Hindawi
Shock and Vibration
Volume 2021, Article ID 8899188, 17 pages
https://doi.org/10.1155/2021/8899188

Research Article

Fault Diagnosis Approach for Rotating Machinery Based on Feature Importance Ranking and Selection

Zong Yuan,[1,2] **Taotao Zhou**,[2] **Jie Liu**,[3,4] **Changhe Zhang**,[5] **and Yong Liu**[2]

[1]*School of Transportation, Wuhan University of Technology, Wuhan 430070, China*
[2]*China Ship Development and Design Center, Wuhan 430063, China*
[3]*School of Civil and Hydraulic Engineering, Huazhong University of Science and Technology, Wuhan 430074, China*
[4]*Nondestructive Detection and Monitoring Technology for High Speed Transportation Facilities, Key Laboratory of Ministry of Industry and Information Technology, Nanjing 210016, China*
[5]*School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China*

Correspondence should be addressed to Changhe Zhang; changhe_zhang@hust.edu.cn

Received 29 August 2020; Revised 16 January 2021; Accepted 21 January 2021; Published 4 February 2021

Academic Editor: Ahmed Ibrahim

The key to fault diagnosis of rotating machinery is to extract fault features effectively and select the appropriate classification algorithm. As a common signal decomposition method, the effect of wavelet packet decomposition (WPD) largely depends on the applicability of the wavelet basis function (WBF). In this paper, a novel fault diagnosis approach for rotating machinery based on feature importance ranking and selection is proposed. Firstly, a two-step principle is proposed to select the most suitable WBF for the vibration signal, based on which an optimized WPD (OWPD) method is proposed to decompose the vibration signal and extract the fault information in the frequency domain. Secondly, FE is utilized to extract fault features of the decomposed subsignals of OWPD. Thirdly, the categorical boosting (CatBoost) algorithm is introduced to rank the fault features by a certain strategy, and the optimal feature set is further utilized to identify and diagnose the fault types. A hybrid dataset of bearing and rotor faults and an actual dataset of the one-stage reduction gearbox are utilized for experimental verification. Experimental results indicate that the proposed approach can achieve higher fault diagnosis accuracy using fewer features under complex working conditions.

1. Introduction

At present, the components of industrial rotating machinery equipment are becoming increasingly complex and compact. As a key component of the transmission system, rotating machinery is an important part of modern industrial machinery equipment, including motors, engines, bearings, and gearboxes [1–3]. When rotating machinery is operating under harsh or complex conditions, its key components are extremely prone to failure, which may cause the shutdown of the entire mechanical equipment, and even endanger the safety of surrounding operators [4, 5]. Therefore, it is significant to construct a fault diagnosis scheme for rotating machinery under complex conditions to accurately detect and diagnose its health or fault state.

Fault diagnosis based on vibration signal analysis is the main research hotspot at present, among which the most critical step is feature extraction [6, 7]. Based on vibration signal analysis, the existing methods mainly extract fault features from the time domain, frequency domain, and time-frequency domain [8, 9]. Time-domain features contain root mean square, mean, standard deviation, kurtosis, etc., which may be valid only for certain fault types [10, 11]. The frequency-domain analysis is mostly based on the Fourier transform (FT) [12]. However, these methods are limited by prior knowledge and experience in practical applications due to the nonlinearity and nonstationarity of the original vibration signal, which makes it difficult for them to effectively mine the fault information hidden in the vibration signal [11, 13].

As a measure of time-domain uncertainty, entropy-based analysis methods have attracted extensive attention of scholars, which have been widely used in image processing, biological analysis, and other fields [14]. These entropies mainly include sample entropy (SE) [15], approximate entropy (AE) [16], permutation entropy (PE) [17], fuzzy entropy (FE) [18], and symbolic dynamic entropy (SDE) [19]. FE uses a Gaussian function to replace the Heaviside function in SE to measure the similarity between two vectors. In recent studies, these entropies are usually combined with some other processing strategies, e.g., multiscale and improved multiscale techniques, and some signal decomposition methods in the time-frequency domain [19–21]. However, the multiscale analysis may result in the absence of part of the frequency band components, which may lead to the loss of some fault information. In addition, traditional signal decomposition methods also have some disadvantages that cannot be ignored.

Research studies based on time-frequency domain analysis have been done for a long time. For example, empirical mode decomposition (EMD) and local mean decomposition (LMD) both are self-adaptive time-frequency decomposition methods [22, 23]. LMD is improved on the basis of EMD to better maintain the local characteristics of the original signal [24]. Variational mode decomposition (VMD) proposed by Dragomiretskiy et al. is a self-adaptive decomposition method that aims to overcome the shortcomings of undershoot, overshoot, mode mixing in EMD [25–27]. However, there are still some defects in VMD, e.g., the massive consumption of computing resources [27, 28]. In addition, these methods are all based on "mode," and some frequency information will be lost, which means that they do not apply to frequency analysis.

Compared with the above methods, the decomposition strategy of wavelet packet decomposition (WPD) is to pass the signal through a series of filters with different central frequencies but the same bandwidth. Therefore, the signal analysis performed by WPD is more refined, especially for the high-frequency components [29]. In [30], an automatic method combining WPD and EMD was proposed to detect the weak defects of rolling bearings. In [31], WPD was combined with PE to extract fault features of rolling bearings. However, how to select the optimal wavelet basis function (WBF) has not been analysed and discussed in these references. The WBF is a group of functions obtained from the expansion and translation, including db wavelets, sym wavelets, and mexh wavelets [32]. Different WBFs are applicable to different analysis objects, and improper selection may affect the accuracy of fault pattern recognition [32, 33]. In this paper, a two-step principle is proposed to select the most suitable WBF for the fault vibration signal. Then, the optimized WPD (OWPD) method is proposed and applied to decompose the vibration signal to obtain its frequency component. In view of the advantage that entropy measure can effectively extract dynamic information of time series, FE is used to extract hidden fault features from decomposed subsignals. Meanwhile, it also has the advantage of being insensitive to background noise and good robustness [18, 20]. Therefore, a novel fault feature exaction method combining OWPD and FE is further proposed in this paper.

After feature extraction utilizing OWPD and FE, a classification algorithm with good performance and computational efficiency is needed to give final diagnosis results. In addition, screening redundant features before fault classification can effectively reduce feature dimension and computational burden and further improve classification accuracy [34]. Traditional classification algorithms include support vector machine (SVM) [35], *K*-nearest neighbor (KNN) [36], artificial neural networks (ANNs) [37], and random forest (RF) [38]. Deep learning (DL) algorithms include convolutional neural network (CNN) [39], autoencoders (AEs) [40], and deep belief network (DBN) [11, 41]. However, these classification algorithms still have some inevitable shortcomings. For example, SVM is not effective for large-scale training samples and sensitive to the selection of parameters and kernel function. RF is easy to overfit in noisy classification or regression problems. The structure and parameters of some DL algorithms, e.g., DBN, are basically determined by human experience, which not only affects the accuracy of diagnostic results but also causes a large amount of computing costs [11, 42]. CatBoost is a new implementation of the gradient boosting decision tree (GBDT) framework [43]. It has the advantages of high efficiency, few parameters, and strong generalization ability and has excellent performance in many machine learning tasks [43–45]. In addition, as an algorithm based on the decision tree, it can obtain the importance of each feature according to the tree model after gradient boosting, and then the valuable features can be effectively selected for model training. Therefore, it is introduced for feature selection to form a feature set that contains the main fault information. To the best of authors' knowledge, CatBoost algorithm is rarely studied in the field of fault diagnosis of rotating machinery. In this paper, it is introduced not only for fault pattern recognition but also for selecting the optimal features.

Finally, the optimization of hyperparameters is also an urgent problem to be solved in the use of the CatBoost algorithm, which usually has a great impact on the performance of the model. The optimization of hyperparameters is to find an acceptable solution for the optimization goal as effectively as possible [46]. Due to the large amount of data and large solution space, the application of traditional solution methods, e.g., grid search and greedy algorithm, has been limited, while intelligent algorithms such as differential evolution have been widely used due to their fast computing efficiency and the ability to obtain global optimal solution [46, 47]. In this paper, Bayesian optimization (BO) algorithm [48] is considered to solve this problem to find the optimal hyperparameters of the CatBoost classifier. It can obtain the global optimal solution through Gaussian process, which has the advantages of high search efficiency and less iteration times, and can be used for the optimization of any black-box function.

Based on the above analysis, aiming to solve the defect that traditional feature extraction methods cannot fully explore the deep-level fault features and to improve the performance of fault pattern recognition, a novel fault diagnosis approach based on feature importance ranking and selection is proposed. In summary, the advantages of WPD in signal decomposition, FE in feature extraction, and CatBoost in fault pattern recognition are fully exploited in the proposed approach. The main contributions can be summarized as follows:

(1) A two-step principle is proposed to select the optimal WBF adaptively according to the characteristics of the mechanical vibration signal.

(2) A fault feature extraction method combining OWPD and FE is proposed, where OWPD is utilized to decompose the vibration signal, and FE is further adopted to form the fault feature set.

(3) CatBoost algorithm is introduced not only for fault pattern recognition but also for feature selection to filter redundant features, which helps to reduce model training time and improve the classification accuracy.

(4) BO algorithm is adopted to solve the optimization problem of hyperparameters in CatBoost. On this basis, the BO-CatBoost algorithm is established and applied to the fault diagnosis of rotating machinery.

The remainder of this paper is organized as follows. Section 2 introduces the theoretical knowledge and methods of the proposed approach. The diagnosis process and the preliminary validation of the proposed fault diagnosis approach with a mechanical fault simulation (MFS) platform dataset are detailed in Section 3. Further experimental verification using another actual dataset of the one-stage reduction gearbox is shown in Section 4. Section 5 contains the conclusions and future research studies.

2. Materials and Methods

2.1. Optimized Wavelet Packet Decomposition

2.1.1. Wavelet Packet Decomposition. Generally, FT has been widely used in traditional vibration signal analysis [49]. However, only the frequency-domain information is retained in FT, while the time-domain information is completely lost, which makes it unsuitable for the analysis of nonstationary time-varying signals. Wavelet transform (WT) can provide information in both frequency and time domains to overcome the deficiency of FT [50]. However, only the low-frequency coefficients will be decomposed again in the WT method, which will cause the problem of missing high-frequency information. WPD was proposed to address this deficiency, where the information in both low-frequency band and high-frequency band is completely preserved [51]. The schematic diagram of a three-layer WPD is shown in Figure 1, and the theory is described as follows.

