IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 18 October 2023, accepted 13 December 2023, date of publication 16 January 2024, date of current version 23 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3354794

RESEARCH ARTICLE

Optimization of Deep Belief Network Based on Sparrow Search Algorithm for Rolling Bearing Fault Diagnosis

DONGHAO XU AND CHENG LI
College of Air Transportation, Shanghai University of Engineering Science, Shanghai 201620, China
Corresponding author: Cheng Li (Licheng@sues.edu.cn)

ABSTRACT This study addresses the randomness of training parameters in the Deep Belief Network (DBN) and proposes an optimization method for rolling bearing fault diagnosis based on the Sparrow Search Algorithm (SSA). SSA is employed to globally optimize the structural and training parameters of the DBN network, effectively resolving the challenge of parameter determination. Simultaneously, vibration signals are extracted from multiple dimensions to capture different types of fault features. These features are derived through Wavelet Transformation (WT) for noise reduction and Intrinsic Mode Functions (IMFs) extraction through Ensemble Empirical Mode Decomposition (EEMD). The fusion of time-domain and frequency-domain dimensional features forms a multidimensional feature set. This comprehensive feature set optimizes the parameters of the deep learning network and significantly improves the accuracy and effectiveness of rolling bearing fault diagnosis. With a remarkable recognition accuracy of 99.17%, this approach outperforms conventional feature sets and mainstream diagnostic methods such as PSO-DBN and SSA-SVM while maintaining high levels of generalization and stability. The introduction of this method represents a significant breakthrough in the field of rolling bearing fault diagnosis.

INDEX TERMS Deep belief network, fault diagnosis, multi-domain features, sparrow search algorithm.

I. INTRODUCTION

China's future manufacturing system planning is marked by critical terms such as the ''14th Five-Year medium and long-term manufacturing development plan'' and ''intelligent manufacturing''. These terms represent the direction of China's manufacturing industry and reflect the global trend in manufacturing development. China has made significant advancements in its manufacturing capabilities, transitioning from being known as the ''world's factory'' to becoming an ''intelligent manufacturing power''. The continuous upgradation of industries and technologies has been crucial in supporting the steady growth of China's economy. Rolling bearings serve as vital components in various large equipment and mechanical parts. They perform several functions under challenging working conditions and operating loads to ensure the safe operation of automated systems. With the continuous increase in the emphasis on the reliability of industrial products, there have been increased demands for more accurate and effective fault diagnosis systems. The AI technologies [1], such as machine learning, deep learning, and pattern recognition, have the advantages of enhancing fault diagnosis accuracy, predictive maintenance, and fault prognosis. These fault diagnosis systems are critical for promptly identifying potential issues in rolling bearings. Through the early detection of faults, they aid in preventing unexpected failures, minimizing downtime, and optimizing maintenance schedules. Moreover, an accurate and effective fault diagnosis system facilitates proactive maintenance strategies, resulting in improved productivity, reduced costs, and enhanced overall system performance. Incorporating the background of big industrial data, national policy guidelines, and current industry trends, researchers have made significant progress by integrating rolling bearing fault diagnosis with machine learning algorithms, neural networks, and deep learning techniques. This integration [2], [3], [4] has effectively addressed the challenges of modern complex industrial equipment.

The associate editor coordinating the review of this manuscript and approving it for publication was Geng-Ming Jiang.

Indeed, the Deep Belief Network (DBN) is recognized as among the most representative and strongly compatible algorithms in various fields, including fault diagnosis. Researchers have increasingly favoured DBN [5] owing to its advantages over traditional fault diagnosis methods. Thus, fault diagnosis has evolved into a multi-method fusion pattern recognition process. Although diagnostic models are crucial to the overall process, the significance of signal processing, feature extraction, and other related techniques [6], [7] cannot be understated.

Typically, the vibration signal of the rolling bearing includes pulse signals, external signals, background noise signals, etc. The complexity of the vibration signal directly corresponds to the progress of the follow-up work. A corresponding signal analysis method is required to perform the related analysis. Wavelet decomposition and reconstruction can effectively remove noise, complete the preliminary optimization of the vibration signal, and retain the original signal characteristics. This [8] renders it suitable for vibration signal processing of rolling bearings. There [9], [10], [11] exist several novel methods for signal processing, among which the most popular is the modal decomposition algorithm. Considering that the parameters of such algorithms affect the signal decomposition effect under manual intervention, this study employed the wavelet analysis algorithm with powerful functions and deep industry experience. In the feature extraction stage, Zhang and Huang [12] used the Empirical Mode Decomposition (EMD) algorithm to decompose the vibration signal into a set of inherent mode functions. Kumar et al. [13] proposed a frequency mode based on the Variational Mode Decomposition (VMD) signal to monitor the bearing health state. Ni et al. [14] identified a superior feature extraction method by comparing VMD, EMD, and Local Mean Decomposition (LMD). Complete Ensemble Empirical Mode Decomposition (CEEMDAN) [15] has been employed to determine specific fault characteristics, and TFR demodulation analysis has been used to obtain accurate fault characteristics. Regarding pattern recognition, Kang et al. [16] proposed the deep domain adaptation method, wherein convolutional and pooling theories were integrated with DBN to solve the problem of multi-state identification of rolling bearings. Xu and Tse [17] combined DBN with the Affinity Propagation (AP) model, which exhibited excellent results compared to traditional fault diagnosis methods. Zhong et al. [18] used the improved fault diagnosis method that combined Ensemble Empirical Mode Decomposition (EEMD) and DBN to achieve fault diagnosis. The intelligent fault diagnosis method of PCA-DBN was proposed [19], which reduced the dimension of complex features before completing the fault diagnosis. Therefore, to obtain a good fault diagnosis result based on the DBN network model [20], the initial parameters of the model must be improved and optimized. Deng et al. [21] proposed an improved quantum-inspired differential evolution (MSIQDE) algorithm, which avoided premature convergence, improved global search ability, and optimized DBN parameters using MSIQDE with global optimization ability. Furthermore, Gao et al. [22] employed the intelligent optimization method Salp Swarm Algorithm to optimize DBN, effectively improving its classification accuracy.

The remainder of this paper was organized as follows. Section II presented different methods applied in the study to extract various features. Section III presented the feature extraction and model optimization process to prepare multi-domain data sets for experiments. Section IV presented the assignment of the experimental data sets and the comparison of results between this paper and the mainstream methods. Section V summarized the whole study and the future work.

II. METHODS

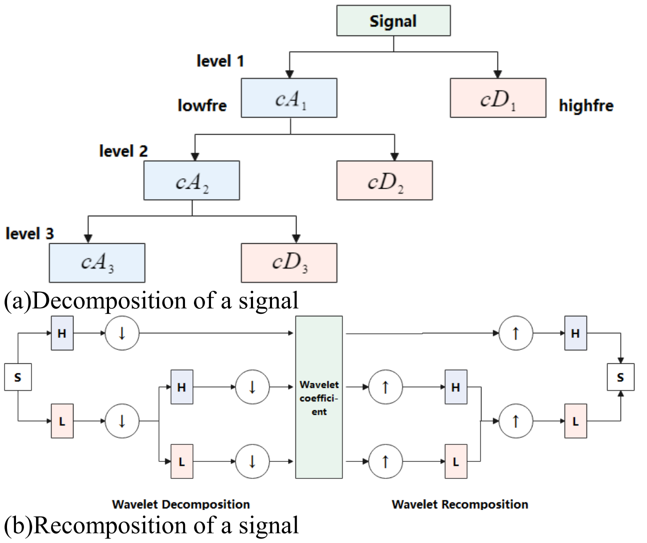
A. WAVELET DECOMPOSITION AND RECONSTRUCTION

The Short-time Fourier Transform (STFT) is an evolution of the traditional Fourier Transform (FT). The size and shape of the window function in STFT remain fixed and independent of time, rendering it unsuitable for analysing time-varying signals. An excessively narrow window function frame can result in poor frequency resolution, whereas a wider frame can result in poor time resolution. This limitation prevents STFT from satisfying the frequency requirements of unsteady signal changes. However [23], the wavelet transform differs from the STFT as it abandons the infinite trigonometric function basis and adopts a finite and decaying wavelet basis. This transformation facilitates both frequency information and accurate time localization. In wavelet transform, frequency information can be obtained while accurately determining the specific time location of a signal. The specific expression is as follows:

$$\mathrm{WT}\left(\alpha,\tau\right)=\frac{1}{\sqrt{\alpha}}\int_{-\infty}^{+\infty}f\left(t\right)*\psi\left(\frac{t-\tau}{\alpha}\right)dt \qquad (1)$$

Eq.(1) shows that in contrast to FT, the wavelet transform incorporates two variables: scale α and position τ. Scale α controls the scaling of the wavelet function, whereas position τ determines its translation. Scale α is inversely proportional to frequency, and position τ corresponds to time. The wavelet analysis method outperforms traditional Fourier analysis in denoising non-stationary signals. Fig.1 shows a diagram of a signal's wavelet decomposition and recomposition.

