Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 9963146, 17 pages
https://doi.org/10.1155/2021/9963146

Research Article

Study on Intelligent Diagnosis of Rotor Fault Causes with the PSO-XGBoost Algorithm

Kai Gu,[1] **Jianqi Wang**,[2] **Hong Qian**,[1,2] **and Xiaoyan Su**[1,2]

[1]*State Key Laboratory of Nuclear Power Safety Monitoring Technology and Equipment, Shenzhen, China*
[2]*School of Automation Engineering, Shanghai University of Electric Power, Shanghai 200090, China*

Correspondence should be addressed to Jianqi Wang; wang_jianqi@foxmail.com

Received 11 March 2021; Revised 6 April 2021; Accepted 10 April 2021; Published 27 April 2021

Academic Editor: Venkatesan Rajinikanth

On basis of fault categories detection, the diagnosis of rotor fault causes is proposed, which has great contributions to the field of intelligent operation and maintenance. To improve the diagnostic accuracy and practical efficiency, a hybrid model based on the particle swarm optimization-extreme gradient boosting algorithm, namely, PSO-XGBoost is designed. XGBoost is used as a classifier to diagnose rotor fault causes, having good performance due to the second-order Taylor expansion and the explicit regularization term. PSO is used to automatically optimize the process of adjusting the XGBoost's parameters, which overcomes the shortcomings when using the empirical method or the trial-and-error method to adjust parameters of the XGBoost model. The hybrid model combines the advantages of the two algorithms and can diagnose nine rotor fault causes accurately. Following diagnostic results, maintenance measures referring to the corresponding knowledge base are provided intelligently. Finally, the proposed PSO-XGBoost model is compared with five state-of-the-art intelligent classification methods. The experimental results demonstrate that the proposed method has higher diagnostic accuracy and practical efficiency in diagnosing rotor fault causes.

1. Introduction

The steam turbine rotor plays an important role in transforming thermal energy into mechanical energy. In a high-speed rotating working station, any defect on the rotor will affect the safe running and even cause serious accidents [1–3]. Therefore, intelligent diagnosis of rotor fault causes is essential besides diagnosing rotor fault categories intelligently.

In the field of industrial intelligent operation and maintenance, the research studies mainly focus on the detection of rotor fault categories [4–6], while the studies on the diagnosis of rotor fault causes are less. The specific rotor fault causes provide a reasonable and practical maintenance decision, ensuring the steam turbine's safe and stable running. The traditional diagnosis of rotor fault causes is mainly based on the expert system [7], but the knowledge is difficult to obtain, and the portability is poor. A serious of running parameters, such as temperature and pressure, can accurately assess the operating status of equipment [8], but they are rarely used to build the intelligent diagnosis system of rotor fault causes. Therefore, the intelligent algorithms can be used to diagnose rotor fault causes and realize the intelligent operation and maintenance depending on running parameters of a rotor.

In essence, diagnosing rotor fault causes is a classification problem, and various intelligent classification methods have been applied. Support vector machine [9] (SVM) is a popular supervised learning algorithm that many researchers have used to train for classification. Jan et al. [10] used SVM classify sensor faults. Lobato et al. [11] used SVM for the classification of the machinery condition. However, the intelligent diagnosis of rotor fault causes is a typical nonlinear problem. Because the principle of SVM is a linear classifier based on maximum interval, it does not work well in solving nonlinear problems. Random forest [12] (RF) and gradient boosting decision tree [13] (GBDT) are two well-known ensemble machine learning methods, and the weak learning model used in them is the decision tree (DT) model. Wang et al. [14] proposed a hybrid approach of a random forest classifier for the fault diagnosis in rolling bearings. Quiroz et al. [15] used random forests to diagnose broken

rotor bar failure in a line start-permanent magnet synchronous motor. Zhu et al. [16] proposed a novel performance fault diagnosis method for SaaS software based on the GBDT algorithm. Zhong et al. [17] used GBDT to predict railway accident types and analyze causes. Although RF and GBDT have advantages such as high classification accuracy, less overfit, excellent generalization performance, and a good explanation, they also have some shortcomings for the intelligent diagnosis of rotor fault causes. RF may not produce good classification for small data or low dimensional data. GBDT uses the first-order Taylor expansion to calculate the loss, which is not accurate enough. On basis of GBDT, the extreme gradient boosting (XGBoost) algorithm was proposed by Chen Tianqi [18]. The XGBoost algorithm introduces second-order derivatives and regularization terms, which improve the accuracy on classification no matter whether the data scale is large or small. Zhang et al. [19] designed a data-driven method for fault detection of wind turbines using XGBoost. Lei et al. [20] diagnosed hydraulic valves by integrating PCA and XGBoost. Wu et al. [21] proposed a method of wind turbine fault diagnosis based on the ReliefF algorithm and XGBoost algorithm in order to improve the accuracy of fault diagnosis on wind turbines. Although the XGBoost has excellent classification results, there are many parameters in the XGBoost model, such as the learning rate, the subsample ration of columns when constructing each tree, the subsample ration of columns for each level, the regularization term on weights, and so on. Different combinations of these parameters determine the performance of the model to a large extent [22]. Usually, the parameter setting of the XGBoost model is to find a set of parameters making the performance best by fixing the values of several parameters and optimizing other parameters by a finite number of exhaustive methods. But different permutations and combinations increase the complexity of the work, and it is difficult to find the optimal parameters. It is an optimization problem to find the most suitable parameters of the XGBoost model. In recent years, various intelligent optimization algorithms have been proposed one by one [23–25]. In [26], an improved PSO-based QEA method was proposed to allocate gate resource. In [27], an enhanced MSIQDE algorithm with multiple strategies was proposed to solve global optimization problems. In [28], an enhanced success history adaptive DE with greedy mutation strategy is employed to optimize parameters of PV models. Aiming at the optimization problems of model parameters, particle swarm optimization (PSO) has simple principle and easy implementation. Many researchers have achieved better results by combing PSO with other classification methods. Wang et al. [29] used PSO to search the optimal architecture of convolution neural networks. Li et al. [30] used PSO to search the penalty factor and kernel function of SVM.

To diagnose rotor fault causes accurately and efficiently, a hybrid model based on the particle swarm optimization-extreme gradient boosting algorithm (PSO-XGBoost) is proposed. XGBoost, a scalable end-to-end tree boosting system, with the second-order Taylor expansion and the explicit regular term, is used as a classifier to diagnose the rotor fault causes. PSO is used to automatically optimize the parameters such as the L1 and L2 regularization terms on weights during the XGBoost model training, which overcomes the low accuracy and low efficiency when using the empirical method or the trial-and-error method to adjust these parameters of the XGBoost model. The hybrid model combined with the advantages of the two algorithms can diagnose rotor fault causes more accurately. Following diagnostic results, maintenance measures referring to the corresponding knowledge base are provided intelligently.

The innovations and main contributions of this study are described as follows:

(1) On basis of fault categories detection, the diagnosis of rotor fault causes is proposed, which has great contributions to the field of intelligent operation and maintenance

(2) A novel hybrid model based on PSO and XGBoost is developed to effectively simplify the parameter adjustment process of the XGBoost model and improve the accuracy of diagnosis

The further detailed structure of this study is summarized in the remaining sections. Section 2 introduces the preliminaries for diagnosis of rotor fault causes. Section 3 conducts an experiment to validate the performance of the proposed method. Conclusions are elaborated in Section 4.