Let j denote the decomposition layer, n denote the frequency factor $(n = 0, 1, 2, \ldots, 2^j - 1)$, and $\psi(n)$ and $\phi(n)$ represent the wavelet function and scale function, respectively. Given $\varphi_0(n) = \phi(n)$ and $\varphi_1(n) = \psi(n)$, the wavelet packet $\varphi_i(n)$ $(i = 0,\ 1,\ 2, \ldots)$ can be defined as

$$\begin{cases} \varphi_{2i}(n) = \sqrt{2} \sum\limits_{k \in Z} h_k \varphi_i(2n - k), \\ \varphi_{2i+1}(n) = \sqrt{2} \sum\limits_{k \in Z} g_k \varphi_i(2n - k), \end{cases} \tag{1}$$

where k is the shift factor, Z is the integer set, h_k denotes the low-pass filter, g_k denotes the high-pass filter, and h_k, g_k is a couple of quadruple mirror filters that satisfies $g_k = (-1)^k h_k$.

For a given time series $x(n)$, let $x_j^p(n)$ $(p = 0, 1, 2, \ldots, 2^j - 1)$ denote its subsignal, which can be represented as a linear combination of the corresponding wavelet function of the wavelet packet:

$$x_j^p(n) = \sum_{k \in Z} d_j^p(k) \varphi_{2_{j-1+p}}(2n - k), \tag{2}$$

where $d_j^p(k)$ denotes the p-th wavelet packet coefficient of the j-th layer, and it can be obtained by the inner product between $x(n)$ and $\varphi_{2_{j-1+p}}(2n - k)$, namely,

$$d_j^p(k) \leq x(n),\ \varphi_{2_{j-1+p}}(2n - k) \geq \int_{-\infty}^{+\infty} x(n) \varphi_{2_{j-1+p}}(2n - k) \mathrm{d}n. \tag{3}$$

An approximation $x_j(n)$ of $x(n)$ with layer j equates the sum of all the subsignals:

$$x_j(n) = \sum_p x_j^p(n). \tag{4}$$

2.1.2. Optimized Wavelet Packet Decomposition. The main idea of the OWPD method is to automatically select the most suitable WBF for vibration signal analysis of rotating machinery according to the proposed two-step principle on the basis of WPD. In general, the main characteristics to be considered in selecting the WBF include orthogonality, compactness, symmetry, and vanishing moment. Considering the characteristics of different wavelet families and vibration signals of rotating machinery, the coif wavelets, db wavelets, and sym wavelets are selected as the candidate WBFs. Figure 2 shows the proposed two-step principle for selecting the optimal WBF, detailed as follows.

Step 1: select the candidate WBFs preliminarily from the same wavelet family according to the principle of maximum energy-to-Shannon entropy ratio (METSE) [52].

(1) Calculate the energy value $E(n)$ of the n-th node:

$$E(n) = \sum_{i=1}^{m} \left| C_{n,i} \right|^2, \tag{5}$$

where i and m are the serial number and total number of discrete points in the n-th node, respectively, and $C_{n,i}$ is the coefficient corresponding to the discrete point.

(2) The Shannon entropy of the n-th node is defined as

$$S_{\text{entropy}}(n) = -\sum_{i=1}^{m} p_i \log_2 p_i, \tag{6}$$

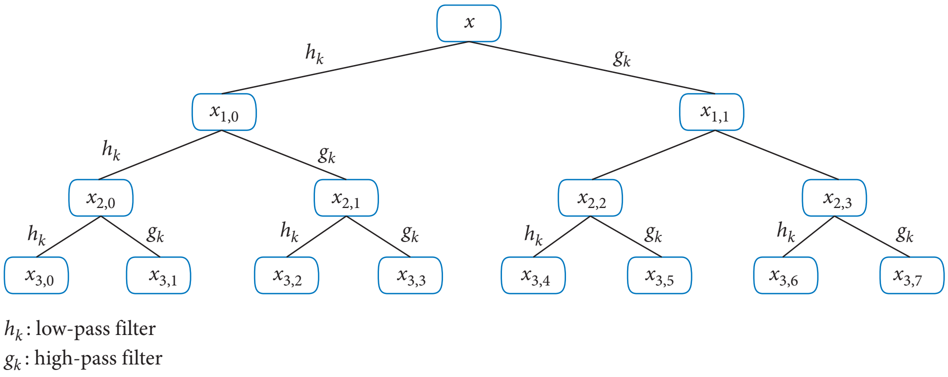


Figure 1: Schematic diagram of the three-layer wavelet packet decomposition.

where p_i is the energy probability distribution of the wavelet coefficients, defined as

$$p_i = \frac{|C_{n,i}|^2}{E(n)}. \quad (7)$$

(3) The ratio of the total energy and the total Shannon entropy of the j-layer WPD is represented as ζ, namely,

$$\zeta = \frac{\sum_{n=1}^{2^j} E(n)}{\sum_{n=1}^{2^j} S_{\text{entropy}}(n)}. \quad (8)$$

According to equations (5)–(8), the candidate WBF with the largest ζ value can be selected from the same wavelet family.

Step 2: select the optimal WBF further from different wavelet families according to the principle of similarity measure.

Firstly, the candidate WBFs from different wavelet families mentioned above are applied to implement WPD, respectively. Then, the signal is reconstructed using the coefficients with the nodes of the last layer. Finally, a standardized Euclidean distance is used to measure the similarity between the original signal x_i and reconstructed signal y_i $(i = 1, 2, \ldots, N)$:

$$d = \sqrt{\sum_{i=1}^{N} \left(\frac{x_i - y_i}{s_i} \right)^2}, \quad (9)$$

where s_i is the standard deviation between x_i and y_i. The smaller the value of d is, the closer the reconstructed signal is to the original signal, and the corresponding WBF is more suitable for signal analysis.

2.2. Fuzzy Entropy.

2.2. Fuzzy Entropy. Given an N-dimensional time series $[\mu(1), \mu(2), \ldots, \mu(N)]$, the phase space dimension and the similarity tolerance are defined as m $(m \leqq N-2)$ and r, respectively. Then, the phase space can be reconstructed as

$$X(i) = [\mu(i), \mu(i+1), \ldots, \mu(i+m-1)] - \mu_0(i), \quad i = 1, 2, \ldots, N-m+1, \quad (10)$$

where

$$\mu_0(i) = \frac{1}{m} \sum_{j=0}^{m-1} \mu(i+j). \quad (11)$$

The fuzzy membership function $A(x)$ is introduced as

$$A(x) = \begin{cases} 1, & x = 0, \\ \exp\left[-\ln(2) \left(\frac{x}{r} \right)^2 \right], & x > 0. \end{cases} \quad (12)$$

For $i = 1, 2, \ldots, N-m+1$, calculate

$$A_{ij}^m = \exp\left[-\ln(2) \cdot \left(\frac{d_{ij}^m}{r} \right)^2 \right], \quad j = 1, 2, \ldots, N-m+1, \ j \neq i, \quad (13)$$

where d_{ij}^m is the maximum absolute distance between window vector $X(i)$ and $X(j)$, that is,

$$d_{ij}^m = d[X(i), X(j)] = \max_{p=1,2,\ldots,m} \left(\left| \mu(i+p-1) - \mu_0(i) \right| - \left| \mu(j+p-1) - \mu_0(j) \right| \right). \quad (14)$$

The function Φ^m is defined as

$$\Phi^m(r) = \frac{1}{N-m+1} \sum_{i=1}^{N-m+1} \frac{1}{N-m} \sum_{j=1, j \neq i}^{N-m+1} A_{ij}^m. \quad (15)$$

Therefore, the FE value of the original time series can be calculated as

$$\mathrm{FE}(m, n, r, N) = \ln \Phi^m(n, r) - \ln \Phi^{m+1}(n, r). \quad (16)$$

2.3. Categorical Boosting

2.3.1. CatBoost for Classification. In the GBDT algorithm, lots of decision trees are combined to produce a model of high accuracy, and the progress can be written as

$$y(x) = \sum_{t=1}^{T} f_t(x, \theta_t), \quad (17)$$

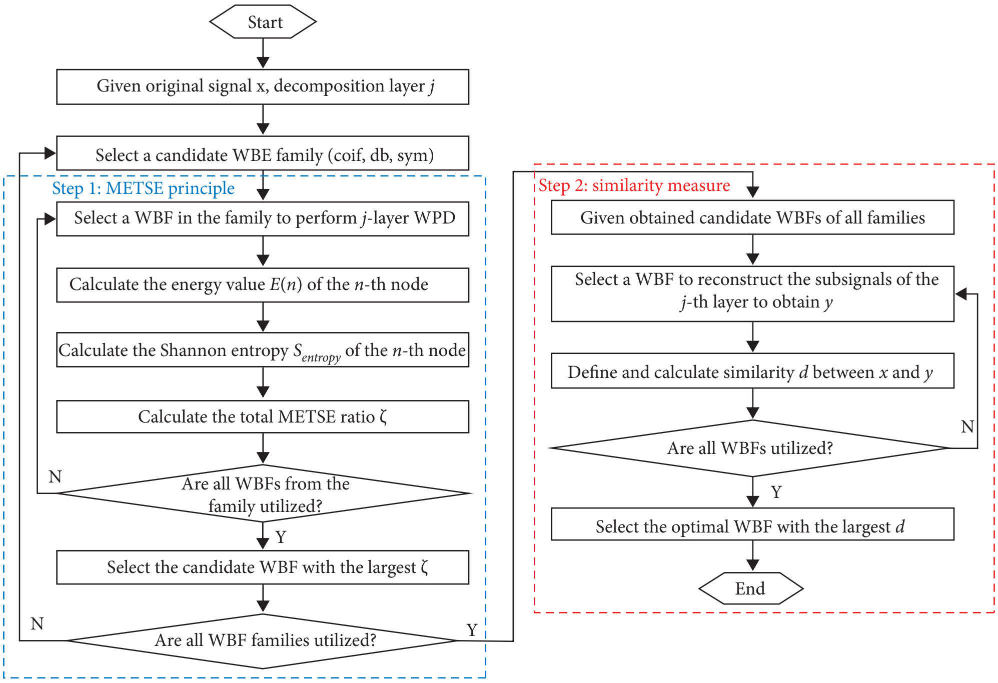

Start
Given original signal x, decomposition layer j
Select a candidate WBE family (coif, db, sym)
Step 1: METSE principle
Select a WBF in the family to perform j-layer WPD
Calculate the energy value $E(n)$ of the n-th node
Calculate the Shannon entropy $S_{entropy}$ of the n-th node
Calculate the total METSE ratio ζ
Are all WBFs from the family utilized?
N
Y
Select the candidate WBF with the largest ζ
Are all WBF families utilized?
N
Y
Step 2: similarity measure
Given obtained candidate WBFs of all families
Select a WBF to reconstruct the subsignals of the j-th layer to obtain y
Define and calculate similarity d between x and y
Are all WBFs utilized?
N
Y
Select the optimal WBF with the largest d
End

Figure 2: The proposed two-step principle of selecting the optimal WBF.

where x is the feature vector, T is the number of trees, θ_t $(t = 1, 2, \ldots, T)$ is a learned parameter, and f_t (x, θ_t) represents the decision trees that are learned.