Wavelet noise reduction facilitates decorrelation. In wavelet analysis [24], signal decomposition is performed using the Mallat tower algorithm, resulting in approximate and detailed signals at each decomposition level. This study employed the Daubechies (Db) wavelet owing to its suitability for rolling bearing fault characteristics. The Daubechies wavelet was characterized by its outstanding orthogonality, effectively reducing information loss during wavelet transformation and inverse operations. Compared to other wavelet functions, it offered a notably refined time resolution in its designated time domain. Furthermore, there was a noticeable enhancement in the smoothness and continuity of the processed



(a)Decomposition of a signal

(b)Recomposition of a signal

FIGURE 1. Wavelet decomposition and recomposition.

signal as the wavelet coefficients increased. Samsingh [25] employed the db4 wavelet to denoise medical images and juxtaposed its performance with other techniques. The superior efficacy of the db4 wavelet was distinctly manifested in his comparisons.

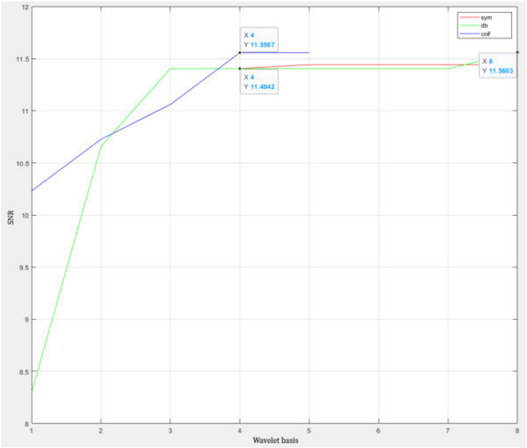


FIGURE 2. Relationship between wavelet basis and SNR.

In Fig.2, when selecting the optimal wavelet base function, the SNR post-Daubechies wavelet denoising experiences a significant surge, plateauing around the db4 mark. Remarkably, all three wavelet functions nearly reach their SNR zenith when the base is set at 4, with only marginal gains observed beyond this point. Employing higher-order wavelet base functions introduced the potential for signal over-decomposition and an upswing in computational demand. Given these findings, this study earmarked the Daubechies db4 wavelet as the focal point for vibration signal denoising. In addition, the number of decomposition levels should be chosen while considering the trade-off between separation effectiveness and noise reduction during reconstruction. A suitable scale of three decomposition levels was selected to address this trade-off. This choice resulted in the clear separation of noise and signal and an excellent final noise reduction effect after reconstruction. Based on the principles above and influencing factors of wavelet noise reduction, the original signal comprising 20,480 sampling points was processed using wavelet noise reduction.

B. ENSEMBLE EMPIRICAL MODE DECOMPOSITION

Limitations, such as mode aliasing and end effects, plague the traditional EMD method in signal processing. These deficiencies can result in periodic signals and a loss of physical meaning, ultimately affecting signal decomposition accuracy. A novel method, called EEMD [26], was proposed to address these issues. The EEMD method tackles the problem of signal decomposition accuracy, particularly mode aliasing, by introducing noise-assisted computation. EEMD involves introducing random noise into a given time-domain signal, performing multiple EMDs on the signal with added noise, and then averaging the resulting IMFs across the decompositions to obtain a more robust decomposition, effectively separating different frequency components within the original signal. The IMF is defined as a single-component signal obtained through decomposing the original signal after noise reduction. The detailed decomposition steps are depicted in Fig.3. Consequently, the EEMD method offers a reliable solution for achieving accurate signal decomposition.

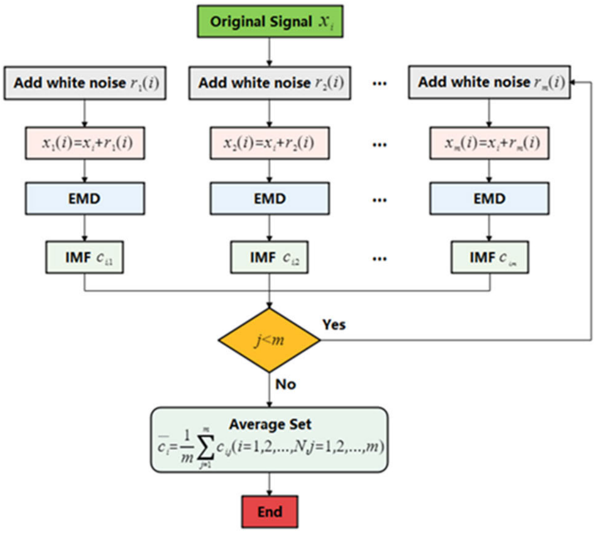


FIGURE 3. Flowchart of the EEMD algorithm.

C. SPARROW SEARCH ALGORITHM

SSA is a swarm optimization algorithm that outperforms existing algorithms regarding convergence speed, stability, and local optima avoidance. It achieves this by simulating the foraging and anti-predation behaviour of sparrow populations. A new meta-heuristic algorithm called the Sparrow Search Algorithm (SSA) [27] is proposed to optimize the operation of microgrids. Furthermore, an improved version of the sparrow search algorithm [28] is utilized to solve time-optimal trajectory problems. Numerous researchers have acknowledged and affirmed the optimization capabilities of SSA, and its feasibility has been demonstrated through comparisons with ample data from multiple regions. The specific flow of the algorithm is as follows.

The population and fitness function composed of n sparrows are expressed as follows:

$$X = \begin{bmatrix} x_1^1 & x_1^2 & \cdots & x_1^d \\ x_2^1 & x_2^2 & \cdots & x_2^d \\ \cdots & \cdots & \cdots & \cdots \\ x_1^n & x_2^n & \cdots & x_n^d \end{bmatrix}, F_x = \begin{bmatrix} f\left(\left[x_1^1 \;\; x_1^2 \;\; \cdots \;\; x_1^d\right]\right) \\ f\left(\left[x_2^1 \;\; x_2^2 \;\; \cdots \;\; x_2^d\right]\right) \\ \vdots \\ f\left(\left[x_1^n \;\; x_2^n \;\; \cdots \;\; x_n^d\right]\right) \end{bmatrix} \tag{2}$$

where d is the dimension of the variable to be solved and f is the fitness function.

In SSA, the efficiency of food discovery is directly proportional to the fitness value. After each iteration of the algorithm, the discoverer in the sparrow population continually searches for food and updates its position and direction. The number of discoverers typically accounts for approximately 10-20% of the total population. The expression for updating the position can be described as follows:

$$X_{i,j}^{t+1} = \begin{cases} X_{i,j}^{t+1} \cdot \exp\left(-\dfrac{i}{\alpha \cdot iter_{max}}\right), R_2 < \mathrm{ST} \\ X_{i,j}^{t} + Q \cdot L, R_2 \geq \mathrm{ST} \\ a \in (0, 1], R_2 \in [0, 1], \mathrm{ST} \in \left[\dfrac{1}{2}, 1\right] \end{cases} \tag{3}$$

where t is the number of iterations, $iter_{max}$ is the maximum number of iterations, L is a matrix of $1 \times d$, α is the random number in the range, R_2 is the early warning value, and the range is the safe value.