2. Materials and Methods

2.1. XGBoost Algorithm. XGBoost [16] is a highly scalable end-to-end tree boosting system, which provides a theoretically justified weighted quantile sketch for efficient proposal calculation, a novel sparsity-aware algorithm for parallel tree learning, and an effective cache-aware block structure for out-of-core tree learning.

For a given dataset with n examples and m features $D = \{(X_i, y_i)\}$ ($|D| = n, X_i \in \mathbb{R}^m, y_i \in \mathbb{R}$), a tree ensemble model uses K additive functions to predict the output. Here, $X_i = [F_i, x_{i1}, x_{i2}, \ldots, x_{is}]$ is the characteristic parameter of the i^{th} sample, composed of fault types $F = \{F_1, F_2, \cdots, F_r\}$ and operating parameters $x = \{x_1, x_2, \ldots, x_s\}$, where r is the number of fault types, and s is the number of operating parameters. The predicted category of fault cause is

$$\widehat{y}_i = \sum_{k=1}^{K} f_k(X_i), \quad f_k \in \mathcal{F}, \tag{1}$$

where K is the number of trees, and f is a function in the functional space $\mathcal{F}$.

The objective function is

$$obj = \sum_{i=1}^{n} l(y_i, \widehat{y}_i) + \sum_{k=1}^{K} \Omega(f_k). \tag{2}$$

The first term $l(y_i, \widehat{y}_i)$ is the training loss function, and the second term $\Omega(f)$ is the regularization term. The training loss measures how predictive the model is with respect to the training data. The regularization term controls the complexity of the model, which helps to avoid overfitting.

Formally, let $\hat{y}_i^{(t)}$ be the prediction of the i^{th} instance at the t^{th} iteration; then, add f_t to minimize the following objective.

$$obj^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{i=1}^{t} \Omega\left(f_i\right) = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t\left(X_i\right)\right) + \Omega\left(f_t\right) + \text{constant}. \tag{3}$$

Second-order approximation can be used to optimize equation (3) in the general setting, i.e.,

$$obj^{(t)} \simeq \sum_{i=1}^{n}\left[l\left(y_i, \hat{y}^{(t-1)}\right) + g_i f_t\left(X_i\right) + \frac{1}{2}h_i f_t^2\left(X_i\right)\right] + \Omega\left(f_t\right) + \text{constant}, \tag{4}$$

where $g_i = \partial_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right)$ and $h_i = \partial^2_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right)$ are the first- and second-order gradient statistics on the loss function. After removing all the constants, the specific objective at step t becomes

$$\sum_{i=1}^{n}\left[g_i f_t\left(X_i\right) + \frac{1}{2}h_i f_t^2\left(X_i\right)\right] + \Omega\left(f_t\right). \tag{5}$$

The definition of the tree $f(X)$ is refined as

$$f_t(X) = \omega_{q(X)}, \quad \omega \in \mathbb{R}^T,\ q\colon \mathbb{R}^d \longrightarrow \{1, 2, \ldots, T\}. \tag{6}$$

Here, ω is the vector of scores on leaves, q is a function assigning each data point to the corresponding leaf, and T is the number of leaves. The regularization term is defined as

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda\sum_{i=1}^{T}\omega_j^2. \tag{7}$$

After reformulating the tree model, the objective value with the t^{th} tree can be written as

$$obj^{(t)} \simeq \sum\left[g_i\omega_{q\left(X_i\right)} + \frac{1}{2}h_i\omega^2_{q\left(X_i\right)}\right] + \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2 = \sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right)\omega_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right)\omega_j^2\right] + \gamma T, \tag{8}$$

where $I_j = \{i|q(X_i) = j\}$ is the set of indices of data points assigned to the j^{th} leaf.

By defining $G_j = \sum_{i\in I_j} g_i$ and $H_j = \sum_{i\in I_j} h_i$, the expression can be compressed as

$$obj^{(t)} = \sum_{j=1}^{T}\left[G_j\omega_j + \frac{1}{2}\left(H_j + \lambda\right)\omega_j^2\right] + \gamma T. \tag{9}$$

In equation (9), ω_j is independent with respect to each other, the form $G_j\omega_j + 1/2\,(H_j + \lambda)\omega_j^2$ is quadratic, and the best ω_j for a given structure $q(X)$ and the best objective reduction are

$$\omega_j^* = -\frac{G_j}{H_j + \lambda}, \tag{10}$$

$$obj* = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j + \lambda} + \gamma T. \tag{11}$$

Equation (11) measures how good a tree structure $q(X)$ is. Typically, it is impossible to enumerate all the possible tree structures q. A greedy algorithm that starts from a single leaf and iteratively adds branches to the tree is used instead. By splitting a leaf into two leaves, the score it gains is

$$\text{Gain} = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{\left(G_L + G_R\right)^2}{H_L + H_R + \lambda}\right] - \gamma. \tag{12}$$

Equation (12) can be decomposed as the score on the new left leaf, the score on the new right leaf, the score on the original leaf, and regularization on the additional leaf, respectively.

2.2. Particle Swarm Optimization Algorithm. The particle swarm optimization algorithm [23] is a popular population-based heuristic algorithm that is inspired by the foraging behavior of birds flocking.

Suppose a population $X = (X_1, X_2, \ldots, X_n)$ of the n particles in a D-dimensional search space, where the i^{th} particle is represented as a D-dimensional vector. According to the objective function, each particle's corresponding fitness of position can be calculated. The individual extremum of i^{th} particle's speed is $V_i = (V_{i1}, V_{i2}, \ldots, V_{i\,D})^T$. Corresponding individual extremum is $P_i = (P_{i1}, P_{i2}, \ldots, P_{i\,D})^T$. During each iteration, the particle updates its speed and position through the extremum of the individual and the extremum of the population as

$$V_{id}^{(t+1)} = \omega \times V_{id}^{(t)} + c_1 \times r_1 \times\left(P_{id}^{(t)} - X_{id}^{(t)}\right) + c_2 \times r_2 \times\left(P_{gd}^{(t)} - x_{id}^{(t)}\right), \tag{13}$$

$$X_{id}^{(t+1)} = X_{id}^{(t)} + V_{id}^{(t+1)}. \tag{14}$$

In equations (13) and (14) above, ω is the inertia weight; $d = 1, 2, \ldots, D$; $i = 1, 2, \ldots, n$; t is the current iteration number; V_{id} is the velocity of the particle; P_{id} is the individual optimum; P_{gd} is the global optimum; c_1 and c_2 are the acceleration constants; and r_1 and r_2 are the subjects to a uniform distribution in the (0, 1) interval.

2.3. Improved XGBoost Algorithm Based on PSO. Although XGBoost has excellent results in many aspects, there are many parameters in it and different combinations of parameters determine the performance of the model to a large extent. PSO has the unique advantage of optimizing the parameters of XGBoost, which can effectively improve the effectiveness and accuracy of diagnosing rotor fault causes. In this study, six parameters that have a great influence on

the model are optimized by PSO. The information of each parameter is given in Table 1.