Given training samples $D = \{(x_k, y_k)\}_1^n$, where n is the number of samples, x_k $(k = 1, 2, \ldots, n)$ is the sample data, and y_k is the true sample label. To learn the model introduced in equation (17), the following objective function needs to be minimized:

$$O(f_t) = \sum_{i=1}^{n} L(y_k, \hat{y}_k) + \sum_{t=1}^{T} \Omega(f_t), \qquad (18)$$

where $\hat{y}_k$ is the predicted sample label, L is the loss function that represents the difference between y_k and $\hat{y}_k$, and Ω is the regular function that is used to punish the complexity of f_t, defined as

$$\Omega = \alpha q + \frac{1}{2}\beta\|\omega\|^2, \qquad (19)$$

where α is the penalty parameter that controls the number of leaf nodes, q is the number of leaf nodes, β is the regularization parameter, and ω is the weight coefficient.

Let g denote the negative gradient of the loss function, and the objective function is minimized in the direction of g, namely,

$$g = -\left[\frac{\partial L(y_k, \hat{y}_k)}{\partial \hat{y}_k}\right]. \qquad (20)$$

Traditional GBDT algorithms generally have the problem of prediction offset, which affects the generalization ability of the model. To overcome this defect, CatBoost was proposed with two notable improvements [43]: (1) the ordered boosting strategy was adopted to obtain the unbiased estimation of the gradient and slow down the prediction offset; (2) the oblivious tree was used as the basic learner to increase the reliability of the model and speed up the prediction. In addition, to better deal with categorical features, the greedy target-based statistics strategy was improved by adding prior terms in CatBoost algorithm, which can be summarized as three main steps: (1) all the sample datasets are randomly arranged; (2) samples with the same category are selected, and the average label of similar samples is calculated; and (3) features of each sample are digitized by adding the prior term and its corresponding weight coefficient. The improved greedy target-based statistics strategy can be expressed as

$$\hat{x}_k^i = \frac{\sum_{j=1}^{n}\left\{x_j^i = x_k^i\right\} \cdot y_i + a \cdot P}{\sum_{j=1}^{n}\left\{x_j^i = x_k^i\right\} + a}, \qquad (21)$$

where x_k^i represents the i-th category feature of the k-th sample, $\hat{x}_k^i$ represents the corresponding numerical feature, P represents the increased prior value, and a represents the weight coefficient $(a > 0)$. The addition of prior values can effectively reduce the noise caused by low-frequency features and avoid the overfitting phenomenon.

2.3.2. CatBoost for Feature Selection. The growth strategies of decision trees are different in different GBDT algorithms. XGBoost uses level-wise strategy, which has the disadvantage of inefficiency [53]. CatBoost uses the symmetric tree strategy to optimize the computation of the leaf value to prevent the model from overfitting. In the case of the basic learner of CatBoost is the tree model, the feature coefficient or importance can be obtained according to a certain evaluation index after training the model, e.g., the change of loss function or prediction values.

For a given feature set $F = \{f_1, f_2, \ldots, f_N\}$, the feature importance of f_i $(i = 1, 2, \ldots N)$ in the trained CatBoost model is calculated by

$$\text{feature}_{f_i} = \sum_{\text{trees,leaf } S_{f_i}} (\nu_1 - avr)^2 \cdot c_1 + (\nu_2 - avr)^2 \cdot c_2, \tag{22}$$

and

$$avr = \frac{\nu_1 \cdot c_1 + \nu_2 \cdot c_2}{c_1 + c_2}, \tag{23}$$

where S denotes the different paths to the leaf nodes in the decision tree, c_1 and c_2 denote the total weight coefficient in the left and right leaves, respectively, and v_1 and v_2 denote the formula value in the left and right leaves, respectively.

2.3.3. BO Algorithm for Optimizing Hyperparameters. In machine learning algorithms, the hyperparameters usually have a strong influence on the performance of the model. There are some inevitable shortcomings in traditional methods of parameter tuning. For example, greedy algorithm can only obtain the local optimal solution, and the uncertainty and nonconvexity of grid search tend to miss global optimality. Different with these methods, BO algorithm can obtain the global optimal solution through Gaussian process, which is considered to find the optimal hyperparameters for the CatBoost model.

The basic thought of BO algorithm is that, for the given data and optimal termination condition (usually, the number of iterations or the expected value of the objective function), Bayesian theory is used to estimate the posterior distribution. The distribution and the information from the previous sampling point are used to select the hyperparameters of the later sampling until that the value of the objective function reaches the maximum globally. Here, the objective function is defined as the maximum of the classification accuracy:

$$OF(\mu) = \arg\max\left(\frac{1}{K}\sum_{k=1}^{K} f(CB, \mu, D_t, D_v)\right), \tag{24}$$

where CB denotes the CatBoost classifier, $\mu = \{\mu_1, \mu_2, \ldots, \mu_n\}$ is the hyperparameters, D_t and D_v represent the training and validation set divided by the K-fold cross-validation, respectively, and $f(\text{CB}, \mu, D_t, D_v)$ is the classification accuracy.

2.3.4. The Proposed Fault Diagnosis Approach. An overview of the proposed fault diagnosis approach for rotating machinery based on feature importance ranking and selection is shown in Figure 3. The specific steps are as follows:

Step 1: the original vibration signals are acquired by accelerometers and the data acquisition system.

Step 2: the optimal WBF is selected according to the proposed two-step principle, and then the original vibration signals are decomposed by OWPD.

Step 3: the FE values of the decomposed subsignals are calculated to form the fault feature set F.

Step 4: CatBoost algorithm is utilized to obtain the importance of each feature in F by a certain strategy. According to the ranking result of feature importance, the candidate features are selected in sequence and combined with the corresponding labels to form dataset S.

Step 5: dataset S is divided into two parts according to a certain proportion: training set and test set.

Step 6: the training set is used to train the CatBoost classifier, and BO algorithm is adopted to optimize the main hyperparameters.

Step 7: the test set is fed into the trained CatBoost classifier to output the diagnostic results.

3. Case I: Experimental Verification with the MFS Dataset

3.1. Experimental Setup and Data Description. To prove the effectiveness of the proposed fault diagnosis approach, a hybrid dataset of bearing and rotor faults collected by the machinery fault simulator (MFS) platform was used for experimental verification [54]. The experiment setup is shown in Figure 4, the MFS is driven by the AC motor with the speed of 2100 rpm, and the power of it is transmitted to the rotating plate and the drive shaft and through the coupling. The sampling frequency is 6 kHz. Through replacing different components, ten different types of datasets are collected with the data acquisition box, including nine fault types and one normal type, with detailed information shown in Table 1. There are 160 samples for each type, and each sample contains 1000 nonoverlapping data points.

To observe the difference between different types of vibration signals, a sample of each fault type is randomly selected to draw the waveform in time and frequency domains, respectively, and the results are shown in Figure 5. It can be seen that all fault types are time-varying and frequency-varying signals, which indicates that the original vibration signals are nonstationary. In addition, in view of the frequency domain, most types of fault information are concentrated in the low-frequency band, while the high-frequency band contains less.

Each sample data is standardized by the z-score method. In addition, missing values are detected in advance. If there is a missing value, it is filled with the Lagrange interpolation formula. The design and improvement of the experimental algorithms are implemented by Python 3.7.3 with a computer configured with Intel Core i5-6000hq CPU and 12G RAM.