Its position update expression can be described as:

$$X_{i,j}^{t+1} = \begin{cases} Q \cdot \exp\left(\dfrac{X_{worst} - X_{i,j}^{t}}{i^2}\right), & i > \dfrac{n}{2} \\ X_p^{t+1} + \left|X_{i,j}^{t} - X_p^{t+1}\right| \cdot A^{+} \cdot L, & i \leq \dfrac{n}{2} \end{cases} \tag{4}$$

where X_p^{t+1} is the optimal position of the discoverer in the global when iteration $t+1$ is reached and X_{worst} is the worst position of the population in the global.

D. DEEP BELIEF NETWORK

DBN is a Restricted Boltzmann Machines (RBM) type that combines low-level features with other nonlinear transformations. DBN is a deep learning model incorporating Neural Networks and Backpropagation Neural Network (BPNN) to capture high-level abstract features. It has been extensively used in various fields, such as classification, prediction, and speech recognition, wherein it has showcased remarkable performance and thus established itself as a leading approach in fault classification and diagnosis owing to its advantageous combination of features. RBM serves as the foundation of the DBN network model. Its structure comprises independent layers without internal communication between them. Each node processes input data units and independently decides whether to pass on the input based on random judgment. The parameters of RBM are randomly initialized, enabling the calculation of the probability for each neuron individually. By multiplying these probabilities, RBM estimates the activation of the entire layer of neurons. Consequently, the computational complexity is reduced, and the connections within the visible and hidden layers are eliminated. Thus, no connections exist between visible or hidden units, as shown in Fig.4.

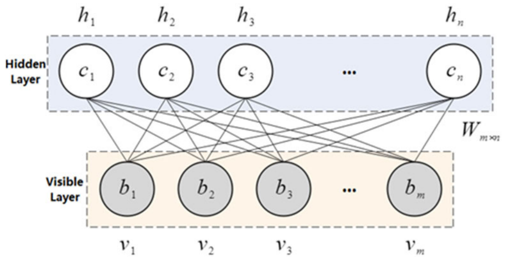


FIGURE 4. RBM infrastructure diagram.

By leveraging the main idea of unsupervised learning, the fault feature identification in the DBN can be achieved by adding a Softmax output layer at the top. This facilitates supervised learning techniques, wherein labels are used to evaluate and analyze the entire dataset. Through this process, effective classification and prediction can be achieved. Regarding the structural characteristics, the training model for the DBN involves establishing the initial model and fixing the weights w and bias values b, c of the first RBM layer. Subsequently, the RBM network of each layer was trained sequentially and stacked on top of each other. This greedy layer-by-layer training process was optimized through multiple iterations, forming the fundamental DBN model.

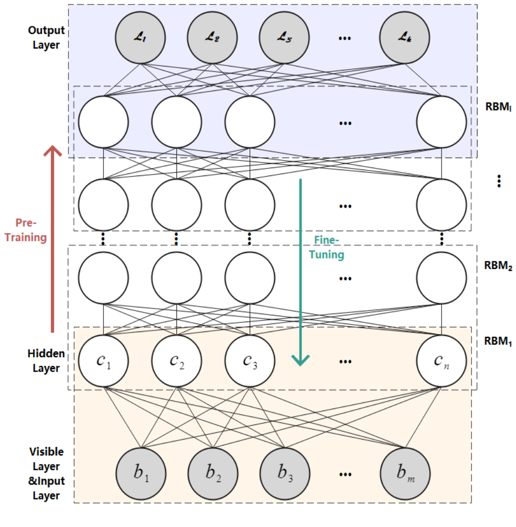


FIGURE 5. DBN structure model.

Fig.5 shows a basic RBM structure composed of m visible and n hidden neurons, where v is the visible layer of input data, h is the hidden layer of feature extraction, w is the weight between layers, and b, c are the bias values of the visual and hidden layers, respectively. The random variable x was divided into two groups to represent it as: $h = \{h_1, h_2, h_3, \cdots, h_n\}$, $v = \{v_1, v_2, v_3, \cdots, v_m\}$, and many connections were omitted inside the visual and hidden layers. The expression of its energy function is expressed as:

$$E(v, h) = -\left(\sum_{i,j=1}^{m,n} v_i w_{ij} h_j + \sum_{i=1}^{m} b_i v_i + \sum_{j=1}^{n} c_j h_j\right) \tag{5}$$

Substituting the energy function into the probability density function yields the final form of RBM:

$$P(x) = P(v, h) = \frac{1}{z} \cdot e^{\sum\limits_{i,j=1}^{m,n} v_i w_{ij} h_j + \sum\limits_{i=1}^{m} b_i v_i + \sum\limits_{j=1}^{n} c_j h_j} \tag{6}$$

where $Z = \sum\limits_{v,h} e^{-E(v,h)}$ is the normalized factor, also known as the partition function, and it represents the sum of energy in all possible cases.

The formation of probabilities in an RBM involves calculating the energy of a specific state and dividing it by the sum of the energies of all possible states. The energy function in RBM follows the Boltzmann distribution, and it facilitates the expression of a joint probability density by continuously calculating the probabilities and connections between the two layers. RBM builds a unified energy model by combining energy functions with related probability distribution functions. The shared weights between the two layers of RBM determine the joint distribution probability.

The states of each unit in the visual and hidden layers are independent of each other, and the conditional probability expression as in Eqs.(7) and (8):

$$P(v, h) = \frac{P(v, h)}{P(h)} = \prod P(v_i h) \tag{7}$$

$$P(h, v) = \frac{P(h, v)}{P(v)} = \frac{\frac{1}{z} \cdot e^{-E(v,h)}}{\frac{1}{z} \cdot \sum\limits_{h} e^{-E(v,h)}} = \prod P\left(h_j, v\right) \tag{8}$$

The neuronal activation probabilities of the visual and hidden layers are defined in Eqs.(9) and (10):

$$P(v_i = 1, h) = \text{sigmoid}\left(\sum_{j=1}^{n} w_{ij} h_j + c_j\right) \tag{9}$$

$$P\left(h_j = 1, v\right) = \text{sigmoid}\left(\sum_{i=1}^{m} w_{ij} v_i + b_i\right) \tag{10}$$

The sigmoid activation function facilitates the hidden layer effect in the DBN network. Instead of receiving a linear function output from the previous layer, each node in the hidden layer transforms the output value using a nonlinear function such as the sigmoid function. This nonlinearity ensures the model's powerful expressive capabilities. Considering that the selected experimental task involved multi-classification, the sigmoid activation function was essential for achieving the hidden layer effect. Therefore, the activation function that yielded the maximum accuracy was chosen. Fig.4 shows the complete training process of a DBN network, beginning from the unsupervised pre-training of the input layer to the top layer. After optimizing the initial parameters of each layer, reverse supervised fine-tuning was conducted in combination with labelled data. During the reconstruction phase, the activation state of the hidden layer served as the input during the backward transmission process. Similar to weight adjustment in the forward transmission process, errors were reconstructed and backpropagated based on weight adjustments. Through continuous iterative learning, the errors were minimized until convergence was reached.

III. FEATURE EXTRACTION AND DBN OPTIMIZATION

A. SIGNAL NOISE REDUCTION

Assume a known signal with sampling frequency $F_s = 5120\text{Hz}$, sampling number $N = 1024$, and sampling step dt of $\frac{1}{F_s}$. A random noise was added in MATLAB using the randn() function to form a signal containing noise, as shown in Eq.(11):

$$\begin{aligned} x(i) = {} & \sin(2 * \text{pi} * 50 * i * \text{dt}) + 0.5 \\ & * \sin(2 * \text{pi} * 1500 * i * \text{dt}) + 1 \\ & * \sin(2 * \text{pi} * 3000 * i * \text{dt}) + 0.1 * \text{randn}(1, 1) \end{aligned} \tag{11}$$

Based on Eq.(11), the time-domain signal comprised components at frequencies of 50, 1500, and 3000Hz, and random noise. Among these, the 50Hz component represents the effective signal, whereas the others correspond to interference noise signals at different frequencies. Wavelet decomposition was applied to obtain wavelet coefficients, which were then used for signal reconstruction. To illustrate the impact clearly, the first 500 signals are shown in Fig.6. The resulting signals exhibited a noticeable effect of smooth noise reduction, ultimately producing a cleaner representation.

