependence in modal decomposition effectiveness and the limited performance of neural network classifiers in handling nonlinear problems, this paper proposes an intelligent classification model based on parameter-optimized VMD and Transformer-SVM. The main contributions are as follows:

1) To address the challenge that VMD decomposition performance is highly sensitive to parameter selection, this paper proposes an integrated Osprey-Cauchy-Sparrow Search Algorithm (OCSSA) that selects optimal VMD parameters by minimizing envelope entropy, thereby enhancing fault features.

2) A hybrid diagnostic model integrating Transformer and SVM is proposed, combining Transformer's global dependency feature extraction capability with SVM's excellent classification performance. Comparative experiments demonstrate the significant advantages of this method in classification performance.

The structure of this paper is organized as follows: Section 2 introduces the relevant theories and the framework of the proposed method. Subsequently, Section 3 presents experimental validation of the proposed method across various performance aspects. Finally, Section 4 provides concluding remarks.

2. Theoretical background

2.1. VMD principles

VMD is a non-recursive signal decomposition algorithm that decomposes a signal into several Intrinsic Mode Functions (IMF) based on a preset decomposition level. The resulting IMFs can be expressed as:

$$u_k(t) = A_k(t)\cos\varphi_k(t), \tag{1}$$

where $u_k(t)$ represents the k-th decomposed component, $A_k(t)$ denotes the instantaneous amplitude, and $\varphi_k(t)$ is the signal phase.

The objective of VMD is to ensure that the sum of the decomposed IMF components equals the original signal while minimizing the total bandwidth of all modal components. Consequently, the constrained variational model is formulated as follows:

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

where ω_k is the center frequency of each IMF component, $x(t)$ is the original signal, $\delta(t)$ represents the Dirac delta function, and $*$ denotes the convolution operation.

To solve this constrained variational problem, the Lagrangian multiplier and quadratic penalty factor are introduced, transforming the formula into:

$$\begin{aligned} L(\{u_k\},\{\omega_k\},\lambda) = \alpha\sum_k \left\|\partial_t\left[\left(\delta(t)+\frac{j}{\pi t}\right)*u_k(t)\right]e^{-j\omega_k t}\right\|_2^2 + \left\|x(t)-\sum_k u_k(t)\right\|_2^2 \\ +\langle\lambda(t), x(t)-\sum_k u_k(t)\rangle. \end{aligned} \tag{3}$$

The above formulation is solved through iterative updates of the values of $\{u_k\}$, $\{\omega_k\}$, λ until the optimal solution is obtained. The update expressions are as follows:

$$\hat{u}_k^{n+1}(\omega)=\frac{\hat{x}(\omega)-\sum_{i\neq k}\hat{u}_i^n(\omega)+\dfrac{\hat{\lambda}^n(\omega)}{2}}{1+2\alpha(\omega-\omega_k^n)^2}, \tag{4}$$

$$\omega_k^{n+1}=\frac{\int_0^\infty \omega|\hat{u}_k^n(\omega)|^2 d\omega}{\int_0^\infty |\hat{u}_k^n(\omega)|^2 d\omega}, \tag{5}$$

$$\hat{\lambda}^{n+1}(\omega)=\hat{\lambda}^n(\omega)+\tau\left[\hat{x}(\omega)-\sum_{k=1}^{K}u_k^{n+1}(\omega)\right], \tag{6}$$

where $\hat{x}(\omega)$, $\hat{u}(\omega)$ and $\hat{\lambda}(\omega)$ are the Fourier transform results of $x(t)$, $u(t)$ and $\lambda(t)$ respectively, and n is the iteration count. The iteration terminates when the following condition is met:

$$\frac{\sum_{k=1}^{K}\|\hat{u}_k^{n+1}-\hat{u}_k^n\|_2^2}{\|\hat{u}_k^n\|_2^2}<\varepsilon, \tag{7}$$

where ε is the convergence threshold.

Although VMD possesses excellent adaptive characteristics and can effectively handle nonlinear signals, its decomposition performance is influenced by the number of modal components and the penalty factor. Manual setting of these parameters often fails to achieve the desired objectives, thus necessitating suitable algorithms for optimizing VMD parameters.

2.2. OCSSA

The Sparrow Search Algorithm (SSA) was proposed in 2020 [28], inspired by sparrows' foraging and anti-predation behaviors. This algorithm exhibits strong global exploration and local exploitation capabilities, demonstrating high stability and fast convergence, with a simple structure and few parameters. However, SSA is still prone to falling into local optima when dealing with complex functions and suffers from over-reliance on the previous generation's position update mechanism. Therefore, OCSSA introduces improvements to SSA primarily through three aspects.

First, for the initial population, logistic chaotic mapping is used to enhance the diversity of the initialized population, as shown in the following equation:

$$y_m=\mu y_m(1-y_m), \tag{8}$$

where $y_m\in(0,1]$, $\mu\in(0,4)$ represents the control parameter of the mapping evolution. As the value of μ increases, the mapping range expands and the distribution becomes more uniform.

Furthermore, the Osprey Optimization Algorithm (OOA) [29] is introduced to address SSA's dependency on previous iteration positions. By simulating the movement of ospreys toward prey, the position update mechanism for explorers in SSA is enhanced, thereby improving global exploration capability. The first-phase global search strategy is as follows:

$$X_{i,j}=x_{i,j}+r_{i,j}\left(SF_{i,j}-I_{i,j}x_{i,j}\right), \tag{9}$$

where SF is the fish selected by the osprey, r is a random number between [0,1], and I is either 1 or 2.

Finally, the Cauchy distribution is employed to replace the position update for followers in SSA. The Cauchy distribution generates greater perturbations compared to the Gaussian distribution, which expands the search scope of SSA and helps avoid convergence to local optima by introducing perturbations throughout the update phase. The position update formula is as follows:

$$P_{ij}(t) = P(t)\big(1 + Cauchy(0,1)\big), \tag{10}$$

where $P(t)$ represents the sparrow's position, and $Cauchy(0,1)$ is a random number generated from the standard Cauchy distribution.

The VMD parameters are optimized using the OCSSA algorithm, with the detailed procedure shown in Table 1.