3.2. Research on Feature Extraction

3.2.1. Parameter Settings of OWPD. Different decomposition layers in OWPD will result in different frequency resolutions of subsignals, which affect the accuracy and time

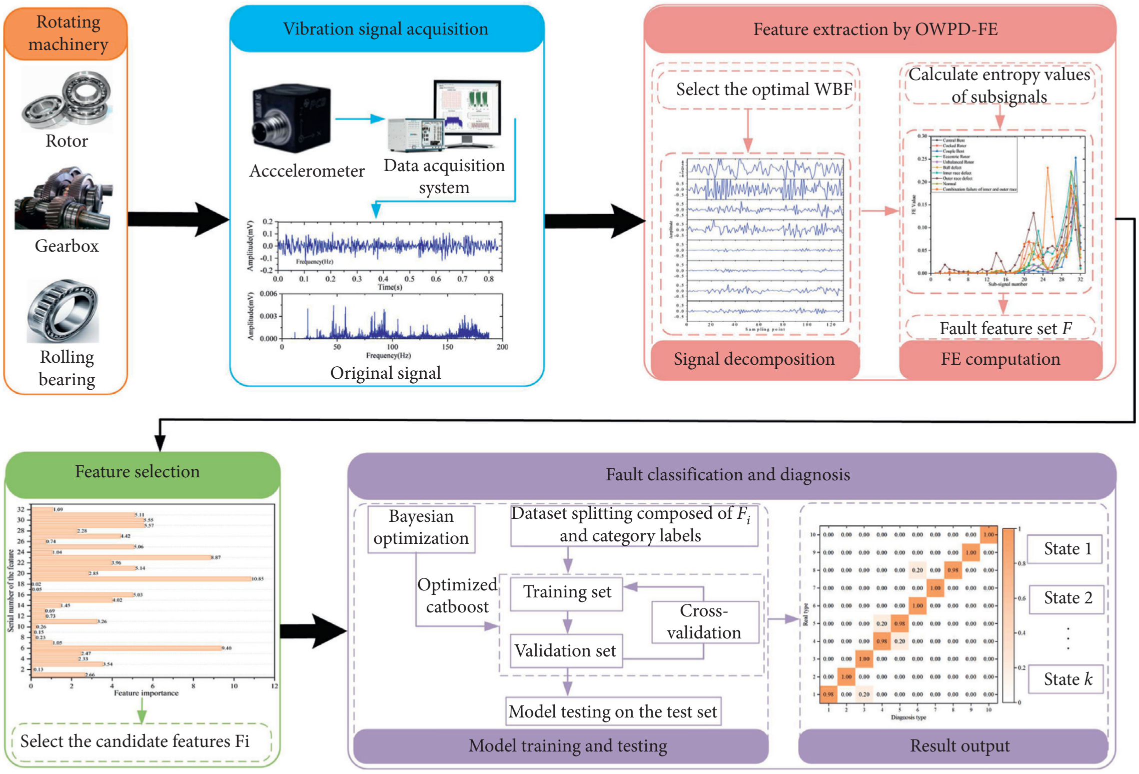


Figure 3: The flowchart of the proposed approach.

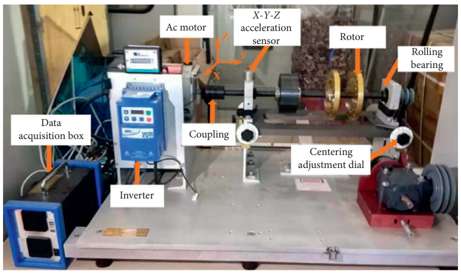


Figure 4: Experimental platform of the MFS dataset [54].

Table 1: Details of ten types of faults.

Fault class	Label
Ball fault	1
Inner race fault	2
Outer race fault	3
Combination fault	4
Normal state	5
Spindle central bent	6
Couple bent	7
Cocked rotor	8
Unbalanced rotor	9
Eccentric rotor	10

consumption of fault diagnosis. If the decomposition layer is l, the frequency resolution of the signal is

$$d_f = \frac{f_s}{2^{l+1}}, \tag{25}$$

where f_s is the sampling frequency, and here, it is 6 kHz. To make sure that d_f is greater than 1 Hz, the value of l needs to be less than 12. In addition, the number of features and calculation time will increase with the increase in the number of subbands. Therefore, l is preliminarily selected as 5, and each sample is averagely separated into 32 parts in the frequency domain. The influence of the value of l on the diagnostic results will be analysed in detail in the following experiments.

The optimal WBF will be selected according to the two-step principle described in Section 2.1.2. To reduce the calculation time, 10 samples are randomly selected under each fault type to form a new dataset, and 5-level WPD is carried out for each sample with different WBFs. In the first step, the total energy-to-Shannon entropy ratio ζ of each sample is calculated, respectively, according to equations (5)–(8), and then the average value of these 100 samples is taken. The results are detailed in Table 2. As can be seen, the WBFs with the largest average ζ value are db7, sym7, and coif3 in the same wavelet family, respectively.

In the second step, the above candidate WBF is used to reconstruct the signal, and the average value of its similarity coefficient d with the original signal is calculated according to equation (9). The results are detailed in Table 3. As can be seen, the original signal is most similar to the reconstructed signal when coif3 WBF is selected. coif WBF has orthogonality and compact support. In addition, compared with db WBF, it has better symmetry. Therefore, it is more effective to extract fault vibration signals of impulsive and nonstationary characteristics.

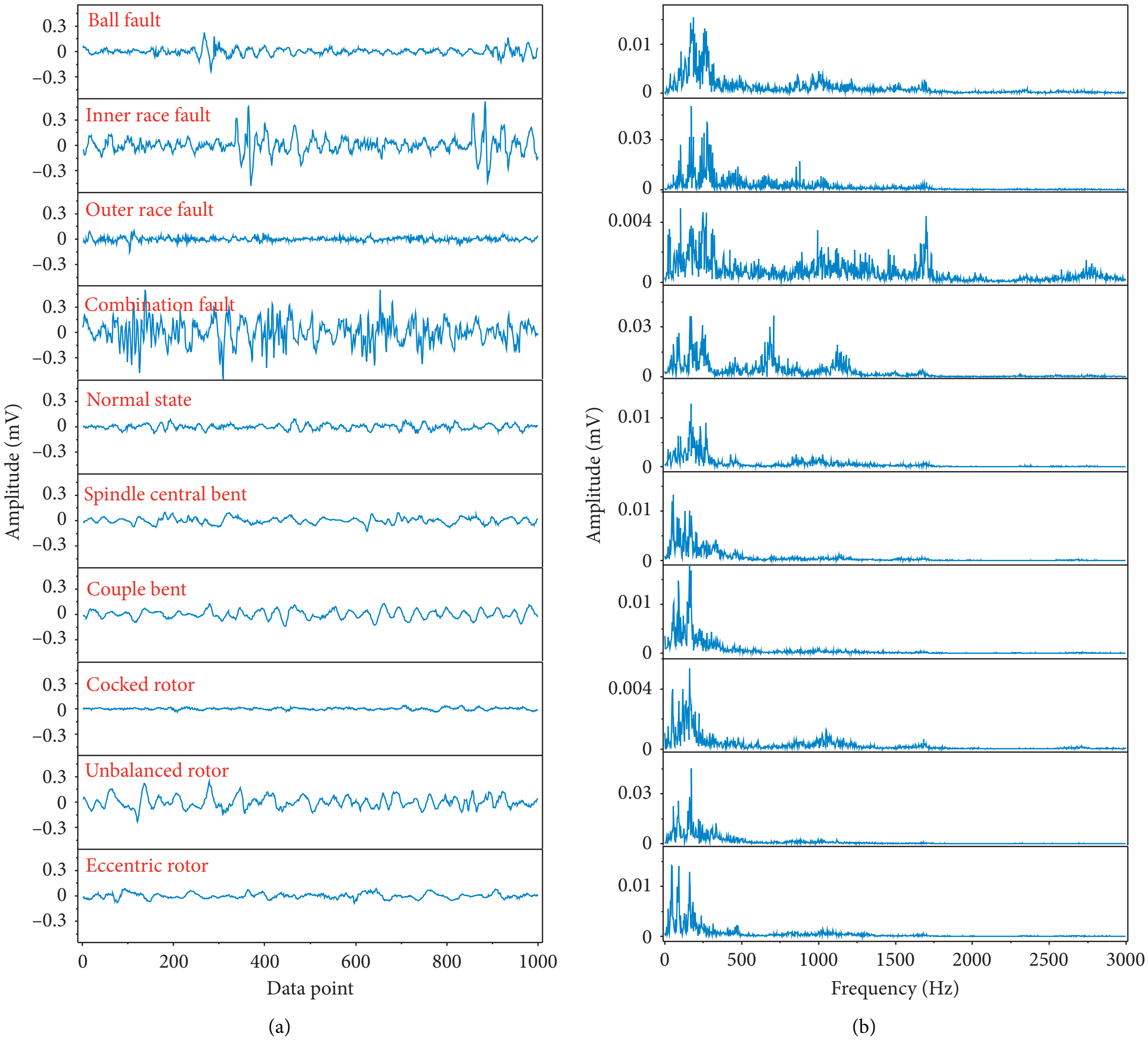


Figure 5: Time- and frequency-domain waveforms of ten fault types. (a) Time domain. (b) Frequency domain.

3.2.2. Parameter Settings of FE. After determining the decomposition layer and optimal WBF, the original signals are decomposed by OWPD. Then, the FE values of these subsignals are calculated separately. That is to say, the dimension of the feature set is 32. The parameters of FE are set according to [14, 18], as shown in Table 4, where STD represents the standard deviation of each sample. Time consumption of feature extraction is 1.29 s/sample. To visualize the calculation results, a sample from each fault type is randomly selected to carry out 5-layer OWPD using coif3 WBF. Figure 6 shows the FE values of the subsignals of different fault types. It can be seen that the difference of FE values among different fault types is quite obvious, which indicates that the proposed OWPD-FE method can effectively extract fault features. In addition, the FE values of some subsignals for certain fault types are almost 0, which also indicates that there is some feature redundancy.

3.2.3. Feature Visualization. t-SNE is the most commonly used algorithm for data visualization and dimensionality reduction [55]. Here, it is adopted to project the extracted 32-dimensional features into two-dimensional (2D) space for visualization. For each fault type, 20 groups of data after dimensionality reduction by t-SNE are selected to draw a scatter diagram, as shown in Figure 7. It is clear that, in the 2D plane, the features of ten fault types overlap little, and the boundaries among different types can be clearly distinguished, which shows that the OWPD-FE method is effective in extracting fault information of bearings and rotors.

3.3. Feature Selection Using CatBoost. An appropriate feature selection method can reduce the feature redundancy and improve the diagnostic performance of the model. In this paper, CatBoost algorithm is considered for this step. According to Section 2.3.2, equations (22) and (23) are used to get the values of importance of the 32 fault features extracted by the OWPD-FE method, according to which the feature importance can be ranked in the descending order. Then, candidate features and their corresponding labels are added sequentially to train and test the model. Finally, through the relationship between the number of features and

Table 2: The average value of the total energy-to-Shannon entropy ratio ζ.

WBF	Average ζ
db1	4.538
db2	5.349
db3	5.660
db4	5.960
db5	5.986
db6	5.951
db7	6.311
db8	5.986
db9	6.285
db10	6.237
sym2	5.349
sym3	5.660
sym4	5.854
sym5	6.044
sym6	5.968
sym7	6.310
sym8	6.151
coif1	5.362
coif2	5.857
coif3	6.319
coif4	6.057
coif5	6.241
—	—
—	—

Table 3: The average value of the similarity coefficient d.