According to Table 1, the velocity vector and the position vector of the i^{th} particle at the t^{th} iteration can be expressed as

$$(V)_i^{(t)} = \left[V_{i,\text{eta}}^{(t)}, V_{i,\text{subsample}}^{(t)}, V_{i,\text{cosample\_bytree}}^{(t)}, V_{i,\text{cossample\_bylevel}}^{(t)}, V_{i,\text{reg\_alpha}}^{(t)}, V_{i,\text{reg\_lambda}}^{(t)}\right], \tag{15}$$

$$(P)_i^{(t)} = \left[P_{i,\text{eta}}^{(t)}, P_{i,\text{subsample}}^{(t)}, P_{i,\text{cosample\_bytree}}^{(t)}, P_{i,\text{cosample\_bylevel}}^{(t)}, P_{i,\text{reg\_alpha}}^{(t)}, P_{i,\text{reg\_lambda}}^{(t)}\right]. \tag{16}$$

The position vector is assigned to the corresponding parameters of XGBoost, and the negative accuracy score of the XGBoost model is used as the fitness value to measure the performance of PSO. The fitness value of the i^{th} particle at the t^{th} iteration is shown as

$$(F)_i^{(t)} = -\frac{1}{N}\sum_{i=1}^{N}[\![\widehat{y}_i = y_i]\!]|_{P_i^{(t)}\longrightarrow \text{XGBoost}}, \tag{17}$$

where $F_i^{(t)}$ represents the negative accuracy score of XGBoost; $[\![\cdot]\!]$ is the indicator function, taking 1 and 0, respectively, when $\cdot$ is true and false; $\widehat{y}_i$ is the prediction label of XGBoost; y_i is the real label of samples; and N is the total number of samples.

The individual optimum of the i^{th} particle at the t^{th} iteration is

$$P_{id}^{(t)} = \min\left(F_i^{(j)}\right), \quad 0 \le j \le t. \tag{18}$$

The global optimum of the i^{th} particle at the t^{th} iteration is

$$P_{gd}^{(t)} = \min\left(P_{kd}^{(t)}\right), \quad 1 \le k \le n, \tag{19}$$

where n is the number of particles.

The XGBoost algorithm and the PSO-XGBoost algorithm are shown in Figure 1. The process of the improved XGBoost algorithm based on PSO is shown in Figure 1(b). Compared with the original XGBoost algorithm shown in Figure 1(a), the PSO-XGBoost final training accuracy score is used as the objective function to search out the optimal parameters. The optimal result can be obtained by running PSO-XGBoost once, while XGBoost needs to be adjusted manually many times, and the optimal result may not be obtained.

The procedures of the proposed method for diagnosis of rotor fault causes, which can be seen from Figure 1(b), are as follows:

Step 1. Initialize the particle swarm. Initialize the particle swarm parameters, including the particle number, learning factors, weighting coefficient, and the maximum number of iterations.

Step 2. Train the XGBoost model. The parameters to be optimized change along with the flying of particles.

Step 3. Calculate and assess the fitness value. The fitness value, originating from the output negative accuracy score of the XGBoost model, is used to evaluate the performance of PSO. A smaller fitness value indicates better performance.

Step 4. Judge the stop condition. Terminate the iteration process and obtain the optimal parameters of the XGBoost model if the number of iterations is reached. Otherwise, proceed to the iterative calculation.

Step 5. Validate the classification model. Use the optimization results to build the XGBoost model and output the results of diagnosing rotor fault causes.

3. Results and Discussion

3.1. Data Description. In this study, 450 sets of operation data related to three kinds of high-pressure rotor faults of a 330 MW unit in a power plant are summarized as example verification. The specifications are given in Table 2. Three kinds of faults are represented by F1, F2, and F3, respectively. They are high-pressure rotor rubbing fault (F1), the mass imbalance fault (F2), and the self-excited oscillation (including oil film half-speed whirl and oil film oscillation) fault (F3) and are taken as objects. C1–C9 represent nine different fault causes. Among them, four causes are leading to rotor rub impact, including rubbing at shaft seal caused by cylinder deformation (C1), rubbing at shaft seal caused by the fast rate of loading up (C2), rubbing at shaft seal caused by a long time of low load remaining (C3), and rotor rubbing with oil baffle (C4); three causes are leading to mass imbalance fault, including inadequate stiffness of bearing pedestal (C5), fracture and falling off of rotating parts (such as blades and coupling windshields) (C6) and other reasons (C7); and two causes are leading to self-excited oscillation fault, including poor stability of bearing (C8) and excessive journal disturbance (C9). A total of 50 groups of data samples for each fault cause constitute the sample set.

In this study, ten running parameters with high correlation with rotor rubbing fault, mass imbalance fault, and self-excited oscillation fault are selected, including the steam temperature of high-pressure cylinder shaft seal and cylinder expansion value of high-pressure cylinder. The details are given in Table 3.

Table 1: Parameters to be optimized.

Parameter	Default value	Range	Explain
Eta	0.3	(0, 1)	Learning rate.
Subsample	1	(0, 1)	Subsample ration of the training instance.
colsample_bytree	1	(0, 1)	Subsample ration of columns when constructing each tree.
colsample_bylevel	1	(0, 1)	Subsample ration of columns for each level.
reg_alpha	0	(0, ∞)	L1 regularization term on weights
reg_lambda	1	(0, ∞)	L2 regularization term on weights.