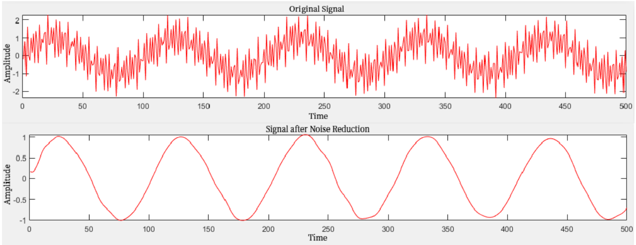


FIGURE 6. Simulation signal noise reduction.

This study employed the wavelet transform to reduce the noise present in the four operating states of the rolling bearing. This technique effectively extracted the smooth and meaningful components of the signal while eliminating the noise. After decomposing the signal using wavelet transform, the wavelet coefficients were used to reconstruct the denoised signal. Fig.7-10 provide a visual comparison between the

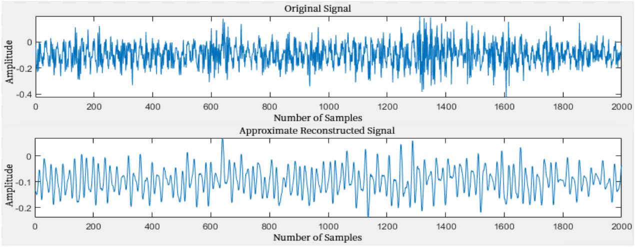


FIGURE 7. Normal signal after noise reduction.

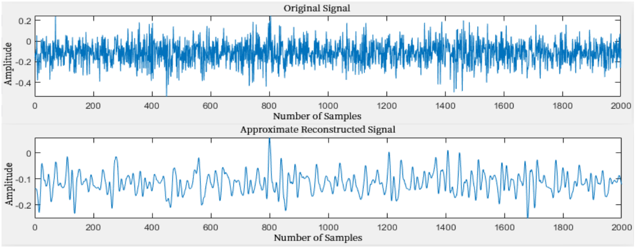


FIGURE 8. Inner race fault signal after noise reduction.

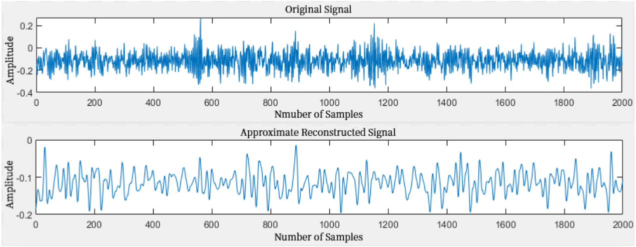


FIGURE 9. Rolling element fault after noise reduction.

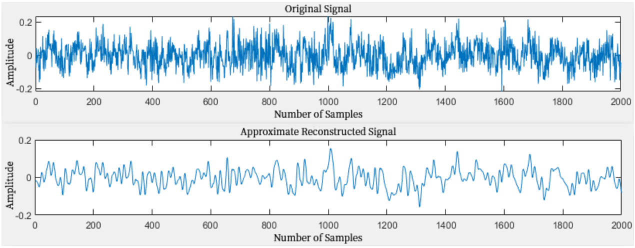


FIGURE 10. Outer race fault after noise reduction.

original signal and the denoised signal obtained through wavelet transform. It is apparent from these figures that wavelet transform exhibits remarkable denoising ability by removing the unwanted noise components from the original vibration signal. The vertical axis represents the amplitude, measured in the unit m/s^2, while the horizontal axis represents the number of samples. To simplify the figures and emphasize the impact of wavelet noise reduction, the first 2000 signals were considered.

Fig.7-10 provide above illustrates the standard vibration signal of the rolling bearing, along with the waveforms of the inner ring fault, rolling element fault, and signal after noise reduction for the outer ring fault. As evident, the vibration signal waveform became smooth and devoid of any sharp or jagged points after noise reduction. This denoising process preserved the essential characteristics of the original signal while effectively eliminating unwanted noise. Consequently, the denoised signal retained the integrity of the practical components while ensuring the removal of invalid noise.

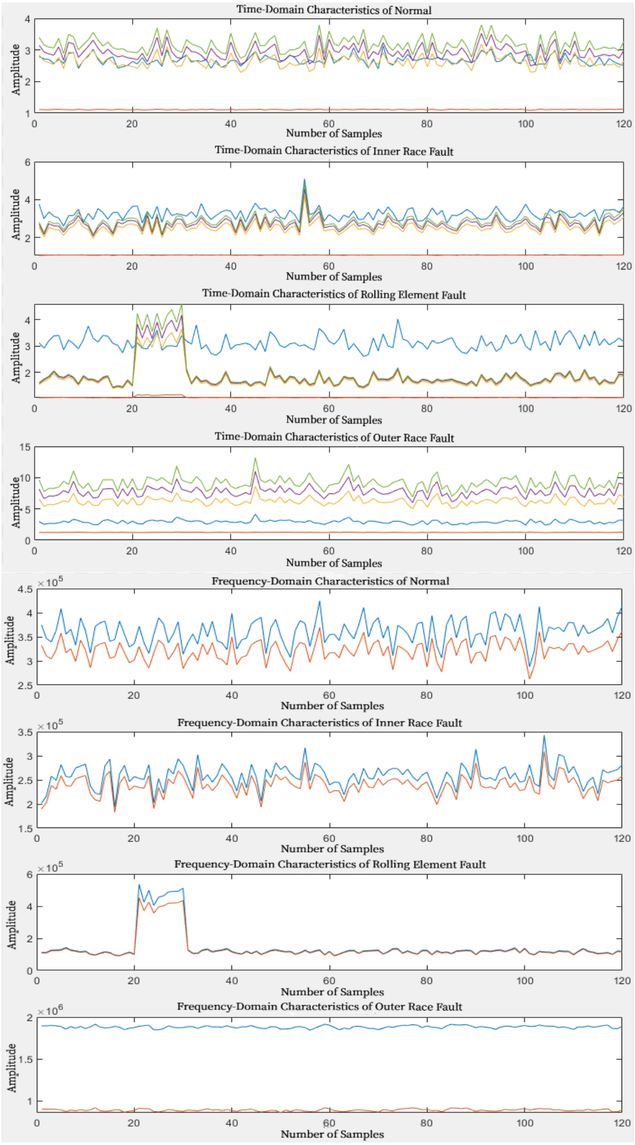


FIGURE 11. Time-domain and frequency-domain feature distribution diagram.

B. FEATURE EXTRACTION

After applying noise reduction, the sample data was subjected to combined time-domain and frequency-domain index analysis. The EEMD was then used to extract IMF energy features. The distribution of these features in the time-domain and frequency-domain is shown in Fig.11, which clearly illustrates the contribution of mean and effective values to fault diagnosis. Fig.12 shows the distribution of IMF energy features, demonstrating the retention of compelling IMF