WBF	db7	sym7	coif3
Similarity coefficient d (10^{-12})	20.036	9.213	7.632

Table 4: Parameter settings of the FE method.

Parameter	Description	Value
Λ	Time delay	1
m	Embedding dimension	2
r	Similarity tolerance	0.15 * STD
n	Gradient of similarity tolerance	2

the classification accuracy, the feature set used to reach the highest classification accuracy can be obtained. Figure 8 shows the normalized calculation results of the importance of 32 features, whose sum is 100.

3.4. Diagnosis Results and Analysis. At first, the dataset is divided into training set and test set, and the ratio of them is set as 3 : 2. That is to say, 96 samples of each fault type are used for model training and the rest of 64 samples for test. In addition, ten-fold cross-validation is conducted on the training set. The main hyperparameters of CatBoost optimized by BO algorithm are shown in Table 5.

Then, to study the effect of the number of features on the classification results, the feature selection process is carried out according to the analysis of Section 3.3. The experimental results are detailed in Figure 9. It can be seen that the model training time is positively correlated with the number of features, which is consistent with the actual experience. The average accuracy of ten-fold cross-validation on the training set has reached 100% when 7 features are selected. When the number of features is 22, the test set accuracy is the highest, reaching 99.17%. However, when all 32 features are used, it decreased by 0.21 percent (only 98.96%). Therefore, the classification accuracy of using 22 features is considered as the final diagnosis result. The time consumption of model training in this case is 10.13 s, which is 1.68 s less than that without feature selection. Experimental results show the reliability of the proposed feature selection method and the effectiveness of the classification algorithm.

As indicated in Figure 10, the confusion matrix of the diagnosis result using 22 features is presented in detail. It is not hard to see that the diagnostic accuracy of all fault types is above 98%, and it reaches 100% for 6 fault types (corresponding category labels are 2, 3, 6, 7, 9, and 10). The experimental results show that the proposed approach can effectively identify the hybrid fault states of the rotor and bearing.

3.5. Comparison of Different Decomposition Layers. In this section, in order to further demonstrate the effectiveness and reasonability of setting decomposition layer l to 5, the influence of the value of l on the diagnostic performance is investigated. Firstly, l needs to be less than 12 according to equation (25). Therefore, l is set to 1 to 8, respectively, to perform OWPD. Then, the parameters of FE are set according to Table 4. The partition of the sample dataset is described in Section 3.4. CatBoost is still applied to feature selection and fault pattern recognition of the 8 datasets, and BO algorithm is used to optimize hyperparameters. The experimental results are shown in Figure 11.

As can be seen from Figure 11(a), the classification accuracy generally presents an upward trend with the increase of l. In detail, when l is equal to 1, the average accuracy of the training set and validation set is very low, only 82.02% and 53.75%, respectively, while the test set accuracy only reaches 57.29%. When l is greater than 2, all the training sets' accuracy reaches 100%. The validation set accuracy reaches the maximum when l is equal to 6 (99.29%). When l is equal to or greater than 5, all the test sets' accuracy reaches 99.17%. As can be seen from Figure 11(b), the time of feature extraction and model training increases with the increase of l. The feature extraction time increases exponentially with the increase of l. Experimental results show that high-quality features containing effective fault information can be obtained by selecting appropriate l to perform WPD. An inappropriate value of l will result in too low classification accuracy (l less than 4) or too high computational cost (l greater than 6). Therefore, it is reasonable to set l to 5 considering the classification accuracy and time consumption comprehensively.

In addition, to illustrate the necessity of feature selection, the diagnostic performance with and without feature selection is compared, and the experimental results are shown in Figure 12. As can be seen from Figure 12(a), there is no feature redundancy when l is less than 4. However, the number of redundant features gradually increases when l is

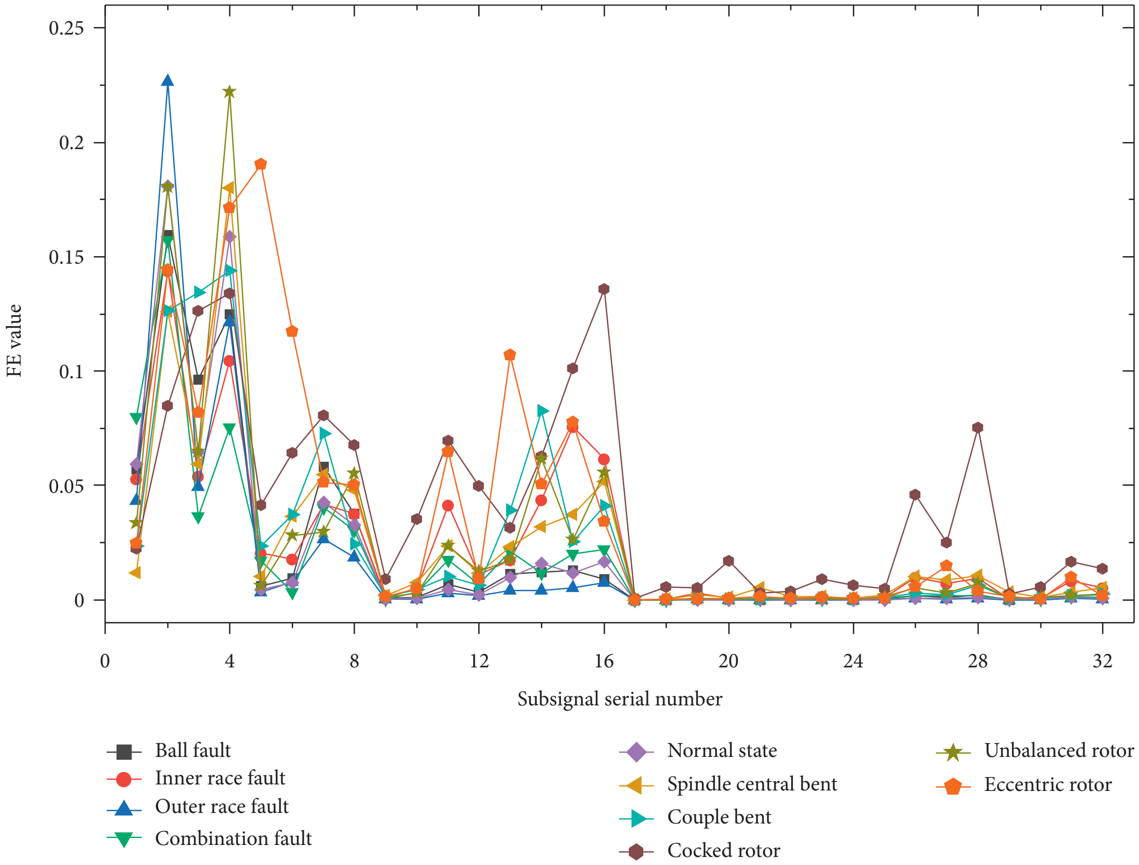


Figure 6: The FE values of subsignals decomposed by OWPD.

greater than 4, which is caused by the increase of frequency bands containing no or less fault information. Accordingly, the implementation of feature selection greatly reduces model training time from the perspective of computational cost. Meanwhile, it can be seen from Figure 12(b) that feature selection can also effectively improve the classification accuracy, especially when l is 5, 7, and 8.

3.6. Comparison of Different Classifiers. In this section, to justify the superiority of the proposed BO-CatBoost algorithm and the applicability of the OWPD-FE method combined with other classifiers, SVM, RF, GBDT, and XGBoost are adopted for comparison. The dataset consisting of 22 high-quality fault features described in Section 3.4 is input into the above classifiers, respectively, for model training and testing. To obtain the optimal diagnostic performance, the main hyperparameters of these classifiers are all optimized by BO algorithm, as detailed in Table 6. The diagnostic results are shown in Figure 13. It can be seen that the average training set accuracy of RF, GBDT, and CatBoost all reaches 100%, while SVM is the lowest, only 97.88%. In terms of the validation set accuracy, SVM has the lowest average accuracy (only 95.27%), while it is the highest of CatBoost (98.39%). RF has the lowest standard deviation, indicating that it is relatively stable. In terms of the test set accuracy, CatBoost is the highest, followed by GBDT and XGBoost (both 98.54%), while SVM performs the worst. In practical applications, different classifiers are applicable to different task requirements. Experimental results show that the diagnostic performance of GBDT algorithm and its variant (XGBoost) is close to that of CatBoost under the premise of setting appropriate hyperparameters. In addition, it is also demonstrated that the OWPD-FE method can extract high-quality fault features that are easy to identify.

4. Case II: Experiment Verification with the One-Stage Reduction Gearbox Dataset

4.1. Experimental Setup and Data Description. Since the working condition of the MFS dataset is relatively simple, this section further verifies the effectiveness of the proposed approach in practical applications through an actual gearbox dataset with more complex working conditions. The experimental platform is composed of a one-stage reduction gearbox, a torque sensor, a servo motor, etc., as shown in Figure 14. Four fault types are formed by processing gears with different crack lengths (0, 5, 10, and 15 mm). The sample frequency is 5 kHz. The details of the dataset can be found in [2]. Here, data collected under 20 different working conditions are used, as shown in Table 7. These data constitute 10 different datasets, as shown in Table 8. Datasets D1 to D9 contain a relatively simple number of working conditions. Composed of all 20 working conditions, D10 is the most complex dataset that is closest to the actual working