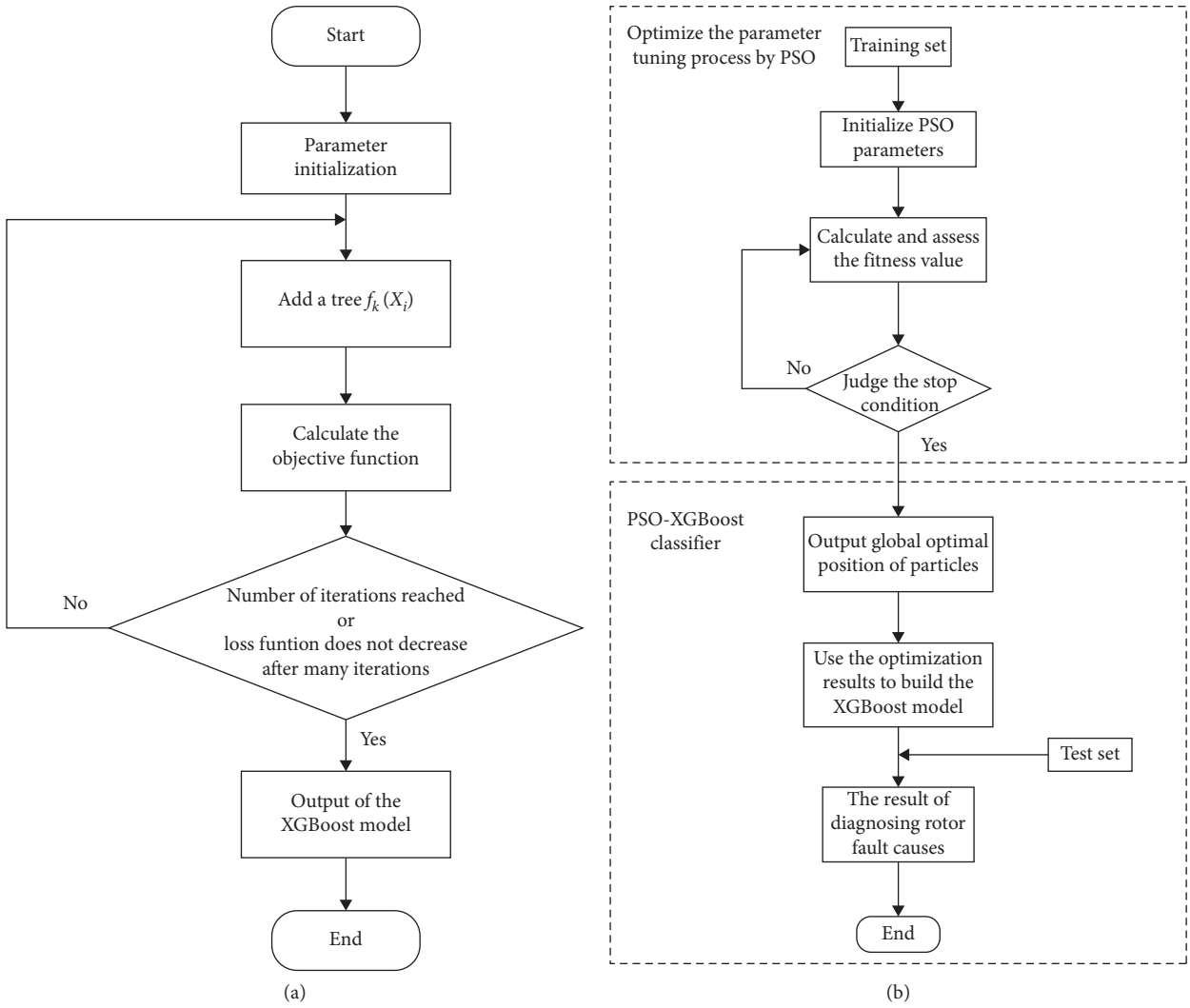

Figure 1: XGBoost and PSO-XGBoost methods. (a) XGBoost algorithm flow chart. (b) The proposed PSO-XGBoost algorithm flow chart.

In all, the input data, composed of running parameters and fault type, have eleven dimensions, i.e., $X_i = [HT1, HT2, HE1, HC1, HT3, HT4, X1, Y1, X2, Y2, F1, F2, F3]$.

3.2. Data Preprocessing. Data preprocessing aims to make the data adapt to the model and match the model's needs. Data preprocessing mainly includes missing value processing, data dimensionless processing (including central processing and scaling processing), classified feature processing (text to digital), and continuous feature processing.

3.2.1. Missing Value Processing. For missing values, in this study, the mean is used to fill the numerical feature, and the mode is used to fill the character feature.

3.2.2. Feature Coding of Character Features. In the original dataset, digits do not represent the fault types in the classification features (rubbing fault (F1), mass imbalance fault (F2), and self-excited oscillation (F3)) and fault cause category labels (rubbing at shaft seal caused by cylinder deformation (C1) and rubbing at shaft seal caused by the fast

Table 2: The specific rotor fault causes.

Number	Fault type	Cause description	Number of samples	Label
1	Rubbing fault (F1)	Rubbing at shaft seal caused by cylinder deformation	50	C1
2		Rubbing at shaft seal caused by the fast rate of loading up	50	C2
3		Rubbing at shaft seal caused by the long time of low load remaining	50	C3
4		Rotor rubbing with oil baffle	50	C4
5		Poor stiffness of bearing pedestal	50	C5
6	Mass imbalance fault (F2)	Fracture and falling off of rotating parts (blades and coupling wind shields)	50	C6
7		Other reasons	50	C7
8	Self-excited oscillation	Poor stability of bearing	50	C8
9	(F3)	Excessive journal disturbance	50	C9

Table 3: Running parameters of a high-pressure steam turbine rotor.

Index	Parameters	Unit	Symbol
1	Steam temperature of high-pressure cylinder shaft seal	℃	HT1
2	Temperature difference between upper and lower cylinders of a high-pressure cylinder	℃	HT2
3	Expansion value of a high-pressure cylinder	mm	HE1
4	The change rate of unit load	Mw/min	HC1
5	High-pressure cylinder temperature	℃	HT3
6	Lubricating oil temperature of high-pressure rotor	℃	HT4
7	Fundamental frequency amplitude of No.1 bearing shaft vibration in X direction of a high-pressure rotor	mm	X1
8	Fundamental frequency amplitude of No.1 bearing shaft vibration in Y direction of a high-pressure rotor	mm	Y1
9	Fundamental frequency amplitude of No.1 bearing pedestal vibration in X direction of a high-pressure rotor	mm	X2
10	Fundamental frequency amplitude of No.1 bearing pedestal vibration in Y direction of a high-pressure rotor	mm	Y2

rate of loading up (C2)). For making the data adapt to the algorithm, the data must be encoded, converting texts to numerical types. The independent fault types (F1, F2, and F3) are transformed into dummy variables by using one-hot coding, namely, F1 [1 0 0], F2 [0 1 0], and F3 [0 0 1]. Labels of fault cause category [C1, C2, . . ., C9] are directly converted into the digital form [0, 1, . . ., 8].

3.2.3. Data Standardization. First, decentralize the data by mean (μ). Then, scale them by standard deviation (σ). The conversion function is given in (10). After the above two steps, the data will follow the standard normal distribution, i.e., $x \sim N(\mu, \sigma^2)$.

$$x^* = \frac{x - \mu}{\sigma}. \tag{20}$$

The preprocessed dataset is given in Table 4.

3.3. Experimental Results. The test set is used to verify the performance of the PSO-XGBoost model. The model is quantitatively evaluated using evaluation indicators, such as the accuracy, confusion matrix, precision, recall, and F1-score [31–33].

The results can be divided into four classes, including true positive (TP), false positive (FP), true negative (TN), and false negative (FN). Here, TP is the correct predicted positive category, FP is the incorrect predicted positive category, TN is the correct predicted negative category, and FN is the incorrect predicted negative category.

Accuracy is simply a ratio of the correctly predicted classifications to the total dataset. In formula, the accuracy ratio is $A = (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN})$.

Precision is the ratio of the system generated results that TP to the system's total predicted positive observations, both TP and FP. In formula, the precision ration is $P = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$.

Recall is the ratio of the system generated results that TP to all categories in the actual class. In formula, the recall ratio is $R = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$.

F1-score is the weighted average of precision and recall, and the calculation formula is $F1 = 2 \times P \times R/(P + R)$.

The confusion matrix is used for evaluating the model when faced with a multiclassification problem. Each column of the confusion matrix represents a predicted category, and the total numbers of data for each column represent the number of data predicted to be in the category. Each row represents the data's actual category, and the total numbers of data for each row represent the number of data instances belonging to that category. For a confusion matrix, the larger the value on the diagonal is, the better the matrix. The smaller the value on other locations is, the better the matrix.