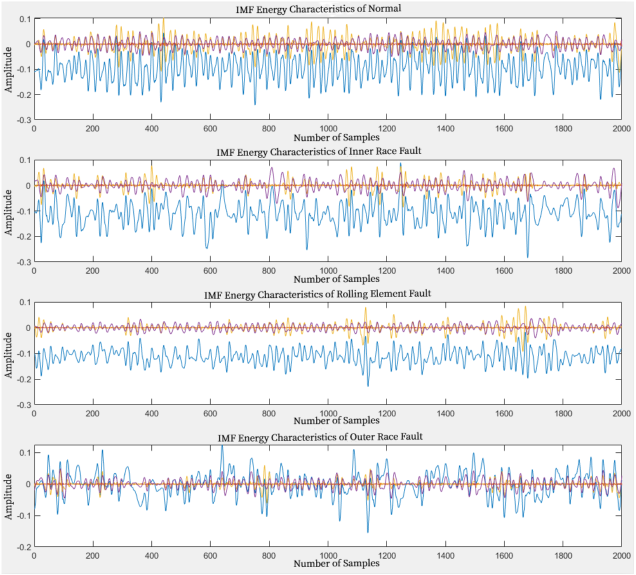


FIGURE 12. **Feature distribution diagram of IMFE.**

TABLE 1. **Feature parameters and expressions.**

Name of Parameter	Expression of Parameter
Effective value	$X_{rms}=\sqrt{\frac{1}{N}\cdot\sum_{i=1}^{1}x_i^2}$
Peak value	$X_p=\frac{1}{I}\cdot\sum_{i=1}^{I}max(x_i)$
Barycentric frequency	$f_{FC}=\frac{\sum_{i=1}^{N}f\cdot S(f)}{\sum_{i=1}^{N}S(f)}$
RMS frequency	$f_{MSF}=\frac{\sum_{i=1}^{N}f^2\cdot S(f)}{\sum_{i=1}^{N}S(f)}$
Frequency variance	$f_{VF}=\frac{\sum_{i=1}^{N}\left(f-f_{FC}\right)^2\cdot S(f)}{\sum_{i=1}^{N}S(f)}$

energy features while avoiding modal aliasing. The characteristic parameters corresponding to these features are listed in Table 1. When obtaining IMF component features through EEMD, an additional white Gaussian noise with a mean square error of 0.25 was introduced, with an overall average of 50 samples. This resulted in different characteristic parameters and energy curves, providing additional bearing information that could be reflected.

In Fig.11, the first one exhibited mean, effective, peak, variance, and skew distribution values. The second one exhibited kurtosis, peak, pulse, and margin factor distribution.

Figure 13 illustrates the variance contribution rates and Pearson correlation coefficients of various modal components post-EEMD decomposition for both denoised and raw signals. The denoised signal maintains a strong linear correlation with the original one and captures distinct frequency details through EEMD, resulting in a smoother curve. This combination of EEMD and db4 wavelet is instrumental in extracting diverse frequency features for bearing fault diagnosis. In contrast, the raw signal exhibits enhanced frequency periodicity after EEMD decomposition but lacks detailed frequency characteristics.

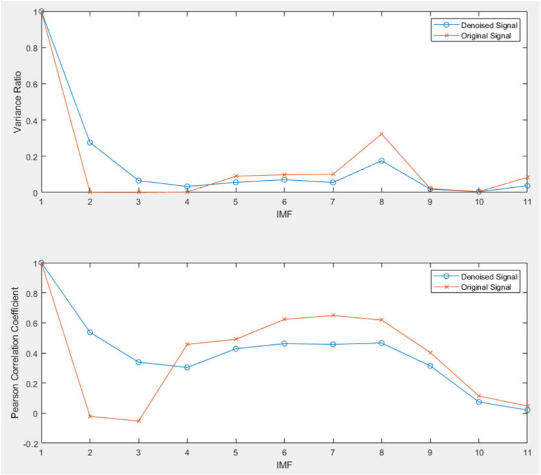


FIGURE 13. **Comparison of IMF related parameter trends.**

C. DBN OPTIMIZED BY SSA

The SSA [29] was employed to optimize the structure and weight parameters of the DBN. The results demonstrate that the recognition rate of the SSA-DBN model surpassed that of other classifiers, with a recognition accuracy approximately 2% higher than that of the unoptimized DBN model. SSA-DBN, VMD and Wigner-Ville distribution (WVD) [30] were used for intelligent fault severity detection. The model achieved an accuracy rate of 98%, indicating its effectiveness in fault detection. Li et al. [31] compared and verified the performance of DBN models combined with different optimization algorithms, including Simulated Annealing (SA), Particle Swarm Optimization (PSO), and SSA. The evaluation results indicated that all three improved DBN models outperformed the original DBN model. However, the SSA-DBN model achieved the highest evaluation accuracy among them.

When proposing a relatively new algorithm, further research and verification are necessary to assess its optimization effectiveness. In the case of the algorithm considered, which was proposed 2-3 years ago, it is expected that more researchers will need to conduct experiments and validate its performance using actual data.

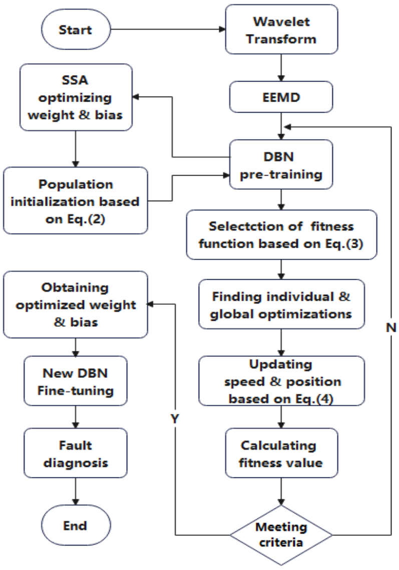

Start
Wavelet Transform
SSA optimizing weight & bias
EEMD
DBN pre-training
Population initialization based on Eq.(2)
Selectction of fitness function based on Eq.(3)
Obtaining optimized weight & bias
Finding individual & global optimizations
N
New DBN Fine-tuning
Updating speed & position based on Eq.(4)
Y
Fault diagnosis
Calculating fitness value
End
Meeting criteria

FIGURE 14. Process diagram of SSA-DBN.

Indeed, a DBN's optimal performance relies on the optimal network structure. While researchers typically set the DBN network structure based on their experience, it may not fully exploit the potential performance of DBN. To address this, a new fault detection model called SSA-DBN was proposed by optimizing DBN with the SSA. The core idea behind using SSA to optimize DBN was to determine the sparrow with the best position and the individual sparrow with the highest fitness. Throughout the iteration process, the parameters of the sparrow were used to determine the optimal network structure of DBN. Subsequently, the labelled data was selected and input into a Softmax classifier for fault classification and diagnosis, as shown in Fig.14. This optimization process aimed to obtain the optimal fault detection model. By integrating SSA optimization with DBN, the SSA-DBN model enhanced the performance and effectiveness of fault detection. It utilized the optimization capabilities of SSA to determine the optimal DBN network structure, resulting in an improved fault detection model.

D. FAULT DIAGNOSIS PROCEDURE

The detailed steps of rolling bearing fault diagnosis based on the SSA-optimized DBN are as follows:

(1)Denoising: Three-layer wavelet packet decomposition and reconstruction were applied to denoise the original vibration signal of the rolling bearing. The denoising effect was remarkable. Further, the practical signal dataset was divided into different states.

(2)Feature Extraction: Time-domain and frequency-domain features were extracted from the practical signal dataset. In addition, IMF energy features were extracted using the EEMD.

(3)Labeling: The sample labels in the entire dataset were marked manually to distinguish different types of sample data. The dataset contained 480 samples with four classifications. The dataset was split into a ratio of 3:1, with 120 samples as the test set and 360 as the training set.

(4)Data Preprocessing: The data were normalized to ensure that the values ranged between [0, 1]. This preprocessing step reduced the computational load of the model. Further, the data were transposed to adapt to the model characteristics. Before training, the maximum number of iterations and the number of sparrows were set in the SSA algorithm, which were 50 and 100. Further, the momentum parameter was set as 0.5, and the learning rate was set as 0.1. Through constant updating and iteration of sparrow positions, the optimal position sparrow with the highest fitness value was determined.

(5)DBN Training: The number of nodes in the input layer corresponded to the dimensionality of the input features, and the number of output nodes was 4, representing the running state of the four types of rolling bearings. A 2-layer RBM was set up. The RBM was trained with 65 iterations, a learning rate of 0.01, and a fine-tuning process of 10 iterations. Further, the integrated feature set was divided into a 3:1 ratio and input into the optimized SSA-DBN network model.

(6)Fault Diagnosis: The labels were set, and the rolling bearing faults were diagnosed using the trained SSA-DBN network model.

(7)Result analyzing: The combination model applied in this study was compared with other mainstream methods to verify the effectiveness of the proposed method.

This paper's diagnostic process and technical roadmap are shown in Fig.15.