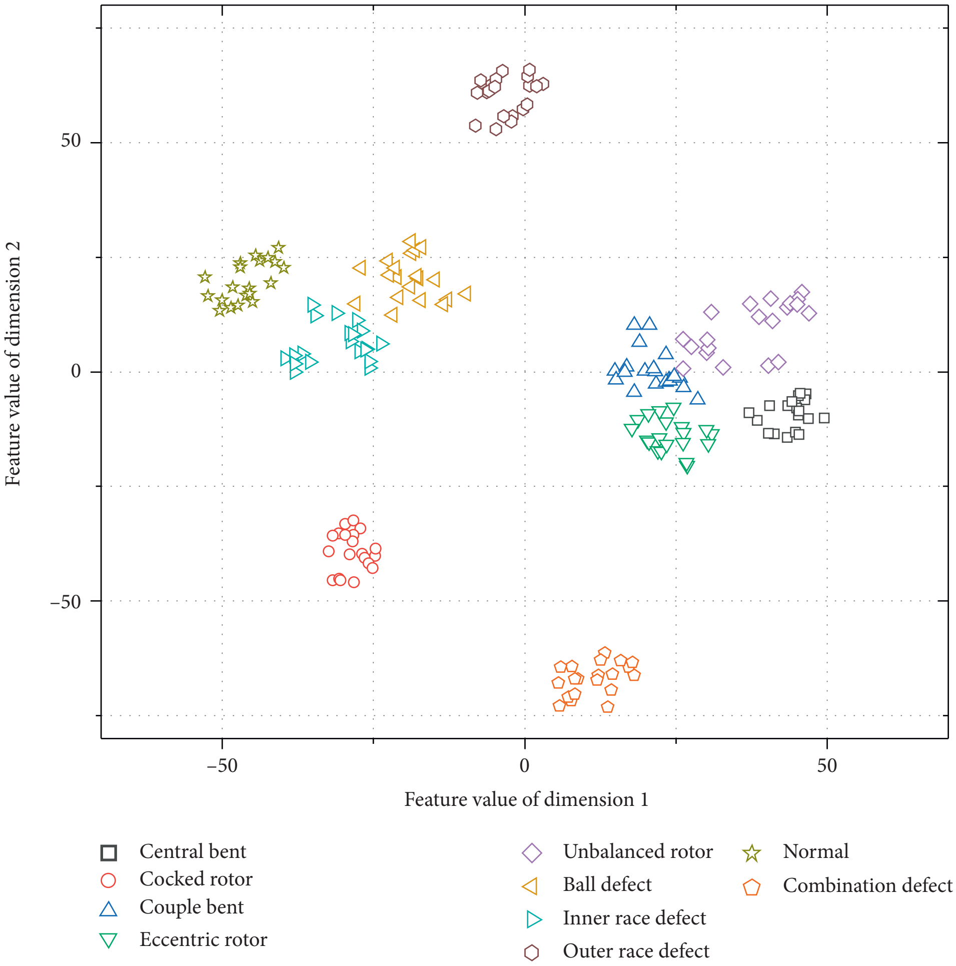


Figure 7: 2D projection of the extracted fault features using t-SNE.

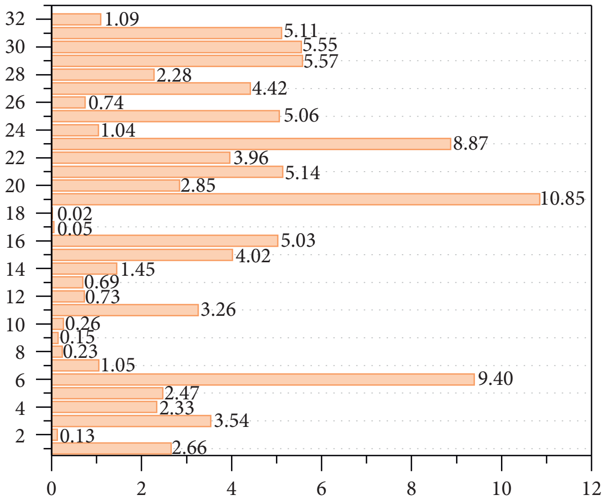


Figure 8: The value of the normalized feature importance.

conditions. For a single working condition, there are 40 samples for each fault type. The total number of samples is 3200, each containing 1500 consecutive data points.

4.2. Diagnosis Results and Analysis. The steps and settings of feature extraction refer to the descriptions in Section 3.2, where sym5 is the optimal WBF. The BO-CatBoost

Table 5: The main hyperparameters of CatBoost optimized by BO algorithm.

Parameter	Value
Loss function	Cross-entropy
Number of estimators	450
L2 regularization	3
Learning rate	0.254
Max depth of one tree	4
Random seed number	50

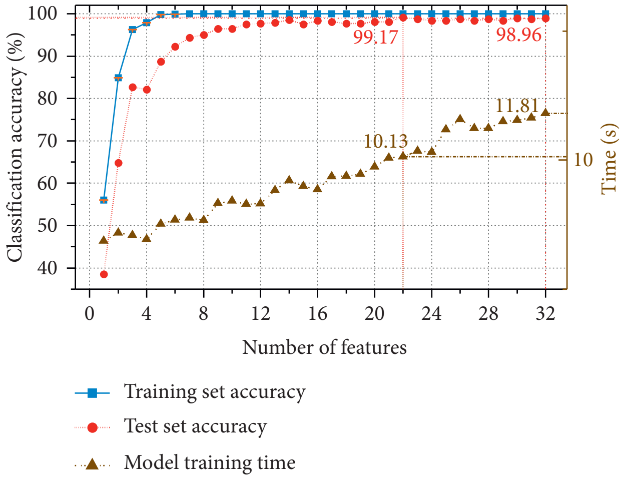
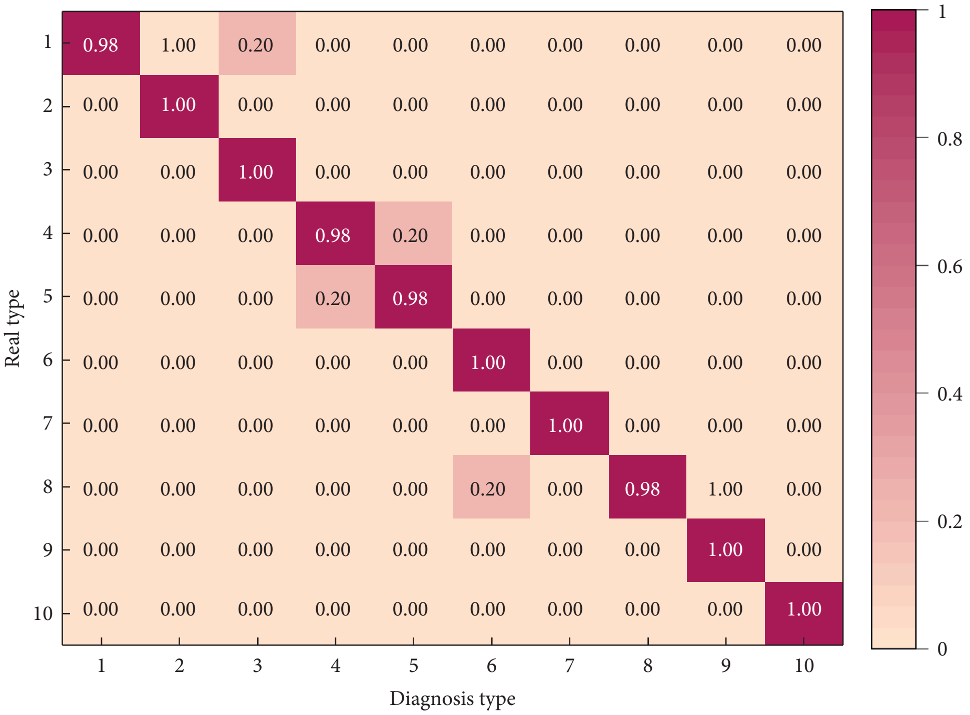


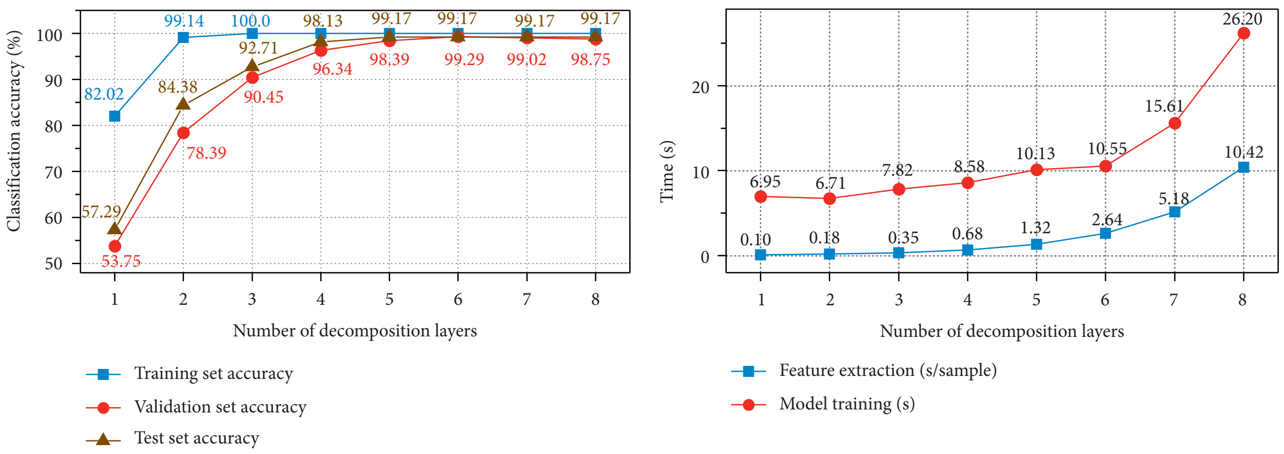
Figure 9: Diagnosis results and time consumption using different number of features.



	1	2	3	4	5	6	7	8	9	10
1	0.98	1.00	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.98	0.20	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.20	0.98	0.00	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
7	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
8	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.98	1.00	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Figure 10: Confusion matrix of using 22 features selected by CatBoost.