The result is shown in Figure 2, and PSO-XGBoost's confusion matrix is shown in Figure 3.

Figure 2 shows that the overall accuracy of the PSO-XGBoost model is 98.52%. From Figure 3, the accuracy of rubbing fault caused by cylinder deformation is 92.86%, the accuracy of rubbing at shaft seal caused by the fast rate of loading up is 92.31%, and the accuracy of three faults caused by other reasons is 100%.

Table 4: Preprocessed dataset.

Number	HT1	HT2	HE1	...	F1	F2	F3	Label
1	−0.31395	−1.0698	−0.14448	...	−0.89442719	1.414213562	−0.534522484	6
2	−1.33696	−0.97908	−0.94812	...	−0.89442719	−0.70710678	1.870828693	7
3	−0.41837	−1.61409	−0.5463	...	−0.89442719	−0.70710678	1.870828693	8
...	...	...	...	...	...	...	...	...
449	1.05843	1.440195	0.900258	...	1.118033989	-0.70710678	-0.534522484	0
450	1.045868	1.149048	-0.38557	...	1.118033989	-0.70710678	-0.534522484	0

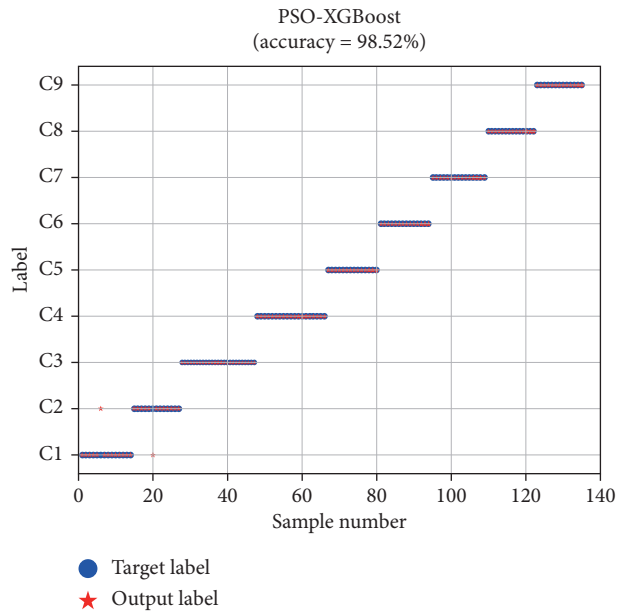

Figure 2: The accuracy of the PSO-XGBoost model.

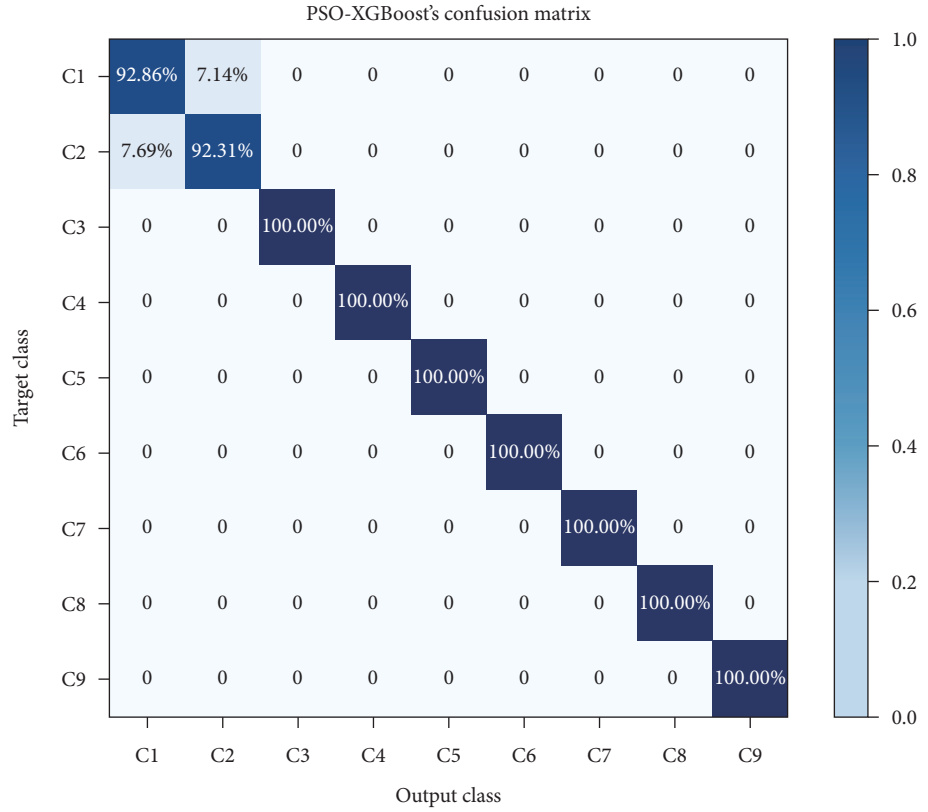

Figure 3: The confusion matrix of the PSO-XGBoost model.

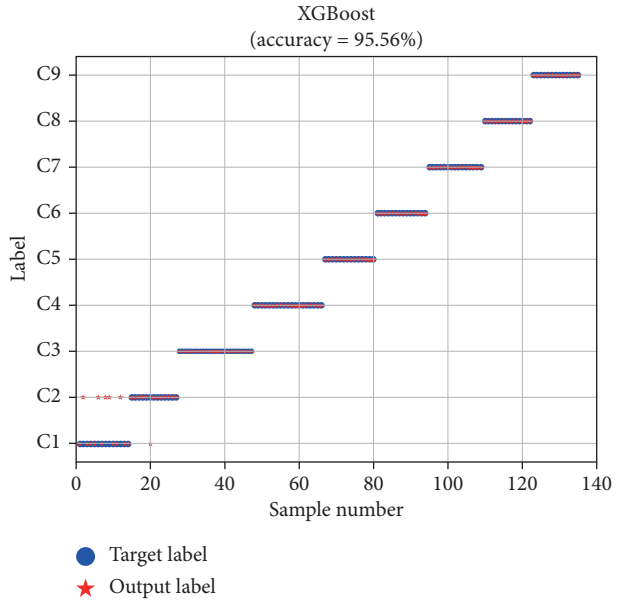

Figure 4: The classification result of the XGBoostmodel.

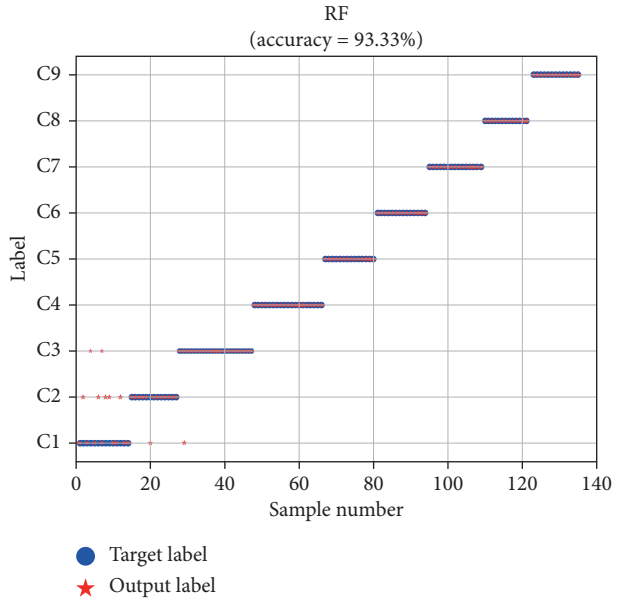

Figure 5: The classification result of the RF model.

Table 5 provides the accuracy, precision, recall, and F1-score for the PSO-XGBoost model. From this table, it can be seen that the accuracy, precision, recall, and F1-score of the proposed method as a whole are all above 98% for the performance of diagnosing rotor fault causes, and it can perform the accurate and comprehensive identification of various categories. Therefore, the proposed method's performance has good results in accuracy, precision, recall, and F1-score.

3.4. Comparative Analysis. An investigation of five different classifiers is performed to verify the superiority of PSO-XGBoost in classification performance, including XGBoost, RF, GBDT, DT, and SVM. The classification results of these algorithms are shown in Figures 4–8 . The results of accuracy are 95.56%, 93.33%, 92.59%, 91.85%, and 84.44%, respectively. Compared with Figure 3, we can conclude that the PSO-XGBoost algorithm is superior to the other five algorithms in classification accuracy.

To have a detailed quantitative analysis related to each classifier's classification results, five confusion matrixes according to five studied classification experiments are introduced for recording the recognition results and the percentage of misclassification of the rotor with different

Table 5: PSO-XGBoost's evaluation indicators.

Accuracy (%)	Precision	Recall	F1-score
98.52	98.52	98.52	98.52

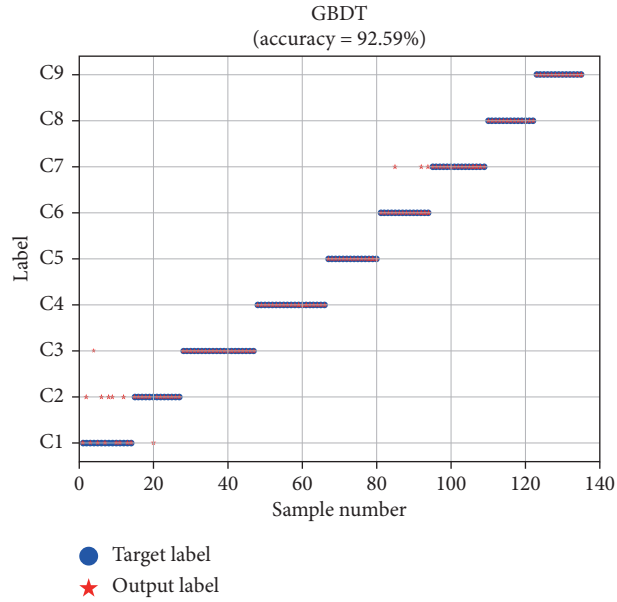

Figure 6: The classification result of the GBDT model.

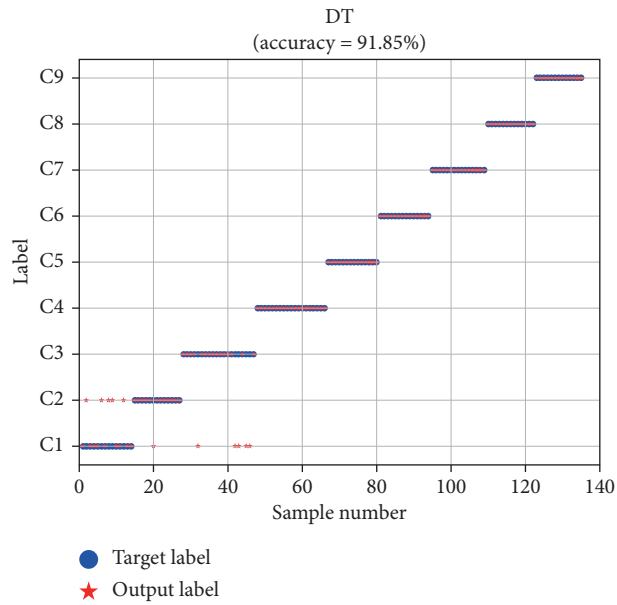

Figure 7: The classification result of the DT model.

fault causes. Figures 9–13 show the confusion matrixes of XGBoost, RF, GBDT, DT, and SVM, respectively.

Figures 10–12 show that the RF model and the DT model are confused with C1, C2, and C3, and the GBDT model is confused with C1, C2, C3, C6, and C7. Figure 13 shows that the SVM model plays the worst performance. Figures 3 and 9 show that the PSO-XGBoost model and the XGBoost model are all confused with C1 and C2, but in category C1, PSO-XGBoost has higher accuracy than XGBoost. Therefore, the PSO-XGBoost model is superior to the other five algorithms. The comprehensive model evaluation indicators are given in Table 6.

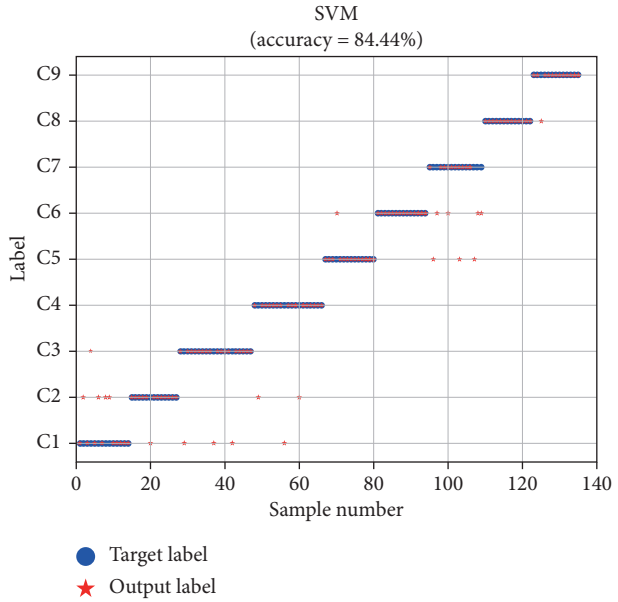

Figure 8: The classification result of the SVM model.

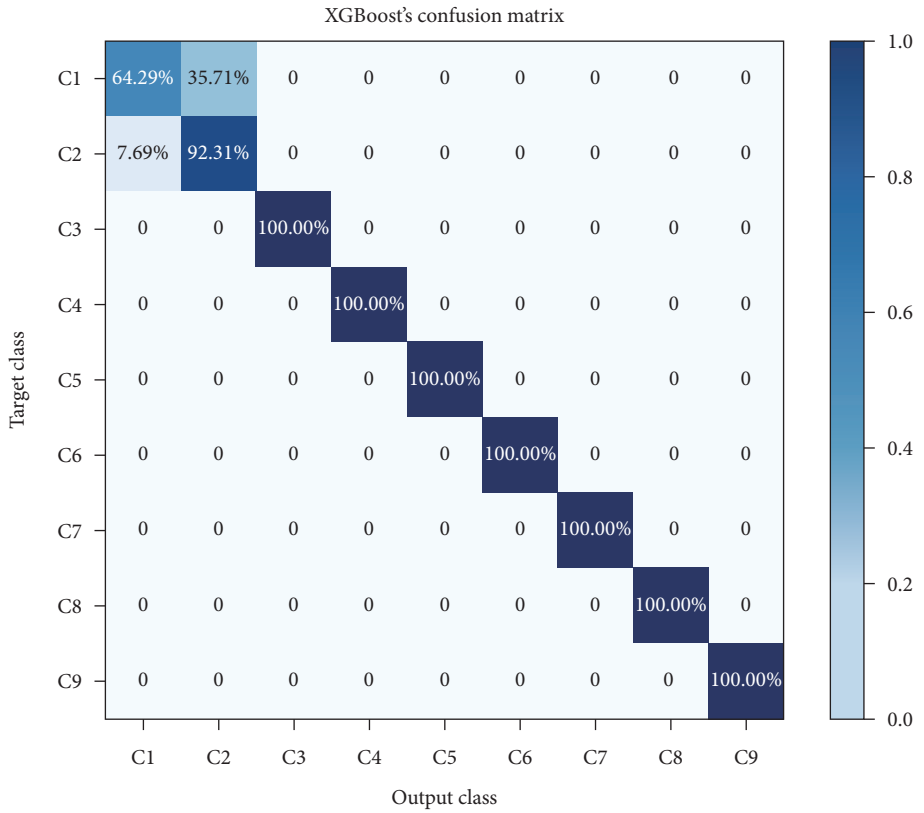

Figure 9: The confusion matrix of the XGBoost model.

The detailed comparison of six algorithms in the accuracy, precision, recall, and F1-score is shown in Figures 14–17.

In the view of Figures 14–17, the SVM model's accuracy, precision, recall, and F1-score are the lowest of the five algorithms because its principle is a linear classifier based on maximum interval, which does not work well in solving nonlinear problems. DT's performance is better than SVM, but its accuracy, precision, recall, and F1-score are slightly lower than other four algorithms because RF, GBDT, and XGBoost all use the DT model as their weak learning models. Except PSO-XGBoost, the performance fault causes the

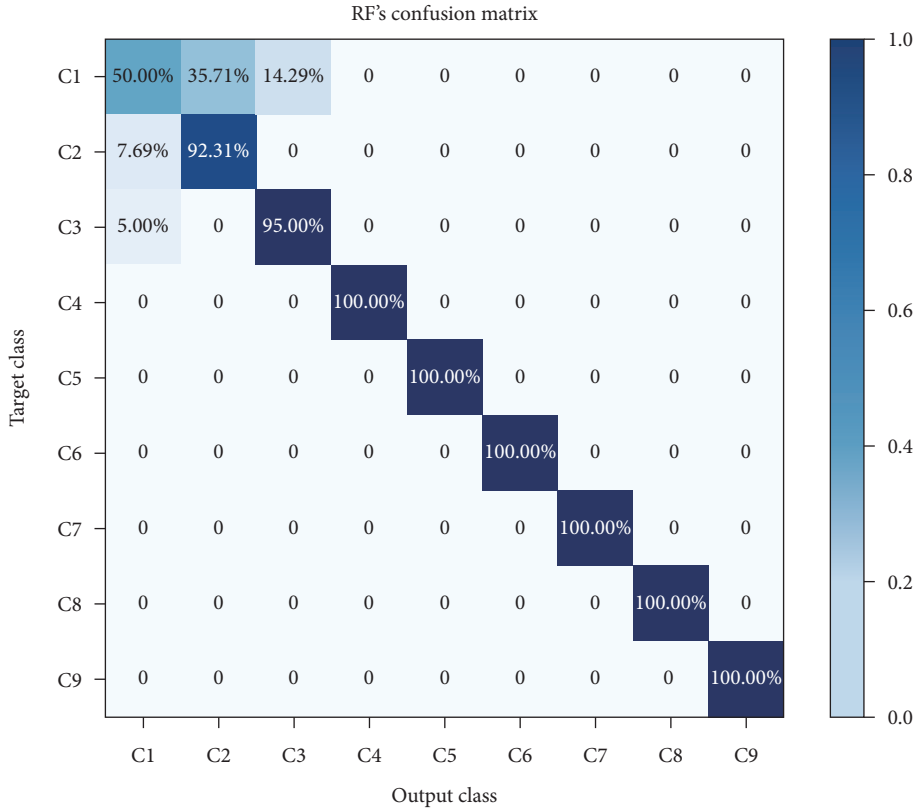

Figure 10: The confusion matrix of the RF model.

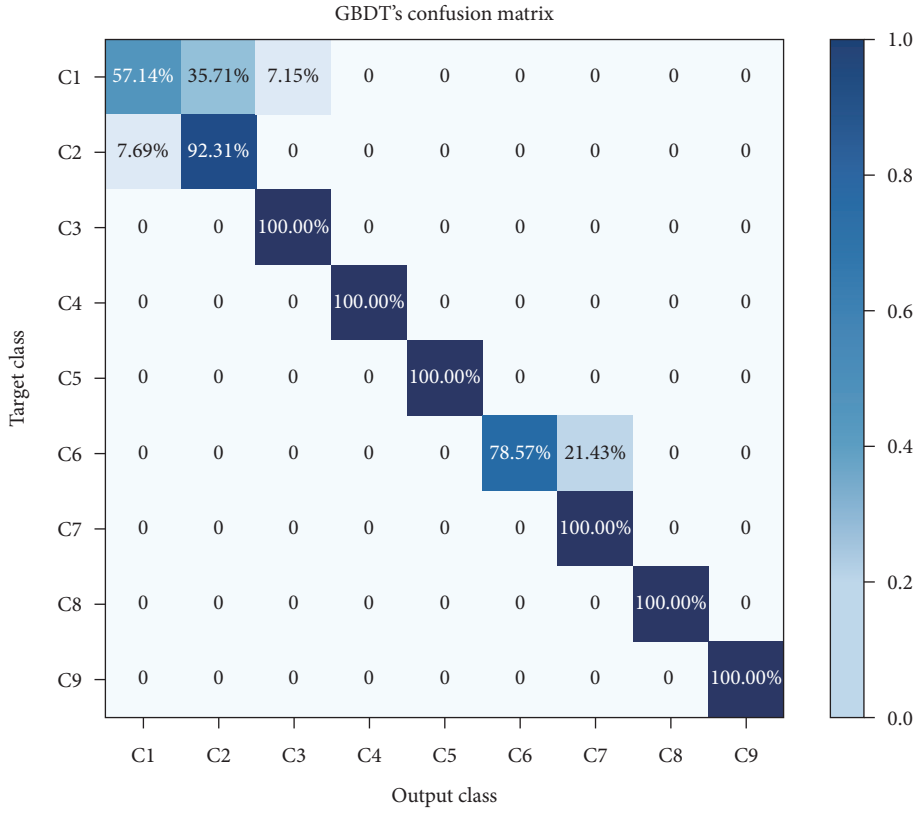

Figure 11: The confusion matrix of the GBDT model.

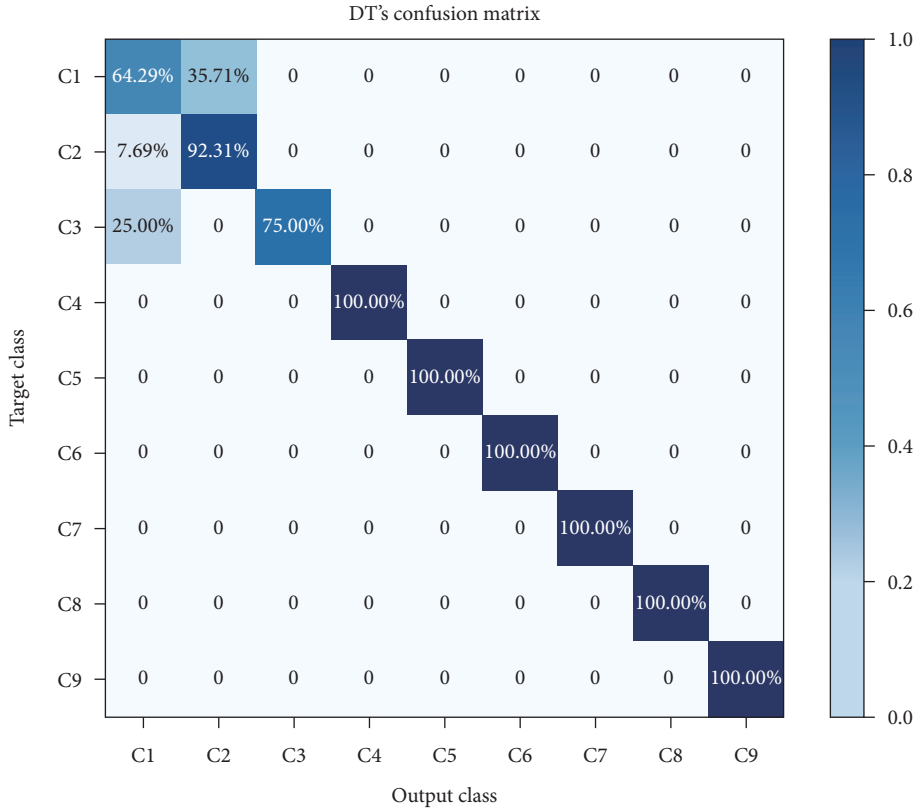

Figure 12: The confusion matrix of the DT model.

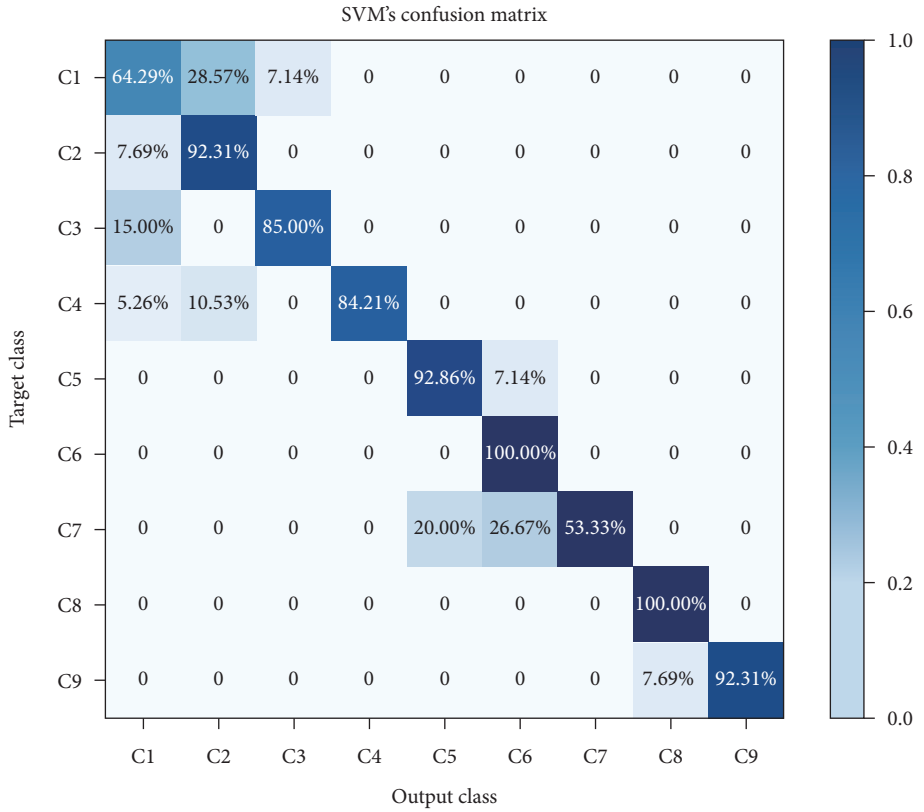

Figure 13: The confusion matrix of the SVM model.

Table 6: Comprehensive model evaluation indicators.

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
PSO-XGBoost	98.52	98.52	98.52	98.52
XGBoost	95.56	96.13	95.56	95.48
RF	93.33	93.45	93.33	92.93
GBDT	92.59	93.46	92.59	92.30
DT	91.85	93.02	91.85	92.02
SVM	84.44	86.90	84.44	84.28

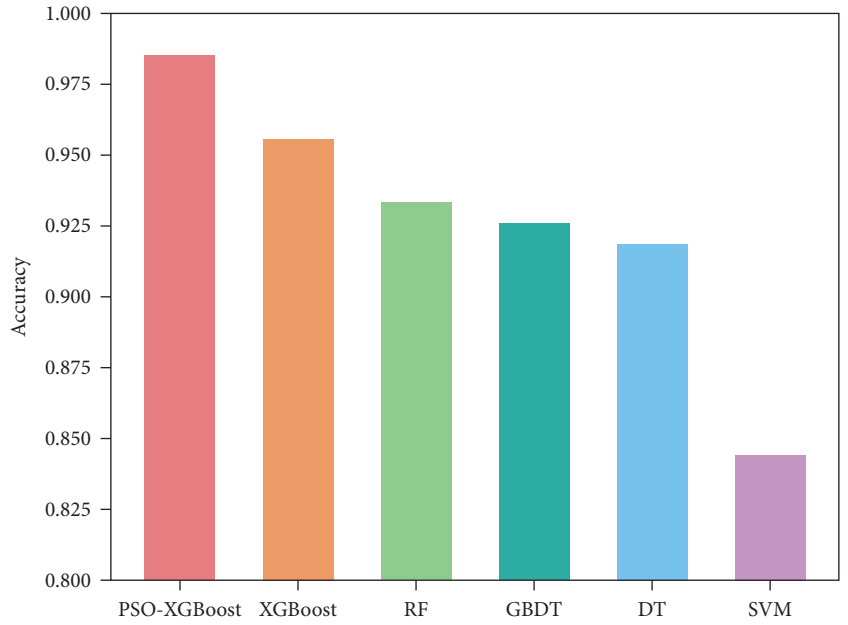

Figure 14: The accuracy of six algorithms.