Table 1. Algorithm flow

```
Input: One-dimensional vibration signal $x$, search range of decomposition level $k$, search range of penalty factor $\alpha$, maximum iteration number $T$, population size $N$, explorer proportion $P$, watcher proportion $S$.
Output: Optimal decomposition level $k$, optimal penalty factor $\alpha$, best IMF components
Set the fitness value to minimize envelope entropy, iteratively calculate $k$ and $\alpha$ with the smallest envelope entropy.
   Envelope entropy formula: $\mathrm{E} = -\sum_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{p_i} \log_2(\mathrm{p_i})$
Perform VMD signal decomposition
While $t < T$ do
For 1: $N$
Initialize population using chaotic mapping
End
Select better individuals based on fitness ranking to guide the search
For 1: $P$
Update sparrow positions using the first-phase position formula optimized by OOA
End
For $(P + 1)$: $N$
Update sparrow positions using Cauchy mutation formula
End
For 1: $S$
Update watcher positions
End
If new position is better, update position, $t = t + 1$
end While
Output optimal decomposition level $k$, optimal penalty factor $\alpha$, and best IMF components
```

2.3. Transformer principles

Transformer is a novel deep neural network primarily based on the self-attention mechanism. During data processing, positional information is first embedded into the data. The data with positional encoding is then fed into the Transformer encoder, which leverages multi-head self-attention layers and multi-layer perceptrons to establish mapping relationships between inputs and outputs.

The multi-head attention mechanism enables parallel feature learning by implementing N independent attention heads in parallel. Each attention head first projects the input into query, key, and value vectors. It then calculates attention weights through dot-product operations, ultimately outputting a weighted sum of the values. The computational process is as follows:

$$MultiHead(Q,K,V) = Con(H_1, H_2 \cdots H_n)W^o, \tag{11}$$

where W^o is the weight matrix of the multi-head attention mechanism, H_i can be represented as:

$$H_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) = soft\max\left[\frac{\left(QW_i^Q\right)\left(KW_i^K\right)^T}{\sqrt{d_k}}\right]\left(VW_i^V\right), \tag{12}$$

where Q is the query matrix, K is the key matrix, V is the value matrix, d_k is the dimension of key vectors.

Meanwhile, layer normalization is added after each layer. The input image is transmitted to the MLP after multi-head attention processing, where the fully connected layer completes the mapping, and finally achieves bearing fault classification through Softmax function operation to output category probabilities.

Transformer achieves efficient parallel computation through its self-attention mechanism and possesses powerful long-range dependency modeling capabilities, enabling effective feature extraction from non-stationary signals. However, the Softmax classification function it employs often fails to deliver satisfactory results when handling nonlinear problems and frequently leads to overfitting issues.

2.4. Support vector machine principles

Support Vector Machine is grounded in statistical learning theory and demonstrates high accuracy in handling small-sample classification tasks. By constructing kernel functions to map the input space into a higher-dimensional space, it creates an optimal classification hyperplane for category separation. For given data x and labels y, the classification hyperplane can be defined as $(w \cdot x) + b = 0$ linear function, and the optimal hyperplane can be constructed by solving the following optimization problem:

$$\begin{cases} y_i[(w \cdot x) + b] \geq 1, \\ \min\Phi(w) = \dfrac{1}{2}\|x\|^2 = \dfrac{1}{2}(w' \cdot w). \end{cases} \tag{13}$$

By converting the aforementioned optimization model into a Lagrange function for solution:

$$L(w, b, \alpha) = \frac{1}{2}\|w\| - \sum \alpha\left(y\left((w \cdot x) + b\right) - 1\right), \tag{14}$$

where $\alpha \geq 0$ represents the Lagrange multiplier. Its dual problem forms a typical convex quadratic optimization model, from which the SVM model parameters ω^* and b^* can be derived. In the feature space of SVM, the distance D between the original feature vector and the optimal hyperplane can be expressed as:

$$D = \frac{|\omega^* \cdot \Phi(x) + b^*|}{\|\omega^*\|}. \tag{15}$$

Maximizing this distance can enhance the robustness of the SVM.

2.5. Proposed methodology

This paper proposes a model based on OCSSA-optimized VMD for feature extraction combined with Transformer-SVM, and applies it to faulty bearing diagnosis. To address the dependency of VMD decomposition on the mode number k and penalty factor α, the model first utilizes an optimization algorithm to adaptively select the optimal VMD parameters. The selected optimal IMFs are then used to compute a set of indicators – including mean, variance, peak value, kurtosis, RMS, crest factor, impulse factor, margin factor, and waveform factor – to construct feature vectors as input for the subsequent neural network. The global modeling capability of

Transformer is then leveraged to further extract deep-level features to aid in final fault classification. Finally, the features extracted by Transformer are fed into a SVM, which utilizes the principle of structural risk minimization to achieve accurate fault pattern classification. The specific workflow is shown in Fig. 1.

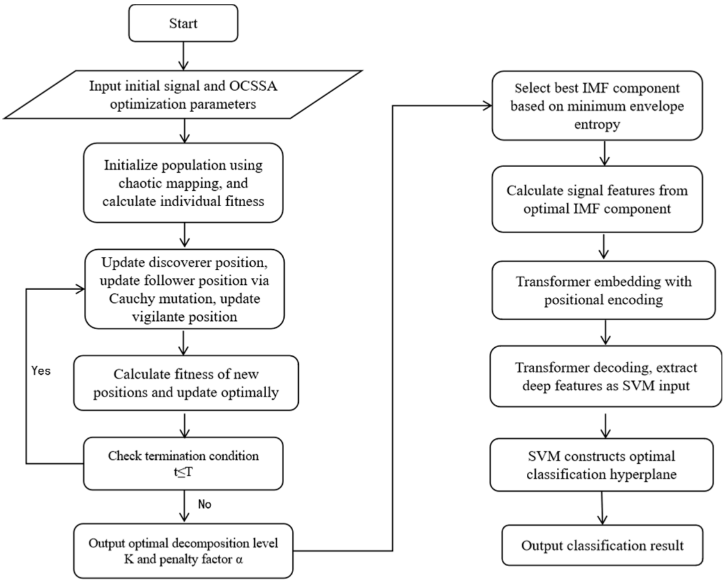

Start
Input initial signal and OCSSA optimization parameters
Initialize population using chaotic mapping, and calculate individual fitness
Update discoverer position, update follower position via Cauchy mutation, update vigilante position
Yes
Calculate fitness of new positions and update optimally
Check termination condition t≤T
No
Output optimal decomposition level K and penalty factor α
Select best IMF component based on minimum envelope entropy
Calculate signal features from optimal IMF component
Transformer embedding with positional encoding
Transformer decoding, extract deep features as SVM input
SVM constructs optimal classification hyperplane
Output classification result

Fig. 1. Overall flowchart

3. Experimental validation

This section validates the performance superiority of the employed optimization algorithm through algorithmic testing, utilizes two public datasets to verify the accuracy of the proposed method, and demonstrates the advantages of the proposed approach by comparing it with different architectures.

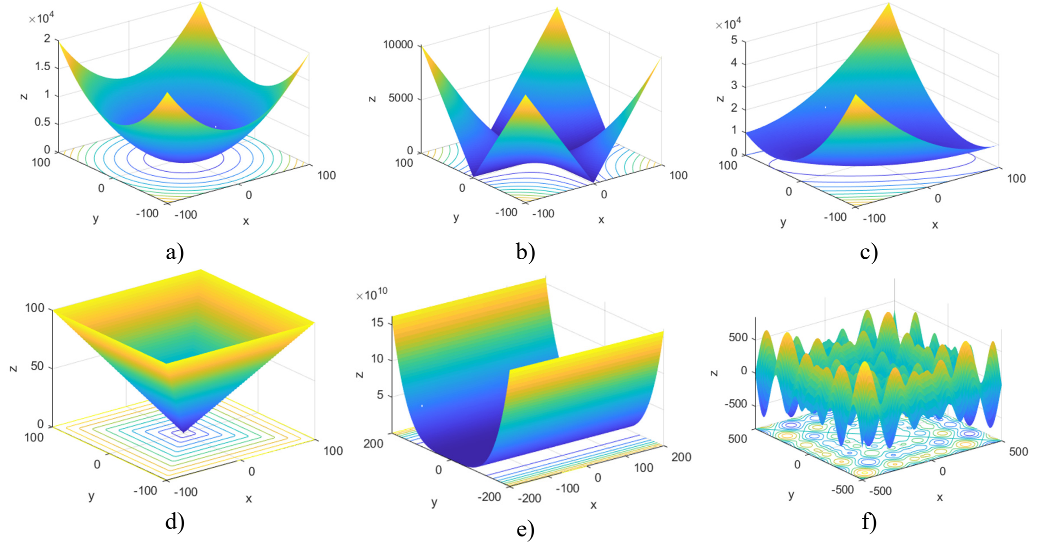
3.1. Algorithm testing

This paper introduces and employs the OCSSA algorithm to optimize VMD parameters. In this subsection, its performance superiority is validated through comparisons with five other common algorithms: Aquila Optimizer (AO) [30], Moth Firefly Optimization (MFO) [31], Honey Badger Algorithm (HBA) [32], Marine Predator Algorithm (MPA), and Sparrow Search Algorithm (SSA) [33].

To comprehensively evaluate the performance of different optimization algorithms, six test functions listed in Table 2 are employed to assess their performance under various optimization scenarios. To ensure a fair comparison, all algorithms are uniformly configured with a population size of 30, a maximum of 1000 iterations, and a problem dimension of 30. Three-dimensional diagrams of the different test functions are shown in Fig. 2.

The solution accuracy and convergence speed of the algorithms are evaluated using the aforementioned test functions. Fig. 3 illustrates the comparison of convergence accuracy and speed among the six algorithms. The letters at the bottom of the figure correspond to the different test functions mentioned above, while the curves in different colors represent the convergence processes of the respective algorithms on each test function. The vertical axis denotes the function value, and the horizontal axis represents the number of iterations.

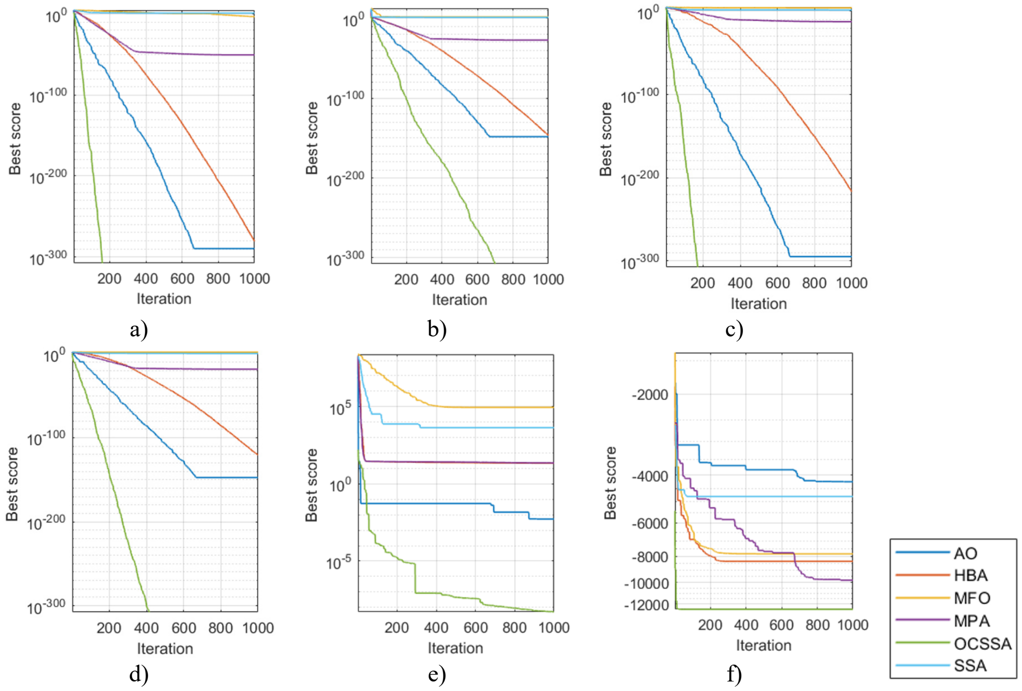
As observed from the figure, OCSSA demonstrates high convergence accuracy and fast convergence speed across all test functions. Its convergence curve descends rapidly and stabilizes within a relatively short time, outperforming comparison algorithms such as SSA, AO, MFO, HBA, and MPA. In test function (e), OCSSA is the first to escape local optima and continues searching toward the global optimum. In test function (f), other algorithms tend to stagnate in local optima with slow subsequent improvement in accuracy, while SSA completely fails to escape local optima. In contrast, OCSSA converges at an extremely fast rate to a significantly higher accuracy level than the other algorithms. Comprehensive analysis confirms that OCSSA achieves the fastest convergence speed and the highest convergence accuracy across various types of nonlinear optimization problems, demonstrating excellent global optimization performance.



a) b) c)

d) e) f)

Fig. 2. 3D diagrams of test functions



a) b) c)

d) e) f)

Fig. 3. Convergence curves of different optimization algorithms

Table 2. Test functions

No.	Test functions	Search range	f_{min}
a	$F_1(x)=\sum_{i=1}^{n} x_i^2$	$[-100,100]^n$	0
b	$F_2(x)=\sum_{i=1}^{n}\lvert x_i\rvert+\prod_{i=1}^{n}\lvert x_i\rvert$	$[-10,10]^n$	0
c	$F_3(x)=\sum_{i=1}^{n}\left(\sum_{j=1}^{i} x_j\right)^2$	$[-100,100]^n$	0
d	$F_4(x)=\max_{i=1}^{n}\lvert x_i\rvert$	$[-100,100]^n$	0
e	$F_5(x)=\sum_{i=1}^{n-1}\left[100\left(x_{i+1}-x_i^2\right)^2+(x_i-1)^2\right]$	$[-30,30]^n$	0
f	$F_8(x)=\sum_{i=1}^{n}-x_i\sin\left(\sqrt{\lvert x_i\rvert}\right)$	$[-500,500]^n$	–125769.5

3.2. Huazhong University of Science and Technology data validation

The experiment utilizes the bearing fault dataset published by Huazhong University of Science and Technology [34]. This dataset was collected using a Spectra-Quest mechanical fault simulator and includes nine condition categories comprising normal, minor faults, and severe faults under different rotational speeds. The signal acquisition setup incorporates triaxial accelerometers and a tachometer. The experimental apparatus is shown in Fig. 4.

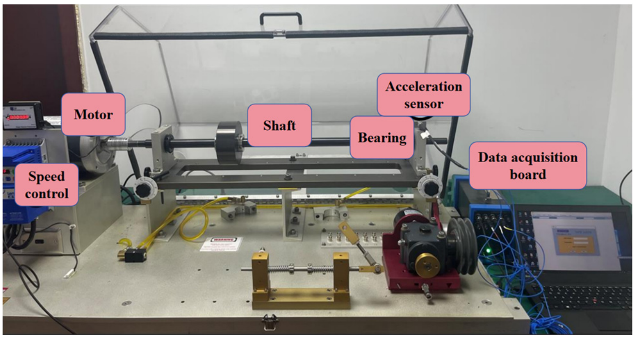


Fig. 4. Bearing test rig at Huazhong University of Science and Technology

The bearing model is ER-16K with a sampling frequency of 25.6 kHz and a single sampling duration of 10.2 seconds. The vibration signal is segmented into data slices every 0.2 seconds with an overlap of 0.05 seconds. Each condition generates 190 samples, with each sample containing 5120 data points. The bearing dataset covers 9 distinct conditions. The complete dataset comprises 1710 samples, of which 1440 are used as the training set and 270 as the test set, as detailed in Table 3.

For the partitioned dataset, following the procedure in Algorithm 1, OCSSA is applied with the objective of minimizing envelope entropy to evaluate the concentration of impact features across IMF components. A lower envelope entropy value indicates that signal energy is more concentrated within fewer impact moments, reflecting more distinct impact characteristics. Based on this, the optimal VMD parameters are obtained and used to decompose the data, from which the optimal IMF component is identified. Nine features are then extracted from this component to form the fault feature vector. The signal resulting from the optimized VMD decomposition is illustrated in Fig. 5.

Table 3. Dataset partition

Fault category	Moderate ball fault	Moderate compound fault	Moderate inner race fault	Moderate outer race fault	Severe ball fault
Label	1	2	3	4	5
Samples	190	190	190	190	190
	Severe compound fault	Normal	Severe inner race fault	Severe outer race fault	
Label	6	7	8	9	
Samples	190	190	190	190	

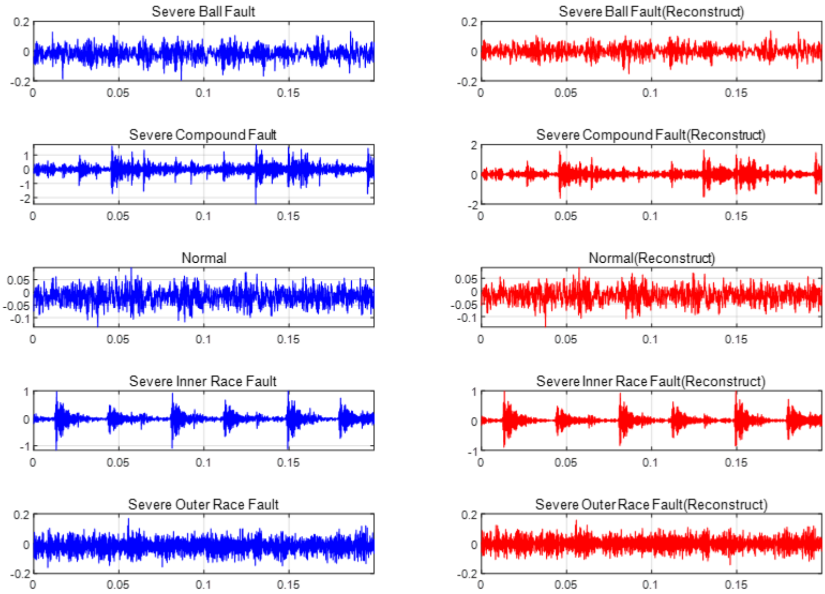
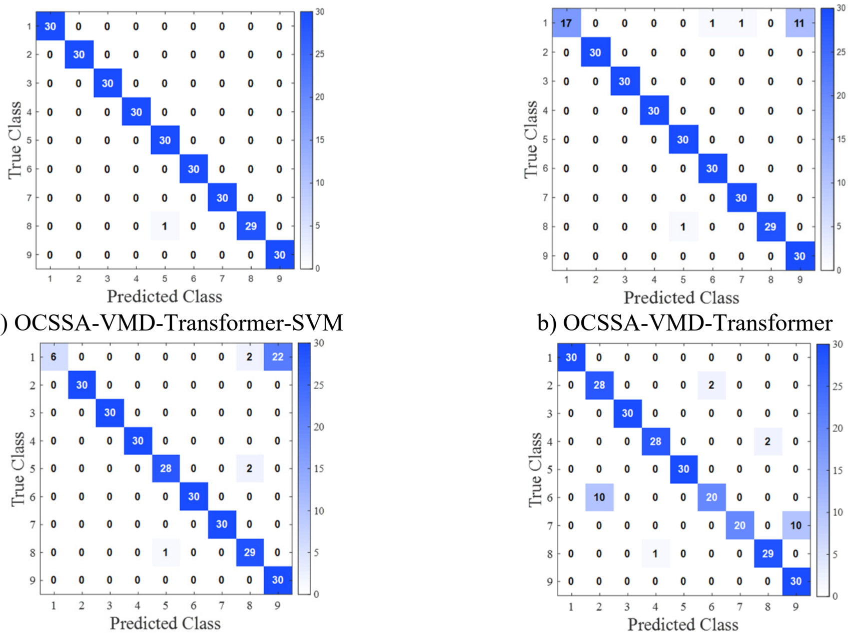


Fig. 5. Comparison of original and reconstructed signals

The figure displays time-domain comparisons of five signal types: normal bearing condition, severe inner race fault, severe outer race fault, severe ball fault, and severe compound fault, showing original signals (left) and reconstructed signals (right). The figure demonstrates that the optimized VMD effectively reduces noise interference. Therefore, selecting the optimal IMF components as computational signals can effectively enhance fault characteristics.

The signal features constructed after OCSSA-VMD processing are input into the established Transformer-SVM network model, with the number of attention heads set to 8, initial learning rate to 0.001, learning rate decay factor to 0.1, and Adam optimizer. The resulting fault classification results are shown in Fig. 6. Since this study primarily focuses on investigating the impact of different features and network models on classification accuracy, all model parameters are assigned fixed values without further parameter optimization.

Fig. 6 displays the classification results of nine bearing fault types under different diagnostic methods. The proposed method achieves the highest accuracy of 99.63 %, significantly outperforming other approaches and demonstrating exceptional capability in identifying subtle fault characteristics. In comparison, the methods in Fig. 6(b) and 6(c) achieve accuracies of only 94.81 % and 90 %, respectively. Both misclassify moderate ball faults as severe inner race faults, primarily due to the oversimplified linear classification layer in the Transformer architecture and the insufficient feature mining by SVM. The method in Fig. 6(d) achieves an accuracy of 90.74 % and shows errors in classifying severe compound faults and normal signals, indicating that standalone VMD decomposition has limitations in extracting critical fault components, thereby affecting subsequent feature learning and classification performance. Comprehensive comparison confirms that the proposed method exhibits significant advantages in both feature extraction and classification decision-making, effectively improving the accuracy of bearing fault diagnosis.



	1	2	3	4	5	6	7	8	9
1	30	0	0	0	0	0	0	0	0
2	0	30	0	0	0	0	0	0	0
3	0	0	30	0	0	0	0	0	0
4	0	0	0	30	0	0	0	0	0
5	0	0	0	0	30	0	0	0	0
6	0	0	0	0	0	30	0	0	0
7	0	0	0	0	0	0	30	0	0
8	0	0	0	0	1	0	0	29	0
9	0	0	0	0	0	0	0	0	30

a) OCSSA-VMD-Transformer-SVM

	1	2	3	4	5	6	7	8	9
1	17	0	0	0	0	1	1	0	11
2	0	30	0	0	0	0	0	0	0
3	0	0	30	0	0	0	0	0	0
4	0	0	0	30	0	0	0	0	0
5	0	0	0	0	30	0	0	0	0
6	0	0	0	0	0	30	0	0	0
7	0	0	0	0	0	0	30	0	0
8	0	0	0	0	1	0	0	29	0
9	0	0	0	0	0	0	0	0	30

b) OCSSA-VMD-Transformer

	1	2	3	4	5	6	7	8	9
1	6	0	0	0	0	0	0	2	22
2	0	30	0	0	0	0	0	0	0
3	0	0	30	0	0	0	0	0	0
4	0	0	0	30	0	0	0	0	0
5	0	0	0	0	28	0	0	2	0
6	0	0	0	0	0	30	0	0	0
7	0	0	0	0	0	0	30	0	0
8	0	0	0	0	1	0	0	29	0
9	0	0	0	0	0	0	0	0	30

c) OCSSA-VMD-SVM

	1	2	3	4	5	6	7	8	9
1	30	0	0	0	0	0	0	0	0
2	0	28	0	0	0	2	0	0	0
3	0	0	30	0	0	0	0	0	0
4	0	0	0	28	0	0	0	2	0
5	0	0	0	0	30	0	0	0	0
6	0	10	0	0	0	20	0	0	0
7	0	0	0	0	0	0	20	0	10
8	0	0	0	1	0	0	0	29	0
9	0	0	0	0	0	0	0	0	30

d) VMD-Transformer-SVM

Fig. 6. Confusion matrices of different algorithms

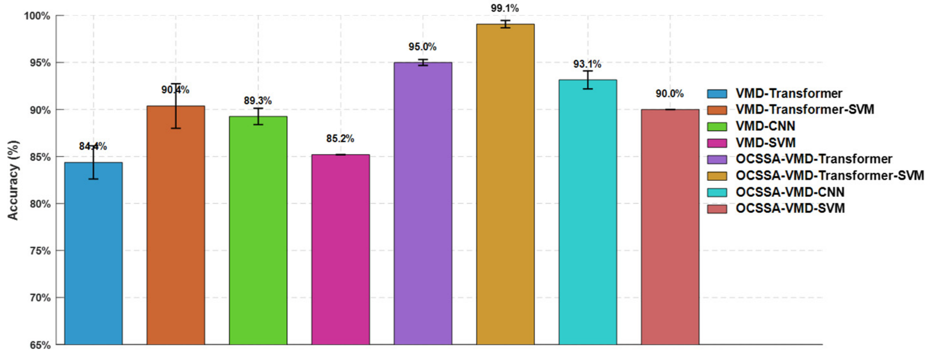


Fig. 7. Classification accuracy of different algorithms

Fig. 7 compares the classification performance of eight diagnostic methods, with all results presented as means and error bars based on five repeated experiments. Compared with conventional CNN and Transformer models, the proposed OCSSA-VMD-Transformer-SVM algorithm achieves the highest classification accuracy of approximately 99.1 %. From the perspective of input features, the performance of SVM, CNN, and Transformer models is significantly dependent on feature quality. When features are extracted using the VMD method optimized by OCSSA, the diagnostic accuracy is markedly higher than that achieved by the non-optimized VMD method, along with a notably narrower error range. This demonstrates that combining deep learning methods with optimized feature extraction strategies can effectively improve recognition accuracy and enhance model stability. From the perspective of network architectures, the Transformer-SVM structure consistently delivers the best performance across different feature inputs. The proposed architecture integrates the strength of Transformer in deep feature extraction with the generalization capability of SVM, thereby significantly improving the model's ability to discriminate multiple fault types under complex working conditions.

Table 4. Time cost of different algorithms

Algorithms	Time cost (S)
VMD-CNN	30.05
VMD-Transform	33.56
VMD-Transform-SVM	33.58
OCSSA- VMD-CNN	30.36
OCSSA- VMD-Transform	33.72
OCSSA- VMD-Transform-SVM	33.73

The time costs of various algorithms are summarized in Table 4, with results averaged over five repeated experiments. While SVM exhibits the lowest time cost, its accuracy is also the lowest; hence, only the time costs of different deep learning algorithms are presented here. Compared to other methods, the proposed OCSSA-VMD-Transformer-SVM model not only achieves the highest classification accuracy but also maintains a time cost comparable to that of the baseline CNN network. In summary, the proposed method optimizes the model from both the feature and architecture levels by integrating improved network architecture and optimized feature extraction, achieving higher fault classification accuracy under comparable time costs.

3.3. Case Western Reserve University data validation

The experiment utilizes the bearing fault dataset released by Case Western Reserve University [35]. The test rig consists of a motor, torque transducer, dynamometer, and control system. Single-point faults were introduced using electric discharge machining with fault diameters of 0.007 inches, 0.014 inches, and 0.021 inches. The fault types include inner race faults, outer race faults, and ball faults, totaling 9 fault conditions. Combined with the normal bearing condition, this forms a 10-class bearing fault dataset. The experimental setup is shown in Fig. 8.

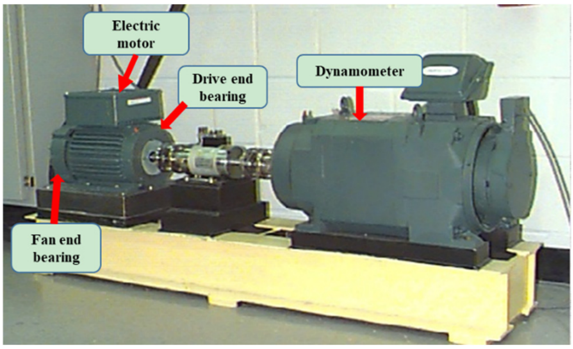


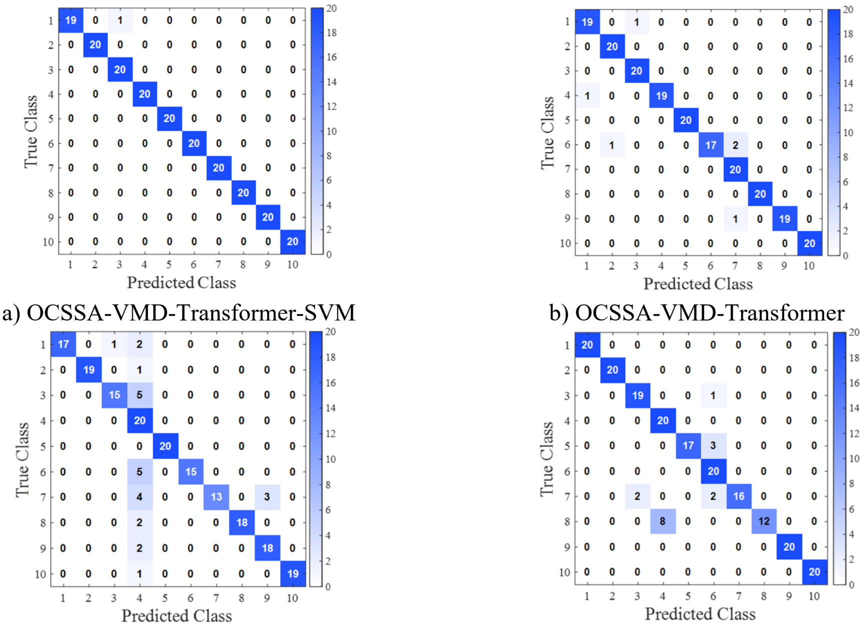
Fig. 8. Case Western Reserve University bearing test rig

The bearing used is SKF6205 with a sampling frequency of 12,000 Hz. The individual data sampling duration varies between 10-20 seconds. Therefore, 2024 data points are uniformly extracted to form one data sample using a sliding window of 1000 data points. Each fault category yields 120 data samples, with 100 allocated for training and 20 for testing. The dataset comprises 1200 samples in total, with 1000 designated as the training set and 200 as the test set, as detailed in Table 5.

Similarly, the classified data is processed using OCSSA-VMD, and the corresponding signal features are extracted as input vectors. The processing methods and parameters remain consistent with those described in Section 3.2. The classification results of four comparative methods are shown in Fig. 9.

Table 5. Dataset partition

Fault category	Normal	0.007 Inner race fault	0.007 Ball fault	0.007 Outer race fault	0.014 Inner race fault
Label	1	2	3	4	5
Samples	120	120	120	120	120
	0.014 Ball fault	0.014 Outer race fault	0.021 Inner race fault	0.021 Ball fault	0.021 Outer race fault
Label	6	7	8	9	10
Samples	120	120	120	120	120



a) OCSSA-VMD-Transformer-SVM

b) OCSSA-VMD-Transformer

c) OCSSA-VMD-SVM

d) VMD-Transformer-SVM

Fig. 9. Confusion matrices of different algorithms

Fig. 9 displays the classification results of the ten-class bearing fault data using different methods. The proposed method achieves an accuracy of 99.5 %, with only one normal sample misclassified as a minor ball fault. In comparison, the methods in Fig. 9(b) and 9(c) achieve accuracies of only 97 % and 87 % respectively, validating the effectiveness of the Transformer-SVM fusion strategy in enhancing classification performance. The method in Fig. 9(d) achieves an accuracy of 92 %, indicating that the unoptimized VMD decomposition underperforms in fault feature extraction, thereby limiting subsequent classification performance. A comprehensive comparison of the four methods demonstrates the significant advantage of the proposed approach in classification accuracy.

Fig. 10 compares the classification performance of eight diagnostic methods on the Case Western Reserve dataset, with all results presented as the average and error bars from five repeated experiments. Compared to CNN and Transformer models under different feature inputs, the proposed OCSSA-VMD-Transformer-SVM model achieved the highest classification accuracy of approximately 99.2 %. From the perspective of feature extraction methods, when features were extracted using the VMD method optimized by OCSSA, diagnostic accuracy was significantly higher than that of the non-optimized VMD method, with a noticeably narrower error range. This indicates that optimizing the input features can more effectively highlight fault characteristics, thereby improving the diagnostic performance of the network. From the dimension of network architecture, the Transformer-SVM structure performed optimally across different feature input conditions; the diagnostic accuracy of the standalone Transformer model and CNN decreased relatively; using SVM alone for classification yielded the poorest results.

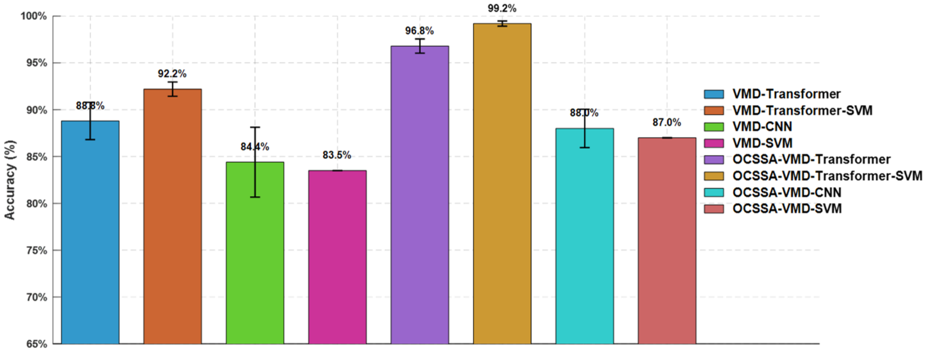


Fig. 10. Classification accuracy of different algorithms

Table 6. Time cost of different algorithms

Algorithms	Time cost (S)
VMD-CNN	20.43
VMD-Transform	17.79
VMD-Transform-SVM	17.81
OCSSA- VMD-CNN	20.84
OCSSA- VMD-Transform	18.01
OCSSA- VMD-Transform-SVM	18.03

The time costs of different algorithms are presented in Table 6. Compared to other methods, the model proposed in this paper exhibits a time cost that is comparable to other approaches, while achieving the highest classification accuracy. In summary, the proposed method can achieve higher fault classification accuracy and stability than other methods under comparable time costs.

4. Conclusions

To address the challenges of non-stationary signal feature extraction and strong noise interference in bearing fault classification, an intelligent classification model based on OCSSA-VMD and Transformer-SVM is proposed. This algorithm employs the improved OCSSA with minimum envelope entropy as the criterion to adaptively optimize the mode number K and penalty factor α of VMD, achieving high-quality signal decomposition while constructing multi-dimensional feature vectors. The global dependency modeling capability of Transformer is leveraged to mine deep feature correlations, combined with the excellent generalization performance of SVM for accurate classification. Experimental analysis compares the convergence curves of OCSSA and five other optimization algorithms on benchmark functions, demonstrating that OCSSA exhibits superior performance over the compared optimization algorithms. Experiments conducted on bearing datasets from Huazhong University of Science and Technology and Case Western Reserve University, in comparison with seven other methods, show that the proposed algorithm achieves classification accuracies of 99.63 % and 99.5 %, respectively. Furthermore, a comparison of time costs among different methods demonstrates that the proposed algorithm achieves significantly higher classification accuracy than the compared methods under comparable time costs, validating its effectiveness and practicality in complex industrial scenarios.

Although the proposed method demonstrates favorable performance in fault classification, it still has certain limitations. This study primarily focuses on improvements in signal processing and network architecture integration, while considerations regarding network parameter optimization and adaptability to variable-speed operating conditions in real industrial scenarios remain insufficient. Therefore, future research will concentrate on variable-speed signal processing and network parameter optimization to provide more comprehensive and reliable diagnostic solutions for industrial applications.

Acknowledgements

The authors have not disclosed any funding.

Data availability

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

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] Y. Jin, C. Qin, Y. Huang, and C. Liu, “Actual bearing compound fault diagnosis based on active learning and decoupling attentional residual network,” *Measurement*, Vol. 173, p. 108500, Mar. 2021, https://doi.org/10.1016/j.measurement.2020.108500

[2] H. Zhou et al., “Hybrid system response model for condition monitoring of bearings under time-varying operating conditions,” *Reliability Engineering and System Safety*, Vol. 239, p. 109528, Nov. 2023, https://doi.org/10.1016/j.ress.2023.109528

[3] I. El-Thalji and E. Jantunen, “A summary of fault modelling and predictive health monitoring of rolling element bearings,” *Mechanical Systems and Signal Processing*, Vol. 60-61, pp. 252–272, Aug. 2015, https://doi.org/10.1016/j.ymssp.2015.02.008

[4] R. Liu, B. Yang, E. Zio, and X. Chen, “Artificial intelligence for fault diagnosis of rotating machinery: A review,” *Mechanical Systems and Signal Processing*, Vol. 108, pp. 33–47, Aug. 2018, https://doi.org/10.1016/j.ymssp.2018.02.016

[5] Y. Zhong, G. Xie, H. Geng, J. Hou, D. Zhao, and W. He, “Thermal analysis for plate structures using a transformation BEM based on complex poles,” *Computers and Mathematics with Applications*, Vol. 161, pp. 32–42, May 2024, https://doi.org/10.1016/j.camwa.2024.02.034

[6] W. Zhao et al., “Multiscale inverted residual convolutional neural network for intelligent diagnosis of bearings under variable load condition,” *Measurement*, Vol. 188, p. 110511, Jan. 2022, https://doi.org/10.1016/j.measurement.2021.110511

[7] M. Cui, Y. Wang, X. Lin, and M. Zhong, “Fault diagnosis of rolling bearings based on an improved stack autoencoder and support vector machine,” *IEEE Sensors Journal*, Vol. 21, No. 4, pp. 4927–4937, Feb. 2021, https://doi.org/10.1109/jsen.2020.3030910

[8] W. Zhang, C. Li, G. Peng, Y. Chen, and Z. Zhang, “A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load,” *Mechanical Systems and Signal Processing*, Vol. 100, pp. 439–453, Feb. 2018, https://doi.org/10.1016/j.ymssp.2017.06.022

[9] F. Kibrete, D. Engida Woldemichael, and H. Shimels Gebremedhen, “Multi-Sensor data fusion in intelligent fault diagnosis of rotating machines: A comprehensive review,” *Measurement*, Vol. 232, p. 114658, Jun. 2024, https://doi.org/10.1016/j.measurement.2024.114658

[10] Y. Xu, X. Yan, B. Sun, and Z. Liu, “Dually attentive multiscale networks for health state recognition of rotating machinery,” *Reliability Engineering and System Safety*, Vol. 225, p. 108626, Sep. 2022, https://doi.org/10.1016/j.ress.2022.108626

[11] Z. Feng, S. Wang, and M. Yu, “A fault diagnosis for rolling bearing based on multilevel denoising method and improved deep residual network,” *Digital Signal Processing*, Vol. 140, p. 104106, Aug. 2023, https://doi.org/10.1016/j.dsp.2023.104106

[12] X. Li, Z. Ma, Kang, and X. Li, "Fault diagnosis for rolling bearing based on VMD-FRFT," *Measurement*, Vol. 155, p. 107554, Apr. 2020, https://doi.org/10.1016/j.measurement.2020.107554

[13] J. S. Rapur and R. Tiwari, "Experimental fault diagnosis for known and unseen operating conditions of centrifugal pumps using MSVM and WPT based analyses," *Measurement*, Vol. 147, p. 106809, Dec. 2019, https://doi.org/10.1016/j.measurement.2019.07.037

[14] D. Yu, J. Cheng, and Y. Yang, "Application of EMD method and Hilbert spectrum to the fault diagnosis of roller bearings," *Mechanical Systems and Signal Processing*, Vol. 19, No. 2, pp. 259–270, Mar. 2005, https://doi.org/10.1016/s0888-3270(03)00099-2

[15] L. Xu, S. Chatterton, and P. Pennacchi, "Rolling element bearing diagnosis based on singular value decomposition and composite squared envelope spectrum," *Mechanical Systems and Signal Processing*, Vol. 148, p. 107174, Feb. 2021, https://doi.org/10.1016/j.ymssp.2020.107174

[16] N. E. Huang et al., "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, Vol. 454, No. 1971, pp. 903–995, Mar. 1998, https://doi.org/10.1098/rspa.1998.0193

[17] Z. Wu and N. E. Huang, "Ensemble empirical mode decomposition: a noise-assisted data analysis method," *Advances in Adaptive Data Analysis*, Vol. 1, No. 1, pp. 1–41, Nov. 2011, https://doi.org/10.1142/s1793536909000047

[18] K. Dragomiretskiy and D. Zosso, "Variational mode decomposition," *IEEE Transactions on Signal Processing*, Vol. 62, No. 3, pp. 531–544, Feb. 2014, https://doi.org/10.1109/tsp.2013.2288675

[19] Y. Li and X. Cheng, "Double-optimized symmetric geometric mode decomposition with dispersion entropy and its application in feature extraction," *Signal Processing*, Vol. 235, p. 110046, Oct. 2025, https://doi.org/10.1016/j.sigpro.2025.110046

[20] F. Zhou, X. Yang, J. Shen, and W. Liu, "Fault diagnosis of hydraulic pumps using PSO-VMD and refined composite multiscale fluctuation dispersion entropy," *Shock and Vibration*, Vol. 2020, pp. 1–13, Aug. 2020, https://doi.org/10.1155/2020/8840676

[21] A. Kumar et al., "Non-parametric ensemble empirical mode decomposition for extracting weak features to identify bearing defects," *Measurement*, Vol. 211, p. 112615, Apr. 2023, https://doi.org/10.1016/j.measurement.2023.112615

[22] X. Wang, D. Mao, and X. Li, "Bearing fault diagnosis based on vibro-acoustic data fusion and 1D-CNN network," *Measurement*, Vol. 173, p. 108518, Mar. 2021, https://doi.org/10.1016/j.measurement.2020.108518

[23] H. Yan, Y. Qin, S. Xiang, Y. Wang, and H. Chen, "Long-term gear life prediction based on ordered neurons LSTM neural networks," *Measurement*, Vol. 165, p. 108205, Dec. 2020, https://doi.org/10.1016/j.measurement.2020.108205

[24] A. Dosovitskiy et al., "An image is worth 16x16 words: transformers for image recognition at scale," *arXiv:2010.11929*, Jan. 2020, https://doi.org/10.48550/arxiv.2010.11929

[25] S. Li et al., "Dconformer: A denoising convolutional transformer with joint learning strategy for intelligent diagnosis of bearing faults," *Mechanical Systems and Signal Processing*, Vol. 210, p. 111142, Mar. 2024, https://doi.org/10.1016/j.ymssp.2024.111142

[26] X. Zhang, M. Zhang, Z. Xiang, and J. Mo, "Research on diagnosis algorithm of mechanical equipment brake friction fault based on MCNN-SVM," *Measurement*, Vol. 186, p. 110065, Dec. 2021, https://doi.org/10.1016/j.measurement.2021.110065

[27] Z. Li, S. Deng, Y. Hong, Z. Wei, and L. Cai, "A novel hybrid CNN-SVM method for lithology identification in shale reservoirs based on logging measurements," *Journal of Applied Geophysics*, Vol. 223, p. 105346, Apr. 2024, https://doi.org/10.1016/j.jappgeo.2024.105346

[28] J. Xue and B. Shen, "A novel swarm intelligence optimization approach: sparrow search algorithm," *Systems Science and Control Engineering*, Vol. 8, No. 1, pp. 22–34, Jan. 2020, https://doi.org/10.1080/21642583.2019.1708830

[29] M. Dehghani and P. Trojovský, "Osprey optimization algorithm: A new bio-inspired metaheuristic algorithm for solving engineering optimization problems," *Frontiers in Mechanical Engineering*, Vol. 8, p. 11264, Jan. 2023, https://doi.org/10.3389/fmech.2022.1126450

[30] A. Faramarzi, M. Heidarinejad, S. Mirjalili, and A. H. Gandomi, "Marine predators algorithm: A nature-inspired metaheuristic," *Expert Systems with Applications*, Vol. 152, p. 113377, Aug. 2020, https://doi.org/10.1016/j.eswa.2020.113377

[31] F. A. Hashim, E. H. Houssein, K. Hussain, M. S. Mabrouk, and W. Al-Atabany, "Honey badger algorithm: new metaheuristic algorithm for solving optimization problems," *Mathematics and*

Computers in Simulation, Vol. 192, pp. 84–110, Feb. 2022, https://doi.org/10.1016/j.matcom.2021.08.013

[32] S. Mirjalili, "Moth-flame optimization algorithm: A novel nature-inspired heuristic paradigm," *Knowledge-Based Systems*, Vol. 89, pp. 228–249, Nov. 2015, https://doi.org/10.1016/j.knosys.2015.07.006

[33] S. Mirjalili, A. H. Gandomi, S. Z. Mirjalili, S. Saremi, H. Faris, and S. M. Mirjalili, "Salp swarm algorithm: a bio-inspired optimizer for engineering design problems," *Advances in Engineering Software*, Vol. 114, pp. 163–191, Dec. 2017, https://doi.org/10.1016/j.advengsoft.2017.07.002

[34] C. Zhao, E. Zio, and W. Shen, "Domain generalization for cross-domain fault diagnosis: An application-oriented perspective and a benchmark study," *Reliability Engineering and System Safety*, Vol. 245, p. 109964, May 2024, https://doi.org/10.1016/j.ress.2024.109964

[35] W. A. Smith and R. B. Randall, "Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study," *Mechanical Systems and Signal Processing*, Vol. 64-65, pp. 100–131, Dec. 2015, https://doi.org/10.1016/j.ymssp.2015.04.021

Chunxue Wei received M.A. degree in School of Mechanical and Electrical Engineering from Lanzhou University of Technology, Lanzhou, China, in 2006. Now she works at Henan Vocational college of Light industry. Her current research interests include control, dynamics and fault diagnosis.