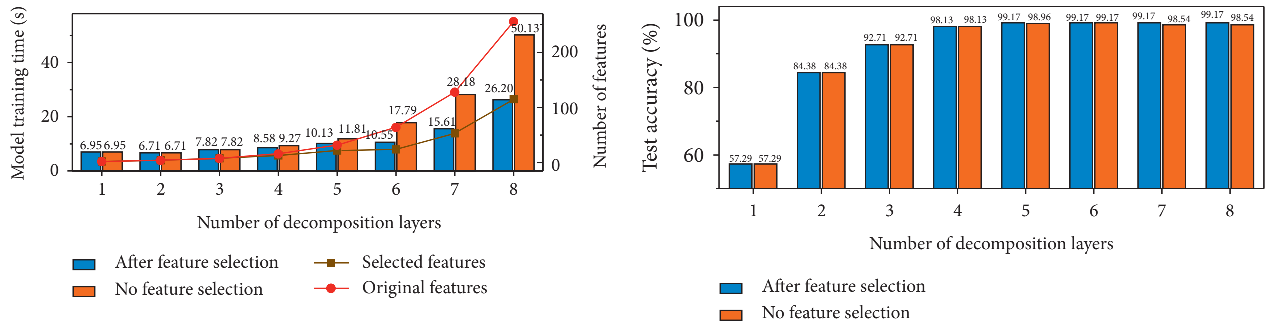
algorithm is still used to identify and diagnose fault types. Figure 15 shows the diagnostic results of ten datasets after selecting the optimal feature subset. For all ten datasets, the training set accuracy reaches 100%. For D1 to D9, the test set accuracy is higher than 96.56% and even higher than 99% on D2, D6, D7, and D9. The test set accuracy of D10 under the most complicated conditions is 98.65%, which is 0.22 percent higher than using the default CatBoost



(a)

(b)

Figure 11: The change of diagnostic results with the number of decomposition layers. (a) Classification accuracy. (b) Feature extraction and model training time.



(a)

(b)

Figure 12: Comparison of diagnostic results with or without feature selection. (a) Model training time and the number of features used for classification. (b) Test set accuracy.

Table 6: The main hyperparameters of different classifiers optimized by BO algorithm.

Classifier	Parameter	Value
SVM	Penalty factor	1799
	Coefficient of the kernel function	0.833
	Kernel function	rbf
	Decision function shape	ovo
RF	Number of estimators	928
	Max depth of one tree	3
	Maximum number of features allowed for a single tree	$0.1N_f$ [1]
	Minimum number of samples needed to split a node	6
GBDT	Number of estimators	208
	Learning rate	0.115
	Max depth of one tree	2
	Subsampling ratio	0.795
XGBoost	Number of estimators	516
	Learning rate	0.129
	Max depth of one tree	4
	Proportion of randomly sampled features in each tree	$0.584N_f$ [1]
	L1 regularization	0.105
	L2 regularization	0.591
	Subsampling ratio	0.1
	Booster model	gbtree
	Objective function	multi: softmax
	Number of classes	10

[1] N_f denotes the total number of features.

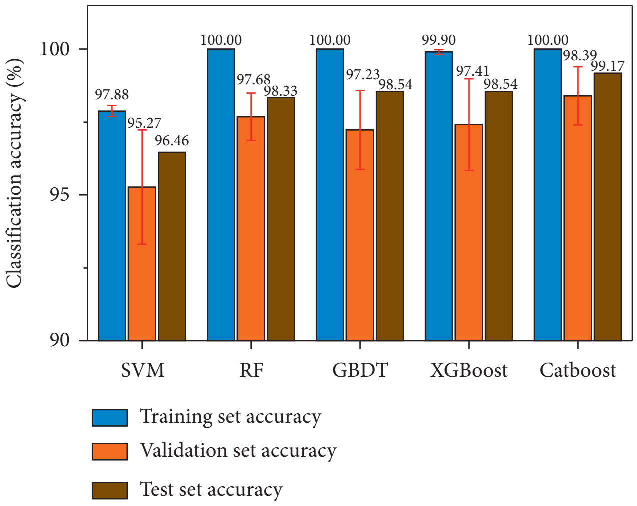


Figure 13: Comparison of classification accuracy of different classifiers.

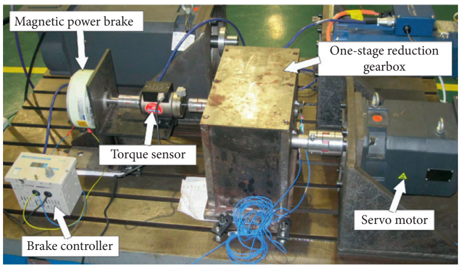


Figure 14: Experimental platform for the one-stage gearbox [2].

Table 7: 20 different working conditions of the one-stage gearbox.

Shaft speed (rpm)	Load (N·m)				
	0	2	4	6	8
600	W_1	W_2	W_3	W_4	W_5
900	W_6	W_7	W_8	W_9	W_{10}
1200	W_{11}	W_{12}	W_{13}	W_{14}	W_{15}
1500	W_{16}	W_{17}	W_{18}	W_{19}	W_{20}

Table 8: Details of 10 datasets used for experimental verification.

Label	Working conditions	Sample size
D1	W_1, W_2, W_3, W_4, W_5	800
D2	$W_6, W_7, W_8, W_9, W_{10}$	800
D3	$W_{11}, W_{12}, W_{13}, W_{14}, W_{15}$	800
D4	$W_{16}, W_{17}, W_{18}, W_{19}, W_{20}$	800
D5	W_1, W_6, W_{11}, W_{16}	640
D6	W_2, W_7, W_{12}, W_{17}	640
D7	W_3, W_8, W_{13}, W_{18}	640
D8	W_4, W_9, W_{14}, W_{19}	640
D9	$W_5, W_{10}, W_{15}, W_{20}$	640
D10	$W_1, W_2, \ldots, W_{20}$	3200

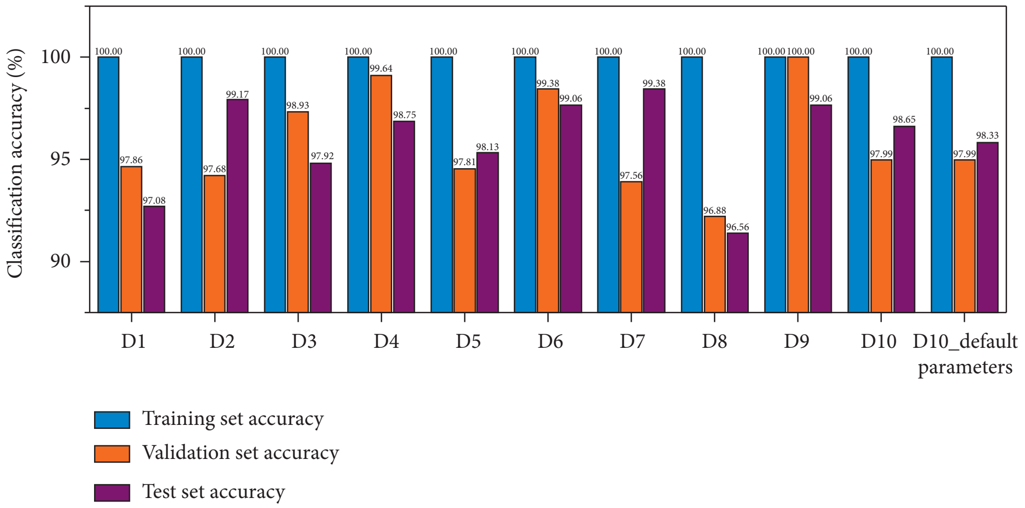


Figure 15: The classification results of the 10 datasets used for experimental verification.

hyperparameters. However, in terms of time consumption, it takes 206.09 s to complete the training process with default hyperparameters, while 6.94 s with hyperparameters optimized by BO algorithm, which greatly improves the computational efficiency. Therefore, experimental results in this study show that the proposed approach is also effective for fault diagnosis of the gearbox.

5. Conclusions

In this paper, aiming at the fault diagnosis of rotating machinery under complex working conditions, a novel approach based on feature importance ranking and selection is proposed. Firstly, the OWPD method is proposed to decompose the vibration signal, where a two-step principle of selecting the optimal WBF is introduced. On this basis, it is combined with FE to extract hidden and high-quality fault features from the decomposed subsignals. Then, in order to filter out redundant fault features that are not conducive to the diagnosis result, the CatBoost model is constructed and preliminarily applied to calculate the importance of each feature for further feature selection. Moreover, the classification model based on BO-CatBoost algorithm can effectively solve the optimization problem of hyperparameters, which can greatly reduce model training time and improve the diagnosis accuracy. Finally, experimental results on the MFS dataset and the one-stage gearbox dataset under complex working conditions demonstrate the practicability and the generalization performance of the proposed approach, and the classification accuracy reaches 99.17% and above 96.56%, respectively. In addition, the robustness of the proposed approach under different working conditions is also verified by the one-stage gearbox dataset.

In the future work, the effect of combining OWPD with other information entropies or some dimensionless time-domain indexes still needs to be discussed. In addition, the vibration data collected by only one acceleration sensor are utilized in this paper, while the fusion of multisensor data may provide more real fault information, which is also worth investigating in the next step.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research was funded by the National Key Research and Development Program of China (no. 2020YFB1711203), the National Natural Science Foundation of China (no. 61703385), the Research Fund (no. 61400020401), and Fundamental Research Funds for the Central Universities (nos. 2019kfyXJJS137, NJ2019007, and NJ2020014).

References

[1] W. Zhang, C. Li, G. Peng, Y. Chen, and Z. Zhang, “A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load,” *Mechanical Systems and Signal Processing*, vol. 100, pp. 439–453, 2018.

[2] J. Liu, Y. Hu, Y. Wang, B. Wu, J. Fan, and Z. Hu, “An integrated multi-sensor fusion-based deep feature learning approach for rotating machinery diagnosis,” *Measurement Science and Technology*, vol. 29, Article ID 055103, 2018.

[3] P. Liang, C. Deng, J. Wu, and Z. Yang, “Intelligent fault diagnosis of rotating machinery via wavelet transform, generative adversarial nets and convolutional neural network,” *Measurement*, vol. 159, Article ID 107768, 2020.

[4] Y. Wang, P. W. Tse, B. Tang et al., "Order spectrogram visualization for rolling bearing fault detection under speed variation conditions," *Mechanical Systems and Signal Processing*, vol. 122, pp. 580–596, 2019.

[5] B. Yang, Y. Lei, F. Jia, and S. Xing, "An intelligent fault diagnosis approach based on transfer learning from laboratory bearings to locomotive bearings," *Mechanical Systems and Signal Processing*, vol. 122, pp. 692–706, 2019.

[6] X. Zhang, J. Wang, Z. Liu, and J. Wang, "Weak feature enhancement in machinery fault diagnosis using empirical wavelet transform and an improved adaptive bistable stochastic resonance," *ISA Transactions*, vol. 84, pp. 283–295, 2019.