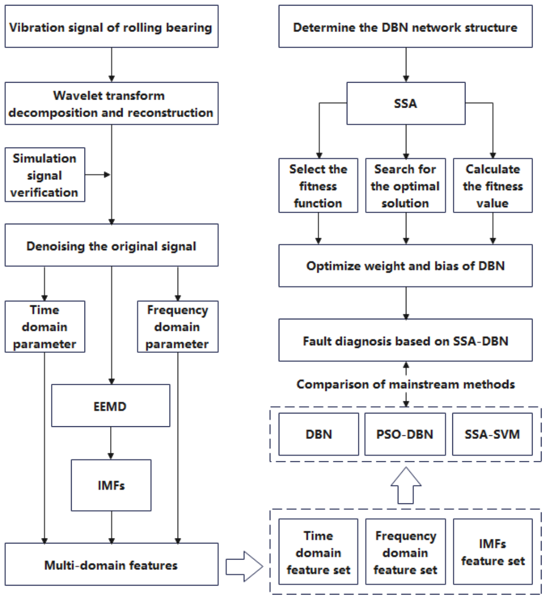

Vibration signal of rolling bearing
Determine the DBN network structure
Wavelet transform decomposition and reconstruction
SSA
Simulation signal verification
Select the fitness function
Search for the optimal solution
Calculate the fitness value
Denoising the original signal
Optimize weight and bias of DBN
Time domain parameter
Frequency domain parameter
Fault diagnosis based on SSA-DBN
Comparison of mainstream methods
EEMD
DBN
PSO-DBN
SSA-SVM
IMFs
Multi-domain features
Time domain feature set
Frequency domain feature set
IMFs feature set

FIGURE 15. Technical roadmap.

IV. FAULT DIAGNOSIS PROCESS BASED ON SSA-DBN

In this study, the data set used was the life cycle data of rolling bearings obtained from the NSF I/UCR Intelligent Maintenance System Center. The data set can be accessed at https://www.nasa.gov. The experiments were conducted using an AC motor rotating at a constant speed of 2000 RPM. Four ZA-2115 double-row roller bearings manufactured by Rexnord were installed on the rotating shaft. Further, acceleration sensors were placed on the horizontal and vertical directions of each bearing to measure and collect the corresponding vibration signals. Each data set recorded the complete life cycle of a bearing, beginning from regular operation and progressing to the point of damage. The sampling frequency for data collection was set at 20 kHz. Each interval between data points was 10 min, resulting in the collection of one sample per interval. The collection time for each sample was approximately 1.024 s, yielding 20,480 data points for each sample.

A. FEATURE SET

After preliminary analysis, the 10-dimensional time-domain feature set $T = \left[\overline{X}\ X_{rms}\ X_p\ D_x\ K_4\ K_3\ L_s\ L_p\ L_\alpha\ L_y\right]^T$ and the 3-dimensional frequency-domain feature set $P = \left[f_{FC}\ f_{MSF}\ f_{VF}\right]^T$ were obtained. In addition, the IMF component of 11 witter collection $E = \left[E_1\ E_2\ E_3 \cdots E_{11}\right]^T$, finally obtained the fusion multi-dimensional multi-domain feature set $L = \left[T\ P\ E\right]$, which contained a total of 24 dimensions. The specific parameter characteristics are presented in Tables 2 and 3.

TABLE 2. Part of time-domain feature parameters.

Fault Type	No.	1	2	3	4
		Effective Value	Degree of Skewness	Pulse Factor	Margin Factor
		X_{rms}	K_3	L_a	L_y
Normal	1	0.108	0.115	3.155	3.401
	2	0.110	0.101	2.935	3.160
Inner Race Fault	1	0.121	0.212	2.702	2.801
	2	0.126	0.011	2.162	2.251
Rollin Element Fault	1	0.119	0.051	1.574	1.602
	2	0.120	-0.009	1.720	1.750
Outer Race Fault	1	0.047	0.037	8.174	9.658
	2	0.045	-0.122	6.630	7.766

B. FAULT DIAGNOSIS

Using the vibration mentioned above signal data, 120 samples were obtained for each of the four working conditions (normal, inner race fault, rolling element fault, and outer race fault). These samples were divided into training and test sets according to the specified proportion. The training set was then used to learn and train the SSA-DBN model with the parameters mentioned earlier. The experimental results of the SSA-DBN model are shown in Fig.16. As evident, high recognition accuracy and effective classification were achieved by the DBN network model optimized by the SSA algorithm. Fig.17 shows the fitness function curve of the SSA-DBN model, indicating the decrease in the objective function value with increasing iteration times. By the second iteration, the objective function value reached its optimal value.

TABLE 3. Frequency-domain feature parameters.

Fault Type	No.	1	2	3
		Barycentric Frequency	RMS Frequency	Frequency Variance
		f_{FC}	f_{MSF}	f_{VF}
Normal	1	206.206	374838.437	332317.529
	2	191.327	346637.347	310031.259
Inner Race Fault	1	108.398	201384.266	189634.090
	2	116.481	216472.593	202904.740
Rollin Element Fault	1	60.540	112314.939	108649.877
	2	60.490	112530.993	108872.004
Outer Race Fault	1	995.530	1895052.292	903971.844
	2	997.765	1895287.995	899752.894

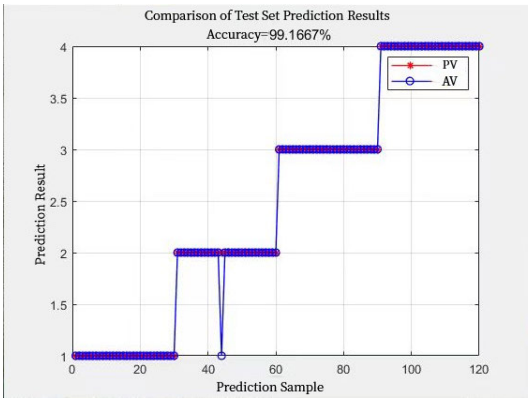


FIGURE 16. Diagnosis result graph based on multi-domain feature set SSA-DBN.

To evaluate the paper's experimental results, a visualization tool, referred to as the Confusion Matrix, was added. The Confusion Matrix is particularly suitable for supervised learning tasks as it facilitates the comparison of the accuracy of classification results. Fig.18 shows the Confusion Matrix used in this study. In the Confusion Matrix, each row represents the actual class labels corresponding to the four bearing states examined in this paper. Each column represented the predicted class labels assigned by the SSA-DBN network model. Based on the definitions of the Confusion Matrix, it can be categorized into four types: True Positive (TP),

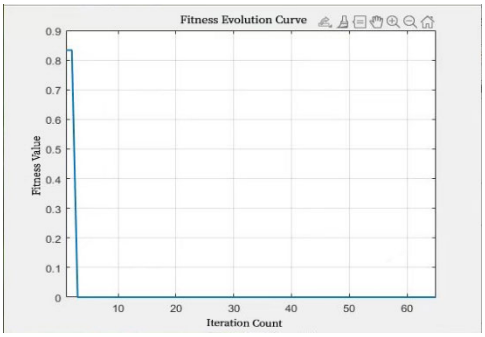


FIGURE 17. Changed curve of fitness function.

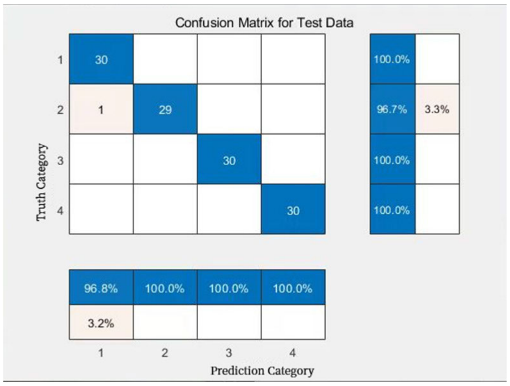


FIGURE 18. Confusion matrix visualization.

