Journal of Engineering Mechanics and Machinery (2024)
Clausius Scientific Press, Canada
DOI: 10.23977/jemm.2024.090105
ISSN 2371-9133 Vol. 9 Num. 1

Identification of time-frequency maps of bearing faults based on hyperparameter optimization SSA-GoogleNet

Ziqiang Luo

School of Mechanical and Automation Engineering, Wuhan University of Science and Technology, Wuhan, 430065, China

Keywords: 2D wavelet transform, convolutional neural network, sparrow optimization algorithm, hyperparameter optimization

Abstract: In order to solve the problem of poor noise immunity of neural network in bearing fault detection and to meet the high demand for adaptive extraction ability of network features in the industrial field, this paper proposes a hyper-parameter optimization method to eliminate the error of human-set parameters. Aiming at the non-smoothness and non-linear characteristics of the bearing fault signal, the two-dimensional wavelet transform (2DWT) is used to extract the feature values of the vibration signal of the bearing fault, and it is proposed to use the CNN convolutional neural network (CNN) to train the fault model. Convolutional Neural Network (CNN) is proposed to be used for fault model training. Firstly, the 2D wavelet transform is applied to the bearing vibration signal to extract the time-frequency map of the vibration signal; secondly, the extracted time-frequency map is used as the training object of the convolutional neural network to train the model, and then multiple convolutional neural network frameworks are used for noise immunity test by adding Gaussian white noise to the original signal to construct the test framework, and then the framework with the best noise immunity is used as the base network; and then the CNN is used as the base network. Sparrow Search Algorithm (SSA), Grey Wolf Optimizer (GWO), and Whale Optimization Algorithm (GWO) to find the best parameters. Validated by the public dataset of Western Reserve University, the method can effectively improve the fault identification accuracy of the model and can get rid of the problem of poor network robustness.

1. Introduction

Bearings in modern industry are widely used in machinery manufacturing, the automotive industry, aerospace, and other fields, playing an important role in mechanical connection, support, and friction reduction. With the development of industrial technology, the requirements for mechanical equipment are getting higher and higher [1], and the use of bearings is getting more and more extensive. The bearing is one of the most easily damaged parts, and the service state of the bearing will directly affect the operation of mechanical equipment, which puts forward high requirements for condition monitoring and fault diagnosis of bearings [2]. In engineering practice, once the bearings fail, it will lead to mechanical equipment cannot operate normally, so real-time fault monitoring and diagnosis of bearings is necessary [3].

Mechanical equipment in the fault state is accompanied by vibration changes that are different from the norm, and bearings in the process of work are produced by the vibration of nonlinear non-smooth characteristics [4], in the event of bearing failure, by the impression of nonlinear, friction and other factors, the vibration signal of the non-smooth characteristics of the vibration signal will be more obvious. Widely used and traditional signal analysis methods, such as the global transform of the Fourier transform are not suitable for the processing of non-smooth signals, wavelet transform, although the nonlinear signal generated by the bearing vibration has good time-frequency localization and multi-resolution characteristics, for the high-frequency part of the re-decomposition ability is weak. The empirical modal decomposition method is suitable for processing nonlinear and nonsmooth signal sequences. It does not need to pre-set any basis function, but it produces the "modal aliasing" problem that is more prominent [5], and the endpoint effect is obvious. It has certain limitations in the engineering application.

The traditional method makes it difficult to solve the bearing fault non-smooth signal analysis, need to introduce a series of adaptive algorithms for fault analysis, the current commonly used fault identification algorithms are more, such as support vector machine (SVM), random forest (RFA), gray correlation analysis (GRA), etc. SVM in the processing of high-dimensional data, sparse representations, and robustness of the theoretical advantages, but its selection of the parameters and kernel function has sensitivity, and the decision boundary does not have a sensitive, and the decision-making boundary. RFA has better generalization ability and robustness and can effectively deal with high-dimensional data and data with noise, but its computational complexity is higher and is prone to underfitting problems, and it is more sensitive to parameter settings, and GRA has a huge defect in classification accuracy.

Machine learning has developed rapidly in recent years, and the development of neural networks has also ushered in a period of rapid development. Machine learning plays an important role in image and audio processing and has become the focus of great attention in the academic and engineering communities. For example, in the acoustic pattern recognition of overlapping acoustic signals, Yanyan Bao et al. found that the use of CNN convolutional neural network can use its own network structure and parameters to adaptively extract the image features of foreground acoustic patterns in the overlapping signals layer by layer, which significantly improves the ability to accurately classify and recognize foreground acoustic patterns in the overlapping signals [6]. Considering the commonality between bearing fault diagnosis and pattern classification and recognition, CNN is applied to bearing fault diagnosis.

CNN consists of a convolutional layer, a pooling layer, and a fully connected layer. The network structure can effectively extract the spatial structure features of the input data, and the structure is adaptive, which can gradually extract the deep-level features, avoiding the subjectivity and limitations of manually constructing features. Considering the relationship between the training data's sample size and the data's dimension, the 2D wavelet transform is used for the feature extraction of the signal before training with the CNN.

In summary, this paper adopts 2D wavelet transform for signal feature extraction, obtains time-frequency diagrams that can reflect the bearing state information, and uses a convolutional neural network to train classifiers to realize fault diagnosis.

2. Basic Theory

2.1 Two-Dimensional wavelet transform

Wavelet transform is one of the techniques that represent a time-frequency analysis method. It utilizes a finite set of long and decaying wavelet basis functions that can provide information about the signal in both frequency and time [7]. The discrete wavelet transform provides a framework for multi-resolution representation compared to the continuous wavelet transform, which allows information at different scales and is more computationally efficient.

The discrete wavelet transform (DWT), on the other hand, can decompose the signal into two parts, high frequency, and low frequency, through the discretization of the continuous wavelet transform (CWT) to the discrete wavelet basis is generally to turn the translation factor b and the expansion factor τ into a power series structure [8], i.e.

$$\begin{cases} b = b_0^i \\ \tau = k b_0^i \tau_0 \end{cases} \tag{1}$$

Where $b_0 \neq 1$; τ_0 is a constant; and i ϵ Z corresponds to the discrete wavelet basis as

$$\varphi_{i,k}(t) = b_0^{-\frac{i}{2}} \varphi\left(b_0^{-i} t - k\tau_0\right) \tag{2}$$

The final discrete wavelet transform is

$$W_f(i,k) = \int f(t)\, \overline{\varphi_{i,k}(t)} dt \tag{3}$$

The 2D discrete wavelet transform decomposes the 2D data into 1D data by decomposing it into 1D data and then performing the 1D discrete wavelet transform, each time the wavelet transform is thought to be performed, the original data will be decomposed into 2 components, the high-frequency and the low-frequency components. Each time the 2D wavelet transform is performed, the original data will be decomposed into 4 components, which are the high frequency components HL, LH, HH, and the low-frequency component LL, and the transform schematic diagram is shown as Fig. 1.

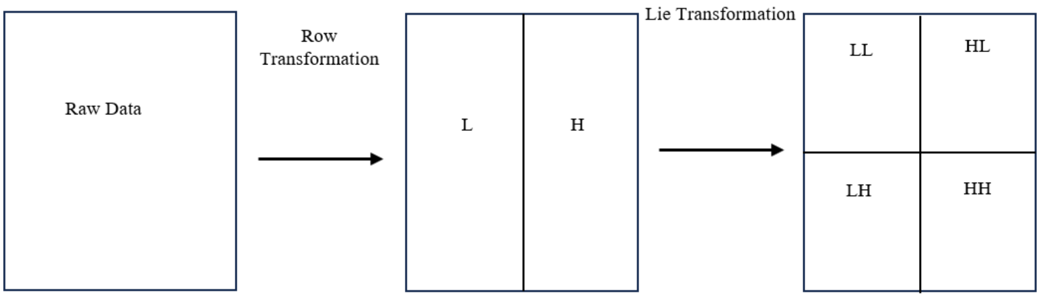


Figure 1: Principle of two-dimensional discrete wavelet transform

The principle of two-dimensional discrete wavelet transform is notated LL as A, LH as H, HL as V, and HH as D. D denotes the high-frequency component of the signal, V denotes the vertical component, H denotes the horizontal component, and A denotes the low-frequency component.

The image belongs to the classical two-dimensional signal, after the original map is subjected to the two-dimensional discrete wavelet transform, the low-frequency components of the image differ less from the original map, and the high-frequency signals reflect less information about the image.

2.2 CNN Convolutional Neural Network

Convolutional neural network (CNN) is a classical deep learning algorithm that generally consists of an input layer, a convolutional layer, a pooling layer, a fully connected layer, and an output layer. The schematic structure is shown in Fig.2. Convolutional neural networks do not need to manually select features [9] and share convolutional kernels, which provides a greater advantage in the processing of high-dimensional data. However, whether the output of forward propagation can meet the requirements, the timely update of other parameters such as weights during backpropagation plays an important role [4], which transforms the problem into finding the parameter with the smallest possible value of the loss function at the search. In deep neural networks, the large number of parameters makes the complexity of the optimization problem rise. An optimization framework is proposed below.

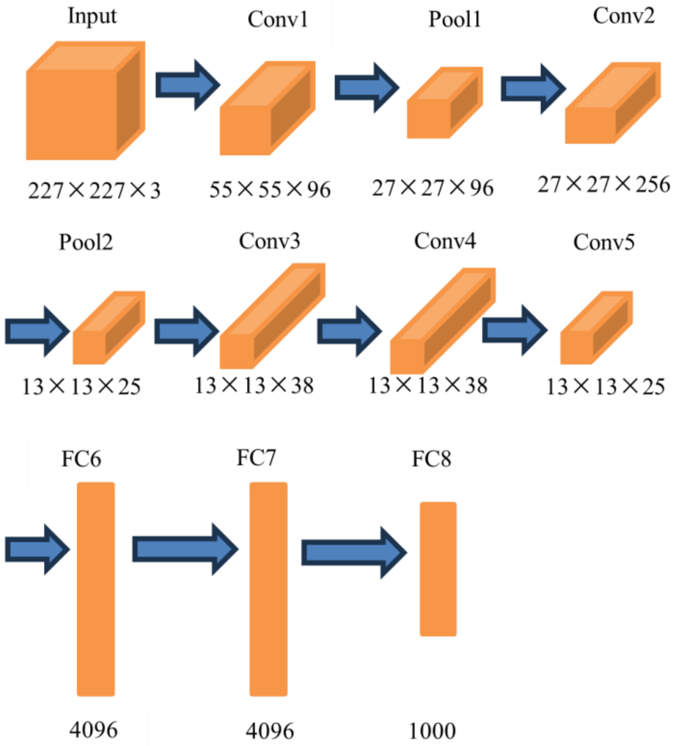


Figure 2: Schematic diagram of CNN convolutional neural network structure

2.3 SSA Sparrow Optimization Algorithm

In the continuous pursuit of solving complex optimization problems, the Sparrow Optimization Algorithm (SSA) emerged as an emerging optimization technique that mimics the group intelligence of nature. SSA originates from an in-depth study of the cooperative work and collective intelligence behavior of sparrow groups and aims to improve the efficiency and accuracy of problem-solving by mimicking such group intelligence [10]. In this section, we will introduce in detail the basic principles of SSA, its mathematical model, and its key features in optimization problems.

SSA draws on the collaborative behaviors of sparrow populations as they search for food and adapt to changes in their environment. Information sharing and collaboration among individuals enable the whole group to converge on the potential solution of the problem more quickly. SSA is adaptive, i.e., it can dynamically adjust the search strategy according to the complexity and characteristics of the problem. This makes the SSA algorithm perform well when facing various types of optimization problems.

The rule for updating the position of a sparrow individual in the search space can be expressed as

$$X_i(t+1) = X_i(t) + V_i(t) \tag{4}$$

where $X_i(t)$ is the position of the ith sparrow individual at moment t and $V_i(t)$ is its velocity vector. The velocity update rule of a sparrow individual can be expressed as

$$V_i(t+1) = \omega V_i(t) + c_1 r_1 \left(P_{best} - X_i(t)\right) + c_2 r_2 \left(G_{best} - X_i(t)\right) \tag{5}$$

where ω is the inertia weight, c_1 and c_2 are where ω is the inertia weight, c_1 and c_2 are the individual and group learning factors, respectively, r_1 and r_2 are random numbers in the range of [0,1], P_{best} is the historical optimal position of an individual, and G_{best} is the historical optimal position of the whole group.

SSA, as an optimization algorithm based on natural group intelligence, demonstrates its potential in solving complex optimization problems by simulating the collective behavior of sparrow groups. Its unique characteristics of group synergy and adaptivity make SSA have a wide range of prospects and potential in multi-disciplinary applications.

3. Experimental Verification

3.1 Experimental Data Processing

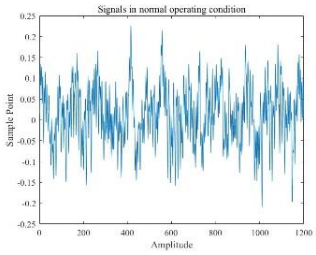
To verify the feasibility of the fault diagnosis method proposed in this paper, a dataset of open-bearing failures from Case Western Reserve University was used as a sample.

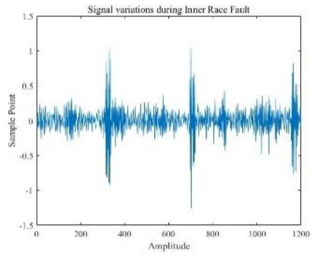
The dataset was obtained at a sampling frequency of 12 kHz and covered four different loading conditions. The inner ring failure, outer ring failure, and rolling element failure of the bearings were simulated, in which each failure type has three damage levels, 0.18 mm, 0.36 mm, and 0.54 mm for the vibration signal data collection.

To validate the diagnostic performance of the algorithm for early faults, a load of Load3 = 3HP/1730 RPM and a damage size of 0.014 ft are selected, and the experimental data set is shown in Table 1. Normal signals are verified against fault signals in three different states. Figure 3 demonstrates the time-frequency plots of the vibration signals for the four bearing states.

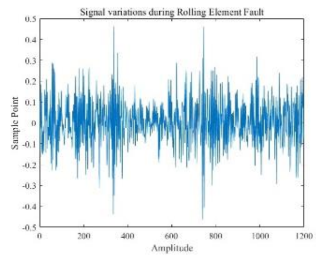
Table 1: Data sets with different bearing states

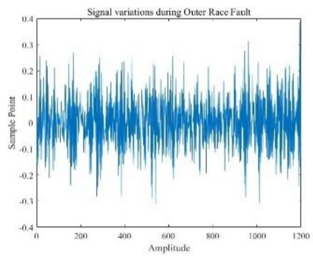
Typology	Label	Sample Size	Training Set	Validation Set
Normal	0	100	70	30
Inner Ring	1	100	70	30
Rolling Element	2	100	70	30
Outer Ring	3	100	70	30





(a) Normal operating condition (b) Inner Race Fault





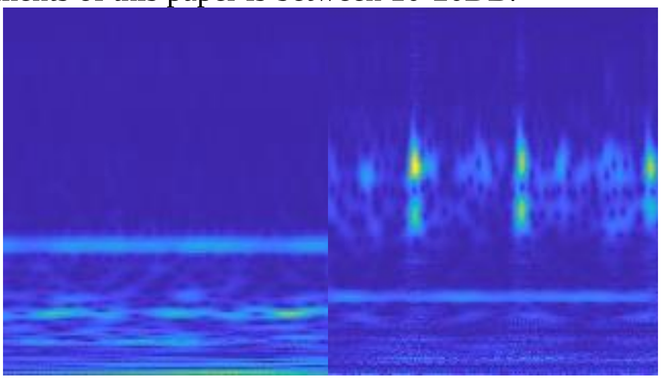
(c) Rolling Element Fault (d) Outer Race Fault

Figure 3: Signal changes in different states

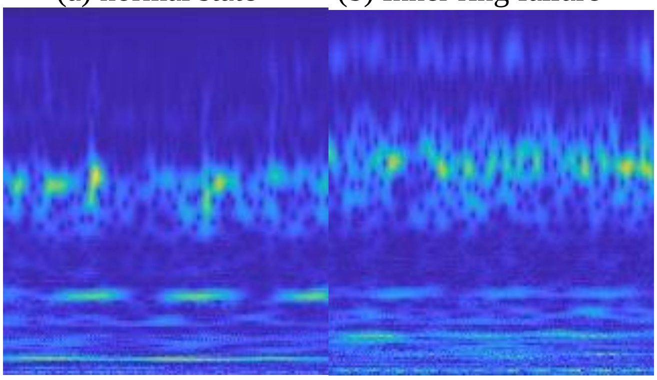
As can be seen from Figure 3, the vibration signal amplitude of the bearing in the normal state is small, and the signal amplitude increases in the fault state. However, the time-domain waveforms of the vibration signals in some states are relatively similar, which, together with the interference of the external environment, makes it difficult to distinguish the fault state of the bearing. Therefore, it is necessary to extract the features of the vibration signal to provide a dataset for the subsequent CNN training.

Combined with the time-frequency diagrams of the vibration signals, we choose the cmor wavelet basis with better adaptability, and perform a two-dimensional wavelet transform on each sample vibration signal under the above four states to obtain the time-frequency diagrams used for CNN training, as shown in Figure 4.

Considering the actual situation of mechanical equipment operation, some time-frequency maps with noise will be collected and used as the training of fault diagnosis model for CNN. The noise range used in the experiments of this paper is between 10-20DB.



(a) normal state (b) Inner ring failure



(c) Rolling body failure (d) Outer ring failure

Figure 4: 2D Wavelet Transform Time-Frequency Plot Base network selection and CNN fault detection model training

Considering the high efficiency and practicality in industrial inspection, lightweight neural networks are used as the base network for training, in this paper, GoogleNet, ShuffleNet, and MobileNet-v2 are used, and the parameters of the three networks are shown in the following Table 2.

Table 2: Base-network parameter

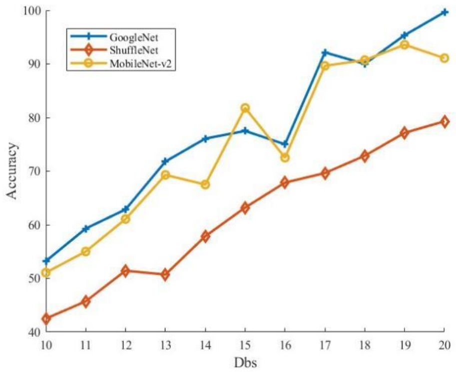
Network	Depth	Size (MB)	Parameters (M)	Input Size
GoogleNet	22	27.0	7.00	224×224×3
ShuffleNet	50	5.2	1.40	224×224×3
MobileNet-v2	53	13.5	3.50	224×224×3

The base network is trained using data without added noise and the pre-trained network model is loaded before CNN training. After several tests and validation, a set of optimal learning parameters are identified as shown in the following Table 3.

Table 3: CNN learning parameter settings

Batch Size	Learning Rate	Learn Rate Schedule	Maximum number of iterations
32	10-4	0.25	10

According to the above learning parameter settings, the pre-trained GoogleNet, ShuffleNet, and MobileNet-v2 network models are brought into the CNN using the Adam adaptive momentum algorithm to sequentially test the classification accuracy of the data with signal-to-noise ratios of 10-20 DB, and the results are shown in Fig. 5.



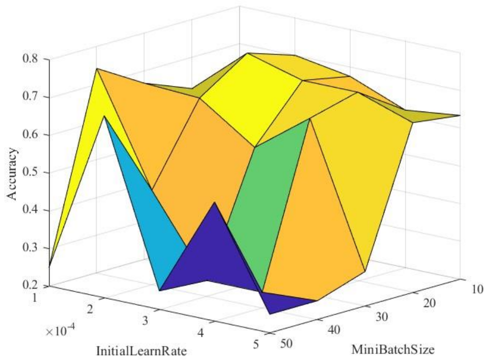


Figure 5: Classification Accuracy of Three Base Networks

Figure 6: Effect of learning parameters on accuracy

From the figure, we can see that the classification accuracies of GoogleNet and MobileNet-v2 are comparable, and the training efficiency of GoogleNet is higher than that of MobileNet-v2; the classification efficiency of ShuffleNet is higher compared to the other two networks, but its classification accuracy is much lower than that of other networks. In the end, GoogleNet is chosen as the base network in this paper.

Considering that the learning parameters of the convolutional neural network may have an impact on the training of the convolutional neural network, it may be worthwhile to use the number of batch sampling and the learning rate as the parameters, and the accuracy rate as the result, and use the grid method for training to test the impact of the learning parameters on the training of the neural network, and the training results are shown in Fig. 6.

In Fig. 6, the accuracy of network training is higher when the learning rate is lower and the number of batch samples is smaller. It is worth noting that when the number of batch samples is 30, the accuracy of network training decreases more significantly, indicating that the setting of parameters has a large impact on the training of neural networks. Considering the time and hardware cost factors, the network needs to be optimized.

3.2 Hyper-parameter optimization

In order to find a suitable algorithm for optimizing the neural network, the following three algorithms, WOA, SSA, and GWO, are used for optimization comparison, and based on their iteration curves, the most suitable algorithm for optimizing the parameters of batch sampling number and learning rate of the neural network used in this paper is found. The parameters such as batch sampling number and learning rate optimization interval are shown in the following Table 4.

Table 4: Optimized parameter setting

Batch Size	Learning Rate	Search Agents	Max Iterations
10~100	10^{-3}~10^{-5}	2	5

According to the data set in Table 4, the three algorithms WOA, SSA, and GWO are optimized, and the iteration curves are shown in Figure 7. The selection of optimization algorithms is carried out based on the results of the iterations.

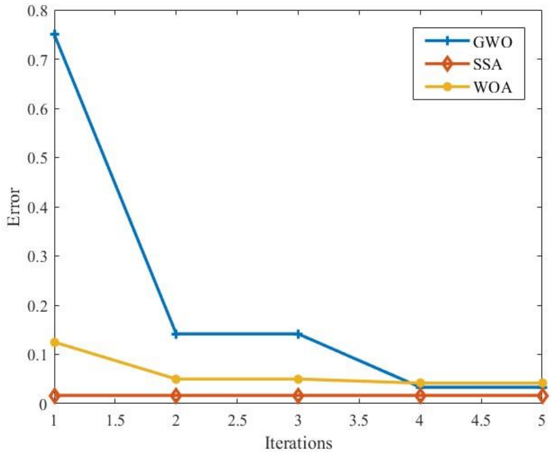


Figure 7: Iteration curves for the three algorithms

In Fig. 7, the gradient of error decline is most pronounced for GWO as the number of iterations increases, and the errors of SSA and WOA change more smoothly. Their error convergence is shown in Table 5.

Table 5: Comparison of errors of three algorithms

Algorithms	Number of iterations	Error (%)
WOA	5	4.17
SSA	5	1.67
GWO	5	3.33

Combined with Fig. 7 and Table 5, the stability and accuracy of SSA is more compared to WOA and GWO. In this paper, SSA is chosen as the optimization algorithm of neural network. Finally, the average correct rate of the SSA-Googlenet-based bearing fault time-frequency image recognition method can reach 98.33%, which verifies that the proposed method in this paper has good diagnostic accuracy.

4. Conclusion of the Experiment

The paper proposes a method for analyzing the time-frequency map of bearing faults based on SSA-GoogleNet and applies the method to the analysis of bearing signals. The results show that the CNN neural network optimized by the SSA algorithm can achieve an accuracy of 98.33% by training the two-dimensional wavelet transform on the time-frequency signal map of bearings with noise interference, which verifies that the method can achieve higher and more reliable diagnostic performance.

The method has the following advantages: feature extraction of the dataset can filter part of the noise, which helps to reduce the time needed to train the model; in addition, the CNN has a strong self-learning ability, which helps to improve the accuracy of the model. At the same time, the SSA optimization algorithm has the characteristic of updating rich population diversity, which reduces the probability of the algorithm falling into the local optimum.

References

[1] Shen J, Xu F. Method of fault feature selection and fusion based on poll mode and optimized weighted KPCA for bearings. Measurement, 2022(194):194. DOI: 10.1016/j.measurement.2022.110950.
[2] Lin Li, Xining Zhang, Shuyu Liu, et al., "Bearing Fault Diagnosis Method Based on Two-Dimensional Empirical Wavelet Texture Domain Feature Adaptive Extraction," Journal of Xi'an Jiaotong University, vol. 55, no. 12, pp. 79-86, 2021.
[3] Gang Yang, Limu Qin, Kun Lv, et al., "Bearing Fault Diagnosis Based on Continuous Cross Wavelet Coherence Analysis and Adaptive CYCBD," Vibration and Shock, vol. 42, no. 21, pp. 17-28, 2023. DOI: 10.13465/j.cnki.jvs. 2023.21.003.
[4] Cong Wu, Mengnan Li, Kun Li, "Elevator Bearing Fault Detection Based on PBO and CNN," Information Technology, vol. 47, no. 4, pp. 73-78, 2023. DOI: 10.13274/j.cnki.hdzj.2023.04.014.
[5] Siyan He, Ya Liu, Xincheng Tian, "Bearing Fault Diagnosis Research Based on Wavelet Packet-AR Spectrum and Deep Learning," Computer Applications Research, vol. 36, no. 6, pp. 1758-1761, 1766, 2019.
[6] Yuetian Wang, Sichao Fu, Qinmu Peng, et al., "Graph Convolutional Neural Network for Multi-Perspective Information Interaction in Semi-Supervised Scenes," Journal of Software, 2023-12-01. DOI: 10.13328/j.cnki.jos.007007.
[7] Junqing Li, Jing Liu, "Motor Bearing Fault Diagnosis Method Combining Convolutional Neural Network and Transfer Learning," Journal of North China Electric Power University (Natural Science Edition), vol. 50, no. 1, pp. 76-83, 91, 2023. DOI: 10.3969/j.ISSN.1007-2691.2023.01.09.
[8] Qinye He, Hanyu Lu, Yongyi Yuan, et al., "Research on Brightness Temperature Downscaling Method Based on Two-Dimensional Wavelet Transform," Journal of Guangxi University (Natural Science Edition), vol. 47, no. 3, pp. 746-755, 2022.
[9] Yanyan Bao, Guangze Yang, Wei Chen, et al., "Voiceprint Recognition of 750kV Transformer and Bushing Discharge Mixed Signal Based on SBSS and CNN," Journal of Southwest Jiaotong University, 2023-12-13.
[10] S. Mirjalili, "Dragonfly algorithm: A new optimization algorithm for solving discrete and continuous optimization problems," Neural Computing and Applications, vol. 27, no. 4, pp. 1053-1073, 2016.