[7] Y. Lan, J. Hu, J. Huang et al., "Fault diagnosis on slipper abrasion of axial piston pump based on extreme learning machine," *Measurement*, vol. 124, pp. 378–385, 2018.

[8] A. Krishnakumari, A. Elayaperumal, M. Saravanan, and C. Arvindan, "Fault diagnostics of spur gear using decision tree and fuzzy classifier," *The International Journal of Advanced Manufacturing Technology*, vol. 89, no. 9–12, pp. 3487–3494, 2017.

[9] Z. Feng, W. Zhu, and D. Zhang, "Time-Frequency demodulation analysis via Vold-Kalman filter for wind turbine planetary gearbox fault diagnosis under nonstationary speeds," *Mechanical Systems and Signal Processing*, vol. 128, pp. 93–109, 2019.

[10] X. Zhang, B. Wang, and X. Chen, "Intelligent fault diagnosis of roller bearings with multivariable ensemble-based incremental support vector machine," *Knowledge-Based Systems*, vol. 89, pp. 56–85, 2015.

[11] W. Deng, H. Liu, J. Xu, H. Zhao, and Y. Song, "An improved quantum-inspired differential evolution algorithm for deep belief network," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 10, pp. 7319–7327, 2020.

[12] Z. Zhang, Y. Wang, and K. Wang, "Fault diagnosis and prognosis using wavelet packet decomposition, Fourier transform and artificial neural network," *Journal of Intelligent Manufacturing*, vol. 24, no. 6, pp. 1213–1227, 2013.

[13] D. Zhen, J. Guo, Y. Xu, H. Zhang, and F. Gu, "A novel fault detection method for rolling bearings based on non-stationary vibration signature analysis," *Sensors*, vol. 19, no. 18, p. 3994, 2019.

[14] A. Humeau-Heurtier, "The multiscale entropy algorithm and its variants: a review," *Entropy*, vol. 17, no. 5, pp. 3110–3123, 2015.

[15] Q. Wu and H. Lin, "Daily urban air quality index forecasting based on variational mode decomposition, sample entropy and LSTM neural network," *Sustainable Cities and Society*, vol. 50, Article ID 101657, 2019.

[16] M. M. Mafarja and S. Mirjalili, "Hybrid binary ant lion optimizer with rough set and approximate entropy reducts for feature selection," *Soft Computing*, vol. 23, no. 15, pp. 6249–6265, 2019.

[17] M. Zanin, L. Zunino, O. A. Rosso, and D. Papo, "Permutation entropy and its main biomedical and econophysics applications: a review," *Entropy*, vol. 14, no. 8, pp. 1553–1577, 2012.

[18] H. Azami, A. Fernández, and J. Escudero, "Refined multiscale fuzzy entropy based on standard deviation for biomedical signal analysis," *Medical & Biological Engineering & Computing*, vol. 55, no. 11, pp. 2037–2052, 2017.

[19] Y. Li, G. Li, Y. Wei, B. Liu, and X. Liang, "Health condition identification of planetary gearboxes based on variational mode decomposition and generalized composite multi-scale symbolic dynamic entropy," *ISA Transactions*, vol. 81, pp. 329–341, 2018.

[20] J. Wang, O. Tawose, L. Jiang, and D. Zhao, "A new data fusion algorithm for wireless sensor networks inspired by hesitant fuzzy entropy," *Sensors*, vol. 19, no. 4, p. 784, 2019.

[21] D. Zhao, S. Liu, D. Gu et al., "Improved multi-scale entropy and it's application in rolling bearing fault feature extraction," *Measurement*, vol. 152, p. 107361, 2020.

[22] P. Flandrin, "Empirical mode decomposition as a filter bank," *IEEE Signal Processing Letters*, vol. 11, pp. 112–114, 2004.

[23] J. Cheng, Y. Yang, and Y. Yang, "A rotating machinery fault diagnosis method based on local mean decomposition," *Digital Signal Processing*, vol. 22, no. 2, pp. 356–366, 2012.

[24] L. Wang, Z. Liu, Q. Miao, and X. Zhang, "Time-frequency analysis based on ensemble local mean decomposition and fast kurtogram for rotating machinery fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 103, pp. 60–75, 2018.

[25] K. Dragomiretskiy and D. Zosso, "Variational mode decomposition," *IEEE Transactions on Signal Processing*, vol. 62, no. 3, pp. 531–544, 2014.

[26] S. Zhang, H. Zhao, J. Xu, and W. Deng, "A novel fault diagnosis method based on improved adaptive variational mode decomposition, energy entropy, and probabilistic neural network," *Transactions of The Canadian Society for Mechanical Engineering*, vol. 44, no. 1, pp. 121–132, 2020.

[27] Y. Miao, M. Zhao, and J. Lin, "Identification of mechanical compound-fault based on the improved parameter-adaptive variational mode decomposition," *ISA Transactions*, vol. 84, pp. 82–95, 2019.

[28] S. Chen, Y. Yang, Z. Peng, S. Wang, W. Zhang, and X. Chen, "Detection of rub-impact fault for rotor-stator systems: a novel method based on adaptive chirp mode decomposition," *Journal of Sound and Vibration*, vol. 440, pp. 83–99, 2019.

[29] R. Paul and A. Sengupta, "Design and application of discrete wavelet packet transform based multiresolution controller for liquid level system," *ISA Transactions*, vol. 71, pp. 585–598, 2017.

[30] A. Tabrizi, L. Garibaldi, A. Fasana, and S. Marchesiello, "Early damage detection of roller bearings using wavelet packet decomposition, ensemble empirical mode decomposition and support vector machine," *Meccanica*, vol. 50, no. 3, pp. 865–874, 2015.

[31] L.-Y. Zhao, L. Wang, and R.-Q. Yan, "Rolling bearing fault diagnosis based on wavelet packet decomposition and multiscale permutation entropy," *Entropy*, vol. 17, no. 12, pp. 6447–6461, 2015.

[32] W. Deng, J. Xu, Y. Song, and H. Zhao, "Differential evolution algorithm with wavelet basis function and optimal mutation strategy for complex optimization problem," *Applied Soft Computing*, vol. 100, Article ID 106724, 2020.

[33] P. K. Kankar, S. C. Sharma, and S. P. Harsha, "Fault diagnosis of rolling element bearing using cyclic autocorrelation and wavelet transform," *Neurocomputing*, vol. 110, pp. 9–17, 2013.

[34] G. I. Sayed, A. E. Hassanien, and A. T. Azar, "Feature selection via a novel chaotic crow search algorithm," *Neural Computing and Applications*, vol. 31, no. 1, pp. 171–188, 2019.

[35] M. Wang and H. Chen, "Chaotic multi-swarm whale optimizer boosted support vector machine for medical diagnosis," *Applied Soft Computing*, vol. 88, p. 105946, 2020.

[36] L. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009.

[37] G. E. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[38] Y. Loozen, K. T. Rebel, S. M. de Jong et al., "Mapping canopy nitrogen in European forests using remote sensing and environmental variables with the random forests method," *Remote Sensing of Environment*, vol. 247, Article ID 111933, 2020.

[39] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of The ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[40] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic flow prediction with big data: a deep learning approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, pp. 865–873, 2015.

[41] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[42] Y. Ding, L. Ma, J. Ma, C. Wang, and C. Lu, "A generative adversarial network-based intelligent fault diagnosis method for rotating machinery under small sample size conditions," *IEEE Access*, vol. 7, pp. 149736–149749, 2019.

[43] L. Prokhorenkova, G. Gusev, and A. Vorobev, "CatBoost: unbiased boosting with categorical features," *Advances in Neural Information Processing Systems*, vol. 31, pp. 6638–6648, 2018.

[44] G. Huang, L. Wu, X. Ma et al., "Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions," *Journal of Hydrology*, vol. 574, pp. 1029–1041, 2019.

[45] W. Liu, K. Deng, X. Zhang et al., "A semi-supervised tri-CatBoost method for driving style recognition," *Symmetry*, vol. 12, no. 3, p. 336, 2020.

[46] W. Deng, J. Xu, and H. Zhao, "An enhanced MSIQDE algorithm with novel multiple strategies for global optimization problems," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pp. 1–10, 2020.

[47] Y. Gao, D. Wu, W. Deng et al., "MPPCEDE: multi-population parallel co-evolutionary differential evolution for parameter optimization," *Energy Conversion and Management*, vol. 228, Article ID 113661, 2021.

[48] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas, "Taking the human out of the loop: a review of bayesian optimization," *Proceedings of the IEEE*, vol. 104, pp. 148–175, 2016.

[49] M. J. Baker, J. Trevisan, P. Bassan et al., "Using Fourier transform IR spectroscopy to analyze biological materials," *Nature Protocols*, vol. 9, no. 8, p. 1771, 2014.

[50] H. Bendjama, S. Bouhouche, and M. S. Boucherit, "Application of wavelet transform for fault diagnosis in rotating machinery," *International Journal of Machine Learning and Computing*, vol. 2, pp. 82–87, 2012.

[51] S. Ekici, S. Yildirim, and M. Poyraz, "Energy and entropy-based feature extraction for locating fault on transmission lines by using neural network and wavelet packet decomposition," *Expert Systems with Applications*, vol. 34, no. 4, pp. 2937–2944, 2008.

[52] R. Yan, *Base Wavelet Selection Criteria for Non-stationary Vibration Analysis in Bearing Health Diagnosis*, University of Massachusetts Amherst, Amherst, MA, USA, 2007.

[53] I. Babajide Mustapha and F. Saeed, "Bioactive molecule prediction using extreme gradient boosting," *Molecules*, vol. 21, no. 8, p. 983, 2016.

[54] Y. Shan, J. Zhou, W. Jiang, and J. Liu, "A fault diagnosis method for rotating machinery based on improved variational mode decomposition and a hybrid artificial sheep algorithm," *Measurement Science and Technology*, vol. 30, Article ID 055002, 2019.

[55] G. C. Linderman, M. Rachh, J. G. Hoskins, S. Steinerberger, and Y. Kluger, "Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data," *Nature Methods*, vol. 16, no. 3, pp. 243–245, 2019.