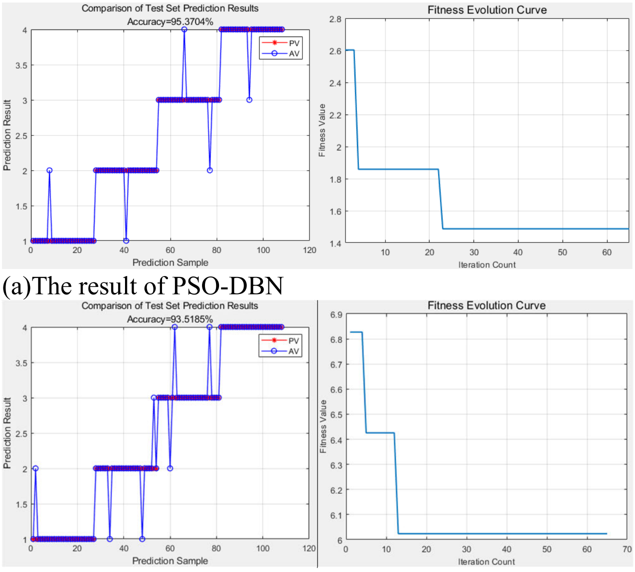
False Negative (FN), False Positive (FP), and True Negative (TN). Notably, in the Confusion Matrix, there is an occurrence of FP type. This implies that in one sample, the actual state is an inner race fault; however, the SSA-DBN network model misidentified it as a normal state, resulting in a misdiagnosis. This observation suggests further refinement in the data processing stage to enhance the overall diagnostic performance.

In contrast, when the SSA was not used to optimize the DBN network model, the diagnostic accuracy rate was only 75% for the same data division and parameters. However, after applying the SSA optimization, the fault diagnosis rate of the DBN network model improved significantly. This demonstrates SSA's effectiveness in enhancing the DBN model's performance for fault diagnosis. Furthermore, to validate the effectiveness of the selected feature set L in this study, all the characteristic parameters were input into the SSA-DBN model for fault diagnosis, and the diagnosis results were compared with mainstream methods. The parameters of different methods were set as follows:

(1)PSO-DBN: The learning rate was set as 0.02, the momentum parameter was set as 0.1, the activation function was the sigmoid function, the number of particle swarms was 15, and the number of particle swarm training iterations was 100. Other parameters were consistent with SSA-DBN.

(2)SSA-SVM: The number of iterations, sparrows, the momentum parameter, and the learning rate were consistent with SSA-DBN. The optimized penalty parameter was 60.309, and the kernel parameter was 0.5694.

The results of PSO-DBN and SSA-DBN are shown in Fig.19. The compared results and accuracy of this analysis are summarized in Table 4.



(a)The result of PSO-DBN

(b)The result of SSA-SVM

FIGURE 19. Results of mainstream methods.

TABLE 4. Diagnosis results of different feature sets.

Feature	Diagnostic Model	Test Set Accuracy	Discrepancy
Time-domain	SSA-DBN	92.67%	-6.54%
Frequency-domain		91.87%	-7.34%
IMFEF		89.44%	-9.78%
Set		99.17%	-
	DBN	75%	-24.36%
	PSO-DBN	95.37%	-3.82%
	SSA-SVM	93.52%	-5.69%

Table 4 shows that the diagnostic accuracy distribution of the SSA-DBN diagnostic model adopted in this study spans different feature sets. Specifically, the performance of the energy features derived from EEMD decomposition, represented by IMFEF, is relatively subpar. However, when temporal and spectral feature sets are combined (i.e., the combined feature set), there is a significant enhancement in the overall diagnostic accuracy, reaching 99.17%.

The percentage difference reveals that, prior to feature integration, the diagnostic precision of a single-dimensional feature set lags behind that of multi-dimensional integrated feature sets by approximately 8%. This underlines the complementary nature of features from different dimensions and domains, highlighting the superiority of multi-dimensional integrated feature sets in fault diagnosis, thereby elevating diagnostic accuracy and robustness. Additionally, Table 4 also offers a lateral comparison against various diagnostic models. The results demonstrate that the DBN model, without structural and parameter optimization, commits significant errors in its diagnosis. Conversely, the diagnostic accuracy of the DBN model, post-optimization with the SSA algorithm, has increased by about 24%, attesting to the necessity of algorithmic optimization. Furthermore, incorporating mainstream optimization algorithm PSO-DBN and mainstream classification model SSA-SVM for comparison, the analysis in conjunction with Figure 17 suggests that the same multi-dimensional integrated feature set, under different optimization algorithms or different classification models, presents variant outcomes with a disparity of around 5%. Both methods have diagnostic errors across four labels, impermissible in practical scenarios. Hence, deep learning models, optimized in structure and parameters with optimization algorithms, outshine shallow machine learning models. The degree of model optimization varies among different optimization algorithms. Empirical evidence confirms that the SSA-DBN diagnostic model employed in this study possesses genuine diagnostic capability and high precision standards.

V. CONCLUSION AND FUTURE WORK

The multi-domain feature set, consisting of time-domain features, frequency-domain features, and IMF energy features, achieved a high accuracy of 99.17% in diagnosing the three fault states (inner ring fault, rolling element fault, outer ring fault) as well as the normal state of rolling bearings. Utilizing this feature set facilitated the efficient diagnosis process in the network model. Furthermore, by comparing different features' impact on the diagnosis results, it was evident that this study's selected feature data set was highly effective. The average diagnosis accuracy rate of 99.08% was obtained after conducting ten experiments, highlighting the robustness of the feature dataset. Compared to the non-optimized diagnosis model, even including the mainstream models, such as PSO-DBN and SSA-SVM, the proposed SSA-DBN model outperformed in performance. This emphasizes the significance of the feature dataset and the diagnosis model, as each domain's feature parameters contribute to diagnostic characteristics and rely on each other. The SSA-DBN model demonstrated excellent fault identification and diagnosis stability, ultimately enhancing overall diagnostic accuracy.

In terms of future work, there are several areas for further improvement and research based on the method proposed in this paper. These aspects can contribute to deploying the proposed method in real-world engineering scenarios and fundamentally contribute to the field. In addition, these suggestions can serve as references for other scholars conducting further research:

(1)Multi-Dimensional Fault Diagnosis: The integration of multiple types of signals, such as vibration, current, temperature, and sound from rolling bearings, can be explored. Various types of sensors can be used for signal acquisition, study sensor layout positions and quantities, and develop suitable algorithms to extract features from different signal types. This research can contribute to establishing a multi-angle, multi-functional system for intelligent fault diagnosis of rolling bearings under different working conditions.

(2)Addressing Sample Scarcity: In practical applications, collecting an adequate number of fault samples can be challenging, resulting in imbalanced datasets. Although deep learning fault diagnosis models can handle multiple types of faults and perform identification and classification tasks, addressing the classification error issue in the presence of imbalanced samples becomes crucial. Further research can focus on developing strategies to mitigate the effects of imbalanced datasets and improve classification performance.

These research directions can extend the current work and contribute to the advancement of fault diagnosis in rolling bearings. Thus, by addressing issues related to multi-dimensional signal analysis and handling imbalanced datasets, further improvements can be made to enhance the practical applicability and effectiveness of fault diagnosis methods.

ACKNOWLEDGMENT

The authors would like to thank the editors and the reviewers for their helpful suggestions, which have greatly improved this article.

REFERENCES

[1] G. Wu, T. Yan, G. Yang, H. Chai, and C. Cao, "A review on rolling bearing fault signal detection methods based on different sensors," *Sensors*, vol. 22, no. 21, p. 8330, Oct. 2022.

[2] X. Zhang, B. Zhao, and Y. Lin, "Machine learning based bearing fault diagnosis using the case western reserve university data: A review," *IEEE Access*, vol. 9, pp. 155598–155608, 2021.

[3] X. Chen, R. Yang, Y. Xue, M. Huang, R. Ferrero, and Z. Wang, "Deep transfer learning for bearing fault diagnosis: A systematic review since 2016," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–21, 2023.

[4] M. Maurya, I. Panigrahi, D. Dash, and C. Malla, "Intelligent fault diagnostic system for rotating machinery based on IoT with cloud computing and artificial intelligence techniques: A review," *Soft Comput.*, vol. 28, no. 1, pp. 477–494, Jan. 2024.

[5] J. Tao, Y. Liu, and D. Yang, "Bearing fault diagnosis based on deep belief network and multisensor information fusion," *Shock Vibrat.*, vol. 2016, pp. 1–9, Aug. 2016.

[6] L. Xia, L. Shang, L. Fan, D. Wang, Z. Xing, and J. Li, "Feature extraction of rolling bearing fault signal of: Rolling mill based on wavelet packet denoising method," *Proc. SPIE*, vol. 12079, pp. 651–656, Dec. 2021.

[7] X. X. Qi, J. W. Ji, and X. W. Han, "Fault diagnosis methods of rolling bearing: A general review," *Key Eng. Mater.*, vols. 480–481, pp. 986–992, Jun. 2011.

[8] B. Jin, T. Yang, M. Tian, and S. Xiong, "Diagnosis for rolling bearing nonlinear fault based on Daubechies wavelet," in *Proc. 7th Int. Symp. Test Meas. (ISTM), Conf.*, vols. 1–7, 2007, pp. 3989–3991.

[9] S. Jing, J. Yuan, X. Li, J. Leng, and Ieee, ''Weak fault feature identification for rolling bearing based on EMD and spectral kurtosis method,'' in *Proc. Int. Conf. Inf. Syst. Comput. Aided Educ. (ICISCAE)*, 2018, pp. 235–239.

[10] R. Wang, Z. Zhang, Z. Xia, J. Miao, and Y. Guo, ''A new approach for rolling bearing fault diagnosis based on EEMD hierarchical entropy and improved CS-SVM,'' in *Proc. Prognostics Syst. Health Manage. Conf. (PHM-Qingdao)*, Oct. 2019, pp. 1–6.

[11] Y. Bu, J. Wu, J. Ma, X. Wang, and Y. Fan, ''The rolling bearing fault diagnosis based on LMD and LS-SVM,'' in *Proc. 26th Chin. Control Decis. Conf. (CCDC)*, May 2014, pp. 3797–3801.

[12] J. Zhang and X. Huang, ''A fault diagnosis approach for rolling bearings based on EMD method and eigenvector algorithm,'' in *Proc. Int. Conf. Intell. Comput.*, 2008, pp. 294–301.

[13] A. Kumar, C. P. Gandhi, G. Vashishtha, P. Kundu, H. Tang, A. Glowacz, R. K. Shukla, and J. Xiang, ''VMD based trigonometric entropy measure: A simple and effective tool for dynamic degradation monitoring of rolling element bearing,'' *Meas. Sci. Technol.*, vol. 33, no. 1, Jan. 2022, Art. no. 014005.

[14] Q. Ni, J. C. Ji, K. Feng, and B. Halkon, ''A fault information-guided variational mode decomposition (FIVMD) method for rolling element bearings diagnosis,'' *Mech. Syst. Signal Process.*, vol. 164, Feb. 2022, Art. no. 108216.

[15] L. Wang and Y. Shao, ''Fault feature extraction of rotating machinery using a reweighted complete ensemble empirical mode decomposition with adaptive noise and demodulation analysis,'' *Mech. Syst. Signal Process.*, vol. 138, Apr. 2020, Art. no. 106545.

[16] S. Kang, W. Chen, Y. Wang, X. Na, Q. Wang, and V. I. Mikulovich, ''Method of state identification of rolling bearings based on deep domain adaptation under varying loads,'' *IET Sci., Meas. Technol.*, vol. 14, no. 3, pp. 303–313, May 2020.

[17] F. Xu and P. W. Tse, ''Combined deep belief network in deep learning with affinity propagation clustering algorithm for roller bearings fault diagnosis without data label,'' *J. Vibrat. Control*, vol. 25, no. 2, pp. 473–482, Jan. 2019.

[18] C. Zhong, J.-S. Wang, and W.-Z. Sun, ''Fault diagnosis method of rotating bearing based on improved ensemble empirical mode decomposition and deep belief network,'' *Meas. Sci. Technol.*, vol. 33, no. 8, Aug. 2022, Art. no. 085109.

[19] J. Zhu, T. Hu, B. Jiang, and X. Yang, ''Intelligent bearing fault diagnosis using PCA–DBN framework,'' *Neural Comput. Appl.*, vol. 32, no. 14, pp. 10773–10781, Jul. 2020.

[20] S. Zhiwu, L. Xia, L. Wanxiang, G. Maosheng, and Y. Yan, ''A rolling bearing fault diagnosis method based on fastDTW and an AGBDBN,'' *Insight Non-Destructive Test. Condition Monitor.*, vol. 62, no. 8, pp. 457–463, Aug. 2020.

[21] W. Deng, H. Liu, J. Xu, H. Zhao, and Y. Song, ''An improved quantum-inspired differential evolution algorithm for deep belief network,'' *IEEE Trans. Instrum. Meas.*, vol. 69, no. 10, pp. 7319–7327, Oct. 2020.

[22] S. Gao, L. Xu, Y. Zhang, and Z. Pei, ''Rolling bearing fault diagnosis based on SSA optimized self-adaptive DBN,'' *ISA Trans.*, vol. 128, pp. 485–502, Sep. 2022.

[23] Y. Lei, J. Lin, Z. He, and Y. Zi, ''Application of an improved Kurtogram method for fault diagnosis of rolling element bearings,'' *Mech. Syst. Signal Process.*, vol. 25, no. 5, pp. 1738–1749, Jul. 2011.

[24] H. Cheng, S. Yu, and L. Cheng, ''Application of wavelet transform in fault diagnosis of rolling bearing,'' in *Proc. 10th Int. Conf. Natural Comput. (ICNC)*, Aug. 2014, pp. 1066–1070.

[25] S. Samsingh, ''Medical image denoising using wavelet transform and image fusion,'' *Proc. Comput.*, vol. 115, pp. 131–138, May 2017.

[26] K. Fang, H. Zhang, H. Qi, and Y. Dai, ''Comparison of EMD and EEMD in rolling bearing fault signal analysis,'' in *Proc. IEEE Int. Instrum. Meas. Technol. Conf. (I2MTC)*, May 2018, pp. 1–5.

[27] A. Fathy, T. M. Alanazi, H. Rezk, and D. Yousri, ''Optimal energy management of micro-grid using sparrow search algorithm,'' *Energy Rep.*, vol. 8, pp. 758–773, Nov. 2022.

[28] X. Zhang, F. Xiao, X. Tong, J. Yun, Y. Liu, Y. Sun, B. Tao, J. Kong, M. Xu, and B. Chen, ''Time optimal trajectory planing based on improved sparrow search algorithm,'' *Frontiers Bioeng. Biotechnol.*, vol. 10, Mar. 2022, Art. no. 852408.

[29] J. He, F. Gao, J. Wang, Q. Wu, Q. Zhang, and W. Lin, ''A method combining multi-feature fusion and optimized deep belief network for EMG-based human gait classification,'' *Mathematics*, vol. 10, no. 22, p. 4387, Nov. 2022.

[30] N. Jia, Y. Cheng, Y. Tian, and F. Yang, ''Intelligent fault severity detection of rotating machines based on VMD-WVD and parameter-optimized DBN,'' *Shock Vibrat.*, vol. 2022, pp. 1–15, Mar. 2022.

[31] J. Li, W. Wang, G. Chen, and Z. Han, ''Spatiotemporal assessment of landslide susceptibility in Southern Sichuan, China using SA-DBN, PSO-DBN and SSA-DBN models compared with DBN model,'' *Adv. Space Res.*, vol. 69, no. 8, pp. 3071–3087, Apr. 2022.

DONGHAO XU was born in Zhejiang, China, in 1997. He is currently pursuing the master's degree with the College of Air Transportation, Shanghai University of Engineering Science, China. His main research interest includes intelligent fault diagnosis.

CHENG LI received the Ph.D. degree in management science and engineering. He is currently a Professor with the College of Air Transportation, Shanghai University of Engineering Science, China. He has published articles in various international journals, such as IEEE Access, *Journal of Coastal Research*, and *Sustainability*. He has successfully managed various national and local sponsored research projects and grants. His research interests include management science, including demand forecast and transportation management.