Intelligent Fault Diagnosis Method for Rolling Bearings Based on SSA-CNN-Transformer

Jinyuan Hu *

School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, China, 430065

* Corresponding Author Email: jyhu825@gmail.com

Abstract. Deep learning has become a key solution in intelligent fault diagnosis, as its ability to learn features directly from raw data addresses the challenges of modeling complex signals in rolling bearings. Traditional Convolutional Neural Networks (CNNs) are constrained by fixed receptive fields and static kernels, which limit their adaptability to dynamic, multi-scale features in vibration signals. Moreover, existing models often lack an adaptive mechanism for evaluating feature importance, which reduces diagnostic robustness in non-stationary and variable operating conditions. This paper introduces the SSA-CNN-Transformer model, which integrates the Sparrow Search Algorithm (SSA) with a self-attention mechanism to address these challenges in intelligent bearing fault diagnosis. The SSA globally optimizes key hyperparameters, improving the efficiency and performance of the model architecture. The CNN module extracts local time-frequency features from vibration signals and performs multi-scale fusion, while the Transformer module captures long-range dependencies, leading to a more accurate and comprehensive representation of fault patterns for precise classification. Empirical evaluations on three publicly available datasets—CWRU, XJTU, and DIRG—demonstrate that the proposed model outperforms current state-of-the-art methods in multiple performance metrics, exhibiting superior diagnostic accuracy and generalization. This work offers valuable insights and a solid foundation for developing intelligent health monitoring systems for real-world industrial applications.

Keywords: Rolling Bearings, Sparrow Search Algorithm, Convolutional Neural Network, Attention Mechanism, Intelligent Diagnosis.

1. Introduction

Rotating machinery, a core component of modern industrial systems, plays a critical role in ensuring the safety, stability, and economic performance of the entire production process. Rolling bearings, essential for power transmission and support, are widely used in high-precision and high-load equipment, where their performance directly impacts the operational reliability of the entire system. It is estimated that approximately 30% of rotating machinery failures originate from rolling bearing damage [1]. Thus, developing a high-precision, robust intelligent health monitoring and fault diagnosis system is crucial for enabling intelligent operation and preventive maintenance of industrial equipment.

In industrial environments, rolling bearings are often exposed to extreme conditions—such as variable loads, high temperatures, humidity, and particulate contamination—which can induce various failure modes, including fatigue spalling and microcrack propagation. The vibration signals of bearings exhibit significant nonlinear and non-stationary characteristics due to factors such as varying rotational speeds, fluctuating loads, and environmental disturbances. These factors significantly challenge the adaptability and discriminative power of traditional frequency-domain and time-frequency-domain signal processing methods. For example, Fast Fourier Transform (FFT) struggles with transient shocks and modulation frequencies due to its fixed window characteristics, making it difficult to capture both local and global information—especially when rotational speeds fluctuate, leading to spectral analysis failure [2]. Furthermore, while Support Vector Machines (SVMs) exhibit good generalization under small sample conditions, their reliance on fixed kernel functions for constructing a single-scale feature space limits accurate modeling of multi-scale fault patterns [3], creating performance bottlenecks in large-scale industrial applications.

The rapid development of artificial intelligence has led to deep learning techniques offering unprecedented advantages in intelligent fault diagnosis, thanks to their superior ability to model nonlinear features and end-to-end self-learning capabilities. Convolutional Neural Networks (CNN) [4], with their multi-layer architecture, extract spatial distribution features from signals and are effective in modeling local fault patterns. Long Short-Term Memory (LSTM) [5] networks effectively retain historical information, making them suitable for time-series pattern recognition. The Transformer structure [6], utilizing the self-attention mechanism, excels at modeling long-range dependencies, offering a novel approach to feature extraction from long sequences. Recently, researchers have extensively explored these models. For instance, the WDD-CNN model was proposed by Eyup et al. [7], achieving a diagnostic accuracy of 96.45% under multi-condition and high-noise environments. The feature extraction capability of CNN was enhanced by Hu et al. [8] using multi-scale convolution kernels, which significantly improved classification performance.; Tang et al. [9] used a Bi-LSTM structure to accurately identify early bearing ball faults, showing excellent predictive capabilities. To overcome the trade-off between computational resources and accuracy, Kangjie et al. [10] designed a lightweight LSTM-Transformer hybrid framework that balances model precision and computational efficiency. Additionally, the Transformer Transfer Learning Network (TTLN) was constructed [11], demonstrating robust cross-condition recognition with minimal target domain samples, thus further extending the application of Transformer models in small-sample fault diagnosis.

However, existing methods still face three major challenges: First, traditional CNNs are constrained by fixed convolution kernel sizes, making it difficult to capture long-range dependencies in long sequences; second, while Transformers excel in modeling global temporal dependencies, they suffer from information loss when handling local fault details; third, the performance of hybrid model architectures heavily depends on manually set hyperparameters, leading to instability and reduced generalization capabilities [12]. Thus, how to integrate local sensitivity and global modeling mechanisms, while improving the model's adaptability through intelligent optimization techniques, has become a frontier challenge in intelligent fault diagnosis research.

To address the difficulty of simultaneously modeling multi-scale fault features and long-range dependencies in rolling bearing vibration signals under complex operating conditions, this paper proposes an intelligent diagnosis model based on a CNN-Transformer hybrid structure optimized by the Sparrow Search Algorithm (SSA). This method integrates the advantages of CNNs and Transformers, and enhances model adaptability and generalization performance through an intelligent hyperparameter tuning mechanism, improving diagnostic robustness and accuracy in non-stationary, multi-condition environments. To overcome the limitations of traditional hybrid models, which rely on manually set hyperparameters, this study introduces the SSA algorithm, which dynamically evolves to globally optimize key hyperparameters (such as convolution kernel size, number of attention heads, etc.), accelerating model convergence while improving its stability and diagnostic performance under complex conditions. The main contributions of this study can be summarized as follows:

(1) A SSA-driven hyperparameter optimization framework is proposed to achieve self-adaptive adjustment of the CNN-Transformer hybrid structure's parameters, replacing traditional manual tuning and significantly improving model training efficiency and performance.

(2) A multi-branch CNN feature extraction module is constructed to efficiently extract multi-scale local features from raw vibration signals, enabling deep fusion and representation of spatial information.

(3) A local-global collaborative modeling mechanism is designed, where CNN-extracted local features serve as input to the Transformer, enhancing the sequence modeling ability through self-attention mechanisms and improving the accuracy and robustness of long-range dependency modeling.

The paper is organized as follows: Section 2 elaborates the overall structure and key module principles of the proposed model; Section 3 presents specific experimental setups and performance

evaluation results, comparing them with existing methods; Section 4 concludes the paper, discussing the engineering significance and future directions of the research.

2. Methods and Models

2.1. Sparrow Search Algorithm Principle

The Sparrow Search Algorithm (SSA) [13] is an innovative optimization method based on swarm intelligence, inspired by the foraging and predation behaviors of sparrow populations. The algorithm simulates the cooperative mechanism between discoverers and followers in the sparrow group, incorporating an adaptive weight adjustment strategy to balance global exploration with local exploitation. This approach demonstrates faster convergence rates and improved robustness compared to traditional optimization algorithms, such as Particle Swarm Optimization (PSO) [14] and Genetic Algorithm (GA) [15].

SSA divides the sparrow population into two distinct roles: discoverers and followers. Discoverers are responsible for global exploration in search of food sources, while followers rely on the information provided by discoverers to conduct more refined local searches, gradually approaching the food source. Discoverers typically exhibit the highest movement speed, enabling them to escape local optima. The position update rule for discoverers is given by the following equation:

$$L_{i,j}^{t+1} = \begin{cases} L_{i,j}^{t} \cdot \exp\left(-\dfrac{i}{\alpha \cdot iter_{max}}\right), & R_2 < S_T \\ L_{i,j}^{t} + Q \cdot \boldsymbol{A}, & R_2 \geq S_T \end{cases} \tag{1}$$

In this equation, $L_{i,j}^{t}$ represents the position of the $i-$th discoverer at the $j-$th iteration; $iter_{\max}$ is the maximum number of iterations; α is a random number between [0,1]; Q is a random value drawn from a uniform distribution within the search space; $\boldsymbol{A}$ is a $1\times d$ unit matrix; and R_2 and S_T are the warning threshold and safety threshold, respectively.

Followers update their positions by moving toward high-quality solutions, as defined by the following equation:

$$L_{i,j}^{t+1} = \begin{cases} Q \cdot \exp\left(\dfrac{L_w^t - L_{i,j}^t}{i^2}\right), & i > n/2 \\ L_P^{t+1} + |L_{i,j}^t - L_P^{t+1}| \cdot \boldsymbol{M}^+ \cdot \boldsymbol{A}, & i \leq n/2 \end{cases} \tag{2}$$

In this equation, L_p^{t+1} represents the current global best position, L_w^t denotes the global worst position, and $\boldsymbol{M}$ is a random matrix of size $1\times d$, with each element randomly set to 1 or -1. The matrx $\boldsymbol{M}^+$ is computed as $\boldsymbol{M}^{\mathrm{T}}(\boldsymbol{M}\boldsymbol{M}^{\mathrm{T}})^{-1}$, and n denotes the population size.

SSA incorporates an alert mechanism to simulate the escape behavior of sparrow populations when facing danger during foraging. This allows individuals trapped in local optima to execute random escape behaviors, improving the algorithm's capability to avoid local minima. The position update rule under the alert mechanism is expressed as:

$$L_{i,j}^{t+1} = \begin{cases} L_b^t + \lambda \cdot |L_{i,j}^t - L_b^t|, & f_i > f_b \\ L_{i,j}^t + \beta \cdot \left(\dfrac{L_{i,j}^t - L_w^t}{|f_i - f_w| + \varepsilon}\right), & f_i = f_b \end{cases} \tag{3}$$

In this context, L_b^t indicates the global best position, λ is the step-size control parameter, and β is a random number within [-1,1]. The variable f_i denotes the fitness of the current individual, while f_b and f_w represent the best and worst fitness values, respectively. A small constant ε is used to prevent division by zero.

SSA dynamically adjusts the weight coefficients of discoverers and followers to improve the efficiency of the search process. Compared with Grid Search (GS) [16] and Random Search (RS) [17], SSA offers a more efficient exploration of the hyperparameter space. In the early evolutionary stages, discoverers are assigned higher weights to strengthen global exploration. As iterations progress, the weights of followers increase to enhance local exploitation. This adaptive mechanism effectively prevents premature convergence and enhances the algorithm's optimization performance, resulting in significantly improved diagnostic accuracy under complex conditions.

2.2. CNN-Based Local Feature Extraction Module

CNNs utilize local receptive fields and weight sharing to efficiently extract discriminative local fault features from raw vibration signals [18]. This module employs a multi-layer convolutional architecture to hierarchically construct feature representations from low to high levels. Convolution, the core component of CNNs, extracts local features from input data using sliding kernels. The mathematical formulation is as follows:

$$y_{i,j}^{l} = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} w_{m,n}^{l} \cdot x_{i+m,j+n}^{l-1} + b^{l} \tag{4}$$

Where $y_{i,j}^{l}$ denotes the feature map output at position (i, j) in the $l-$th layer; $w_{m,n}^{l}$ is the kernel weight; $x_{i+m,j+n}^{l-1}$ is the input from the previous layer; b^{l} is the bias; M and N are the height and width of the kernel, respectively.

In rolling bearing fault diagnosis, pooling layers play a critical role in convolutional neural networks. They reduce the spatial dimensionality of vibration feature maps, thereby decreasing computational complexity. Typically placed between consecutive convolutional layers, pooling effectively preserves fault-related features crucial for diagnosis. It also mitigates overfitting and improves generalization. Given the characteristics of bearing fault diagnosis, max pooling is used as follows:

$$p_{i,j}^{l} = \max_{(m,n)\in R_{i,j}} (y_{m,n}^{l}) \tag{5}$$

Where $R_{i,j}$ represents the pooling window. The activation function employed is the Swish function [19]:

$$f(x) = x \cdot \sigma(\beta x) = \frac{x}{1+e^{-\beta x}} \tag{6}$$

Here, β is a learnable parameter initialized to 1.0 in the experiments.

Accurate extraction of local spatial features in CNNs is essential for fault detection. This is accomplished by using a multi-branch convolutional structure that combines features via concatenation, facilitating multi-scale feature fusion. Local response normalization (LRN) is subsequently applied, and the resulting formula is given as:

$$y_{i,j}^{k} = \frac{x_{i,j}^{k}}{(1+\frac{\alpha}{N}\sum_{k'=max(0,k-N/2)}^{min(K-1,k+N/2)} (x_{i,j}^{k'})^2)^{\beta}} \tag{7}$$

In this equation, $x_{i,j}^{k}$ and $y_{i,j}^{k}$ represent the input and output features, respectively. N is the number of adjacent channels for normalization, K is the total number of channels in the feature map, α is the scaling factor, typically set to 0.0001, and β is the exponent coefficient, typically set to 0.75. Local response normalization achieves competitive inhibition along the channel dimension, thus emphasizing fault-sensitive features.

2.3. Global Temporal Modeling with Transformer

Transformer [20] is a deep learning framework based on the self-attention mechanism, demonstrating significant advantages in time-series signal processing. The self-attention mechanism allows the model to compute relationships between any two positions in the sequence, enabling it to capture global dependencies. This mechanism computes interactions between three components: query (Q), key (K), and value (V). It autonomously learns the strength of dependencies in vibration features at different time points, using scaled dot-product attention. The formula is given by:

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{T}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{8}$$

In this equation, $\boldsymbol{Q} \in \mathbb{R}^{n\times d_k}$ is the query matrix, $\boldsymbol{K} \in \mathbb{R}^{m\times d_k}$ is the key matrix, and $\boldsymbol{V} \in \mathbb{R}^{m\times d_k}$ is the value matrix, $\sqrt{d_k}$ is a scaling factor used to prevent gradient vanishing. For vibration signals $\boldsymbol{X} \in \mathbb{R}^{L\times d_{\text{model}}}$, the attention calculation process is given by:

$$\boldsymbol{X}_{\boldsymbol{out}} = \text{Attention}(\boldsymbol{X}W^{Q}, \boldsymbol{X}W^{K}, \boldsymbol{X}W^{V}) \tag{9}$$

Here, W^{Q}, W^{k}, W^{v} are the learnable parameters.

The multi-head attention layer is a key component of the Transformer framework, enabling multi-dimensional feature extraction through parallel attention mechanisms. To enhance feature extraction, this module is extended to hhh parallel attention heads, as shown in the following formula:

$$\begin{cases} \text{MultiHead}(Q,K,V) & = \text{Concat}(head_1,...,head_h)W^{O} \\ head_i & = \text{Attention}(QW_i^{Q}, KW_i^{K}, VW_i^{V}) \end{cases} \tag{10}$$

In this context, $\text{MultiHead}()$ refers to the output of the multi-head attention mechanism, and W^{o} is the parameter matrix used to concatenate $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$. The multi-head design enables the model to capture time-frequency features in different subspaces, focusing on both local and global patterns, which improves robustness against noise interference.

To improve training stability, a pre-normalization (Pre-LN) residual structure is adopted. Layer normalization is applied before each sub-layer, followed by the calculation of the attention or feed-forward network outputs. Finally, residual connections are used to combine the input and output, as shown mathematically by:

$$\begin{aligned} X' &= X + \text{Dropout}(\text{MultiHead}(\text{LayerNorm}(X))) \\ X'' &= X' + \text{Dropout}(\text{FFN}(\text{LayerNorm}(X'))) \end{aligned} \tag{11}$$

In this formula, $\mu = \frac{1}{d}\sum_{i=1}^{d} x_i$ is the mean of the features, and $\sigma^2 = \frac{1}{d}\sum_{i=1}^{d}(x_i-\mu)^2$ is the variance. γ, $\beta \in \mathbb{R}^{d}$ are the learnable scaling and shifting parameters, and $\varepsilon = 10^{-5}$ is a constant for numerical stability. $\odot$ represents element-wise multiplication.

2.4. SSA-CNN-Transformer-Based Intelligent Bearing Diagnosis Model

The SSA-CNN-Transformer intelligent bearing diagnosis model proposed in this study comprises three core modules: the SSA hyperparameter optimization module, the CNN local feature extraction module, and the Transformer global temporal modeling module, as shown in Figure 1. By introducing the SSA, based on the sparrow foraging behavior mechanism, the model automatically optimizes key hyperparameters, such as the CNN convolution kernel size and the number of attention heads in the Transformer, during training. This improves the model's convergence speed and performance stability under complex operating conditions.

In feature extraction, the CNN module utilizes a multi-scale convolutional structure combined with the local response normalization mechanism to precisely capture local fault features, such as transient impacts and periodic modulations, in vibration signals. The Transformer module then takes the high-dimensional feature sequence extracted by the CNN as input and employs the multi-head self-attention mechanism to model long-range temporal dependencies. The incorporation of positional encoding and residual structures enhances the model's ability to capture potential evolutionary patterns in non-stationary signals.

The overall framework facilitates an end-to-end intelligent diagnosis process, from hyperparameter adaptive configuration to the fusion of local and global features. This not only improves fault recognition accuracy and robustness but also offers an efficient and feasible solution for rolling bearing health monitoring in complex industrial scenarios.

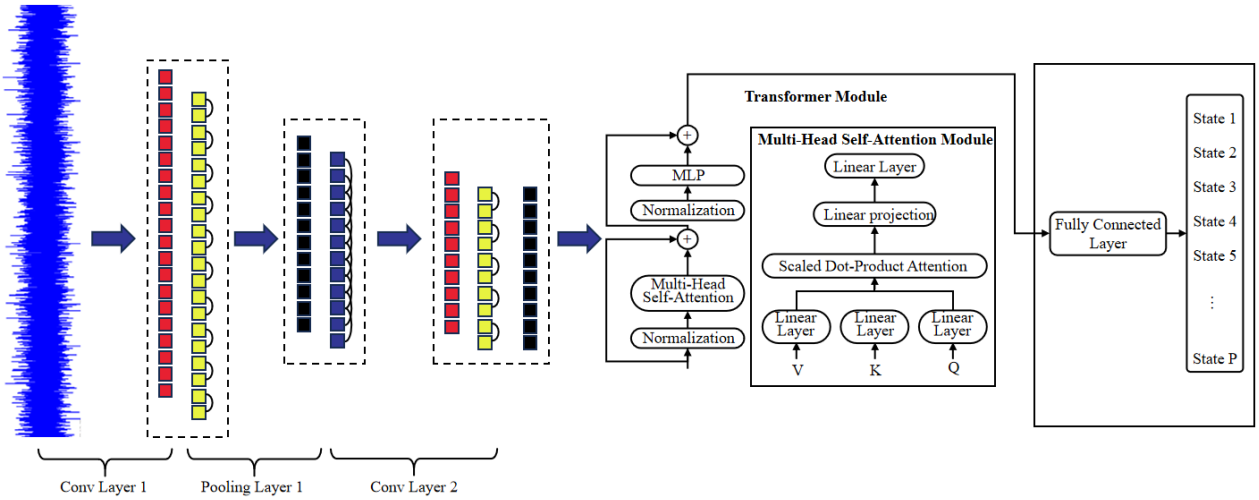


Figure 1. SSA-CNN-Transformer-Based intelligent bearing diagnosis framework

3. Experimental Design and Performance Evaluation

3.1. Dataset Description

This study selects three representative bearing fault public datasets to validate the generalization performance and adaptability of the proposed SSA-CNN-Transformer model under multi-source and multi-operating conditions. These datasets encompass vibration signals from industrial, laboratory, and aerospace high-speed environments. The details are as follows:

(1) Case Western Reserve University Bearing Dataset (CWRU Dataset) [21]: Collected by the Department of Mechanical Engineering at Case Western Reserve University, this dataset is widely used in rolling bearing fault diagnosis research. The dataset includes four bearing states: Normal (Normal), Inner Race Fault (IF), Outer Race Fault (OF), and Ball Fault (BF), all of which are preset single-point damages. Vibration signals were recorded using a 16-channel vibration recorder, with various load combinations (0hp, 1hp, 2hp, 3hp) and rotational speeds (1730, 1750, 1772, 1797 r/min), at a sampling frequency of 12kHz. Power and speed data were synchronized using torque sensors. This dataset exhibits typical controllability of operating conditions and fault repeatability.

(2) Xi'an Jiaotong University Bearing Dataset (XJTU Dataset) [22]: Collected from the lifetime prediction experimental platform at Xi'an Jiaotong University, this dataset includes a drive motor, gearbox, controller, and load unit, capable of simulating bearing operating conditions across their entire lifecycle. The experiment designed four bearing fault states: OF, IF, BF, and Combined Fault (CF), with a sampling frequency of 20480Hz. To ensure consistency, only bearing data were selected for modeling and analysis. The data contains abundant degradation information and noise interference, making it ideal for assessing the model's robustness under varying operating conditions and life stages.

(3) Politecnico di Torino Aerospace Bearing Dataset (DIRG Dataset) [23]: Collected by the Department of Mechanical and Aerospace Engineering at Politecnico di Torino, Italy, this dataset focuses on the performance degradation of high-speed aerospace bearings under extreme conditions. The data were recorded on a dedicated high-speed test bench, capturing vibration signals at speeds up to 30,000 rpm, under various damage levels, sensor arrangements, and load conditions. The sampling frequency is 51200Hz. Bearing damage was induced using Rockwell tools through indentation, with fault levels ranging from 0A (normal) to 6A (severe damage), totaling seven categories. In this study, 14 status signals were selected under two typical load conditions, at a shaft frequency of 200Hz, to serve as evaluation samples for high-dynamic and complex operating conditions, thoroughly testing the model's generalization ability.

3.2. Data Preprocessing and Experimental Setup

In order to ensure both the efficiency of model training and the reproducibility of results, all experiments were carried out on a local computing server featuring a 13th Gen Intel® Core™ i9-13900H CPU and an Intel® Iris® Xe Graphics unit. The entire experimental procedure was implemented in a Python 3.9 environment. Model construction and training were carried out using the PyTorch framework, and evaluation metrics were computed with the Scikit-learn machine learning library, establishing a comprehensive end-to-end training and validation pipeline.

During the data preprocessing stage, Min-Max normalization was applied to map all vibration signal samples to the [0,1] range, standardizing the scale of various input features and facilitating faster model convergence. Furthermore, to improve training stability, the dataset was partitioned into 60% for the training set, 10% for the validation set, and 30% for the test set, ensuring that model tuning and performance evaluation remained independent. As depicted in Figure 2, both the loss and classification accuracy rapidly converged within 50 iterations, indicating that the proposed model exhibits strong fitting capability and stability during training under complex data distributions.

For performance evaluation, this study employs three key classification metrics to assess model performance comprehensively: accuracy, which evaluates overall discriminative ability; recall, which measures the identification of positive class samples; and F1-score, which balances precision and recall, making it especially useful for addressing imbalanced class distributions in complex classification tasks. By analyzing these multidimensional metrics in coordination, the proposed model's generalization ability and diagnostic accuracy across various operating conditions and diverse sample sets can be thoroughly evaluated.

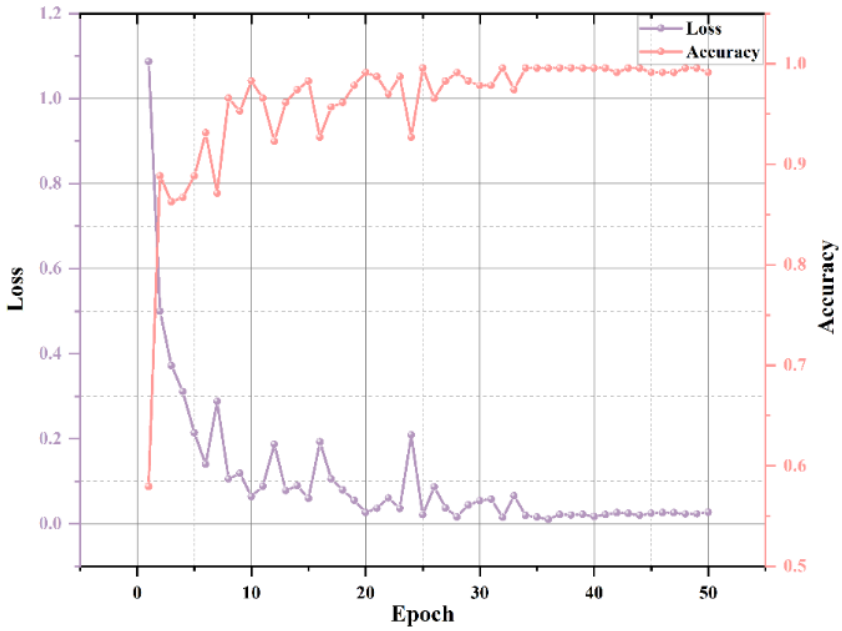


Figure 2. Convergence curves of loss and accuracy during training

3.3. Analysis of experimental results

In deep neural networks, hyperparameter configuration plays a critical role in determining convergence speed, representational capacity, and generalization ability. To address the subjectivity and uncertainty of manual tuning, this study adopts the Sparrow Search Algorithm (SSA) to perform global optimization of critical structural and training hyperparameters in the CNN-Transformer model. The objective is to enable intelligent, systematic, and performance-optimized hyperparameter configuration.

Given the model's multi-module architecture, six key hyperparameters were selected for optimization: the number of layers and initial channel count in the CNN branch; the hidden dimension and number of encoder layers in the Transformer module; the number of attention heads in the multi-head attention mechanism; and the learning rate used during training. The hyperparameter search space was defined within a reasonable range based on prior knowledge, ensuring both search efficiency and comprehensive coverage of potential optimal solutions. Table 1 summarizes the search ranges and corresponding optimal values for each hyperparameter obtained through SSA.

The optimization results indicate that the CNN module adopts a three-layer architecture, which effectively mitigates overfitting while preserving strong feature extraction capability. The initial number of channels is optimized to 27, enhancing the resolution of low-level features and suppressing redundancy. The Transformer module's hidden dimension is optimized to 54, achieving a balanced trade-off between model complexity and representational capacity. The Transformer's encoder depth and the number of attention heads are both optimized to 2, indicating that a lightweight structure is sufficient to capture global dependencies in vibration signals. This demonstrates the effectiveness and robustness of the self-attention mechanism in capturing fault-related temporal features. Furthermore, the learning rate is optimized to 0.0003, which significantly improves convergence speed and generalization performance compared to conventional values.

Table 1. Hyperparameter optimization results

Hyperparameter	Search Range	Optimal Value
Number of CNN Layers	[1-4]	3
Initial Number of CNN Channels	[16-128]	27
Transformer Hidden Dimension	[16-128]	54
Number of Transformer Encoders	[1-4]	2
Number of Attention Heads	[1-4]	2
Learning Rate	[0.0001-0.1]	0.0003

Overall, the SSA-driven optimization enables the proposed SSA-CNN-Transformer model to achieve a well-balanced configuration between structural complexity and training stability. Experimental results demonstrate that the optimization strategy not only improves classification accuracy across diverse datasets but also significantly enhances training convergence. These findings validate the broad applicability and practical value of SSA in hyperparameter optimization for deep diagnostic models. From the comparison between prediction data and actual data, the BP neural network has better prediction performance and relatively small error, which can meet the demand completely, and has fast prediction speed and convenient operation.

3.4. Results and Analysis

3.4.1. Comparison of Multi-Model Performance and Advantage Validation

To comprehensively assess the fault diagnosis performance of the proposed SSA-CNN-Transformer model, five benchmark methods were selected for comparison across three representative publicly available bearing datasets. These methods include the K-Nearest Neighbors (KNN) algorithm, Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), standard Convolutional Neural Network (CNN), and the unoptimized CNN-Transformer model. Table 2 presents the classification accuracy results of each method across different datasets.

Table 2. Performance comparison of different models on three datasets

Methods	CWRU	XJTU	DIRG
KNN	85.62	78.65	81.66
SVM	81.47	74.37	78.42
MLP	90.65	84.64	86.73
CNN	92.93	89.01	90.35
CNN-Transformer	96.81	91.82	93.14
SSA-CNN-Transformer	99.52	94.98	97.85

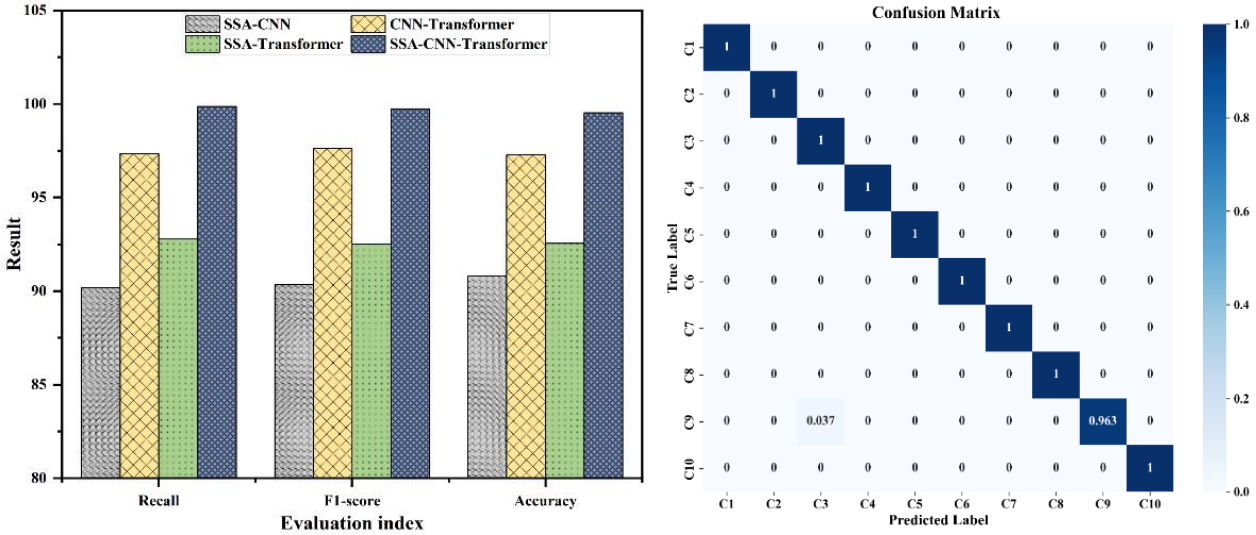
The experimental results show that traditional machine learning models, such as KNN and SVM, exhibit relatively low classification performance across all three datasets, highlighting their limited ability to model high-dimensional, nonlinear vibration features. Specifically, the accuracy of SVM on the CWRU, XJTU, and DIRG datasets were 81.47%, 74.37%, and 78.42%, respectively, which is significantly lower than that of deep learning-based models. While KNN performs slightly better on certain datasets, such as CWRU, its overall stability and generalization ability remain limited. In comparison, the MLP model shows slight improvement due to its nonlinear mapping capability; however, it remains suboptimal in non-stationary signal environments due to its limited ability to model temporal features.

Further comparison shows that the standard CNN model outperforms the shallow models, validating the effectiveness of the convolutional structure in spatial feature extraction. After introducing the Transformer structure, the model’s global modeling capability is significantly enhanced, and the CNN-Transformer model's accuracy further improves, reaching 96.81%, 91.82%, and 93.14% on the three datasets, respectively. This demonstrates the advantages of the local-global collaborative architecture.

It is worth emphasizing that the proposed SSA-CNN-Transformer model outperforms the CNN-Transformer model, achieving the best performance across all three datasets, with an accuracy improvement of 2–5 percentage points. This fully demonstrates the critical role of intelligent hyperparameter optimization in enhancing the model's expressive power and generalization performance. In conclusion, the SSA-CNN-Transformer not only maintains outstanding performance under conventional conditions but also demonstrates excellent robustness and transferability under more challenging cross-platform and cross-load conditions, validating its broad adaptability and application potential in practical engineering fault diagnosis.

3.4.2. Ablation Study

To evaluate the contribution of each component in the proposed SSA-CNN-Transformer model, an ablation study was conducted. Three simplified variants were developed: SSA-CNN (containing only the CNN architecture with SSA optimization), SSA-Transformer (containing only the Transformer architecture with SSA optimization), and CNN-Transformer (a complete model without SSA optimization). All models were evaluated under identical experimental settings and dataset configurations using three key metrics: recall, F1-score, and accuracy, as illustrated in Figure 3(a).



(a) Results of the ablation study (b). Visualization of the confusion matrix

Figure 3. Results of the experiment

The SSA-CNN model demonstrated baseline-level performance across all three metrics. This suggests that while CNN effectively captures local features, it lacks the capacity to model global temporal dependencies. The SSA-Transformer model exhibited slightly improved performance, achieving an F1-score of 92.53% and accuracy of 92.56%, highlighting the Transformer's advantage in modeling global contextual dependencies.

After integrating the CNN and Transformer modules, the CNN-Transformer model achieved a significant performance improvement, with an F1-score of 97.62% and accuracy of 97.27%. This confirms the effectiveness of combining local convolutional features with global attention mechanisms in joint modeling. Notably, the full SSA-CNN-Transformer model, incorporating SSA-based hyperparameter optimization on top of the fusion architecture, achieved the best performance across all metrics. It consistently outperformed all other ablation variants, demonstrating superior diagnostic accuracy and robustness.

Additionally, a confusion matrix was analyzed to provide a more intuitive evaluation of the model's diagnostic performance, as shown in Figure 3(b). In summary, the ablation study clearly validates the independent utility of each module and the synergistic effect of their integration. In particular, the incorporation of the SSA optimization strategy significantly enhances feature representation and parameter adaptability, serving as a key enabler for high-performance intelligent fault diagnosis.

4. Conclusion

To address the challenge of insufficient accuracy in rolling bearing fault diagnosis under non-stationary and multi-condition environments, where multi-scale fault features and long-range dependencies in vibration signals are difficult to jointly model, we propose an intelligent diagnostic model based on SSA-CNN-Transformer to improve diagnostic robustness and accuracy in complex operating conditions.

The model utilizes SSA for global adaptive optimization of key hyperparameters, significantly improving the efficiency and performance of the model's configuration. In feature extraction, the CNN

module effectively captures local time-frequency features from vibration signals, while a multi-scale fusion strategy enhances feature representation. The Transformer module is incorporated to model the global dependencies in long time series, enabling deep feature representation and high-accuracy classification of complex fault patterns. The performance of the SSA-CNN-Transformer fault diagnosis model was compared to that of other mainstream models. Experimental results show that, the SSA-CNN-Transformer model demonstrates significant advantages in fault diagnosis, validating its broad applicability and potential for real-world engineering fault diagnosis.

However, in practical engineering applications, high-quality and accurately labeled training data are often difficult to obtain, limiting the broader adoption of the model. Therefore, a key direction for future research is to develop an effective diagnostic model based on existing equipment maintenance data. In addition, considering the imbalance of class distribution commonly encountered in real-world fault diagnosis scenarios, future work will explore advanced strategies such as stratified sampling, data augmentation, and cost-sensitive learning to further improve the model's robustness and diagnostic performance for minority classes.

References

[1] Zhiqin Z, Yangbo L, Guanqiu Q, et al. A review of the application of deep learning in intelligent fault diagnosis of rotating machinery [J]. Measurement, 2023, 206.

[2] Fang X, Zheng J, Jiang B. A rolling bearing fault diagnosis method based on vibro-acoustic data fusion and fast Fourier transform (FFT) [J]. International Journal of Data Science and Analytics, 2024, (republish): 1 - 10.

[3] Zhao S, Liang X, Wang L, et al. A fault diagnosis method for analog circuits based on EEMD-PSO-SVM [J]. Heliyon, 2024, 10 (18): e38064 - e38064.

[4] Feng L, Zhu Y, Xu S, et al. Open-circuit fault diagnosis of DC charging pile rectifier based on sparse data and CNN-ISSA-BiLSTM [J]. Energy Reports, 2025, 133024 - 3034.

[5] Bharatheedasan K, Maity T, Kumaraswamy L, et al. Enhanced fault diagnosis and remaining useful life prediction of rolling bearings using a hybrid multilayer perceptron and LSTM network model [J]. Alexandria Engineering Journal, 2025, 115355 - 369.

[6] Wu M, Zhang J, Xu P, et al. Bearing Fault Diagnosis for Cross-Condition Scenarios Under Data Scarcity Based on Transformer Transfer Learning Network [J]. Electronics, 2025, 14 (3): 515 - 515.

[7] Eyup S, Sezgin K, Suleyman U. A new deep learning model combining CNN for engine fault diagnosis [J]. Journal of the Brazilian Society of Mechanical Sciences and Engineering, 2023, 45 (12): 644.

[8] Hu H, Feng F, Zhu J, et al. Research on Fault Diagnosis Method Based on Improved CNN [J]. Shock and Vibration, 2022, 2022.

[9] F. X T, B. Y L. Integration of gradient least mean squares in bidirectional long short-term (LSTM) memory networks for metallurgical bearing ball fault diagnosis [J]. Metalurgija, 2024, 63 (3-4): 403 - 406.

[10] Kangjie C, Ting Z, Jueqiao H. Advanced hybrid LSTM-transformer architecture for real-time multi-task prediction in engineering systems [J]. Scientific Reports, 2024, 14 (1): 4890 - 4890.

[11] Wu M, Zhang J, Xu P, et al. Bearing Fault Diagnosis for Cross-Condition Scenarios Under Data Scarcity Based on Transformer Transfer Learning Network [J]. Electronics, 2025, 14 (3): 515 - 515.

[12] Yu G, Yu J L, Hui Z. Efficient Hyperparameter Optimization for Convolution Neural Networks in Deep Learning: A Distributed Particle Swarm Optimization Approach [J]. Cybernetics and Systems, 2020, 52 (1): 36 - 57.

[13] Xue J, Shen B. A novel swarm intelligence optimization approach: sparrow search algorithm [J]. Systems Science & Control Engineering, 2020, 8 (1): 22 - 34.

[14] Liu R, Wang X, Su C, et al. Bearing fault diagnosis method based on variational mode decomposition optimized by CS-PSO [J]. Journal of Vibration and Control, 2024, 30 (5-6): 973 - 987.

[15] Yang X, Jiang A, Jiang W, et al. Abnormal Detection and Fault Diagnosis of Adjustment Hydraulic Servomotor Based on Genetic Algorithm to Optimize Support Vector Data Description with Negative Samples and One-Dimensional Convolutional Neural Network [J]. Machines, 2024, 12 (6): 368.

[16] Mustafa A, Yunus A. A novel hybrid PSO- and GS-based hyperparameter optimization algorithm for support vector regression [J]. Neural Computing and Applications, 2023, 35 (27): 19961 - 19977.

[17] Arbi J S, Rehman U Z, Hassan W, et al. Optimized machine learning-based enhanced modeling of pile bearing capacity in layered soils using random and grid search techniques [J]. Earth Science Informatics, 2025, 18 (4): 332 - 332.

[18] Yu T, Ren Z, Zhang Y, et al. A rolling bearing fault diagnosis method based on a new data fusion mechanism and improved CNN [J]. Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability, 2024, 238 (6): 1156 - 1169.

[19] B A, Kalirajan K. An intelligent magnetic resonance imagining-based multistage Alzheimer's disease classification using swish-convolutional neural networks. [J]. Medical & biological engineering & computing, 2024, 63 (3): 1 - 15.

[20] Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need [J]. arXiv, 2017.

[21] CWRU: Case Western Reserve University Bearing Data Center Website, URL http://csegroups.case.edu/bearingdatacenter/home

[22] XJTU: The emerging graph neural networks for intelligent fault diagnostics and prognostics: A guideline and a benchmark study.

[23] DIRG: The Politecnico di Torino rolling bearing test rig: Description and analysis of open access data, Mech. Syst. Signal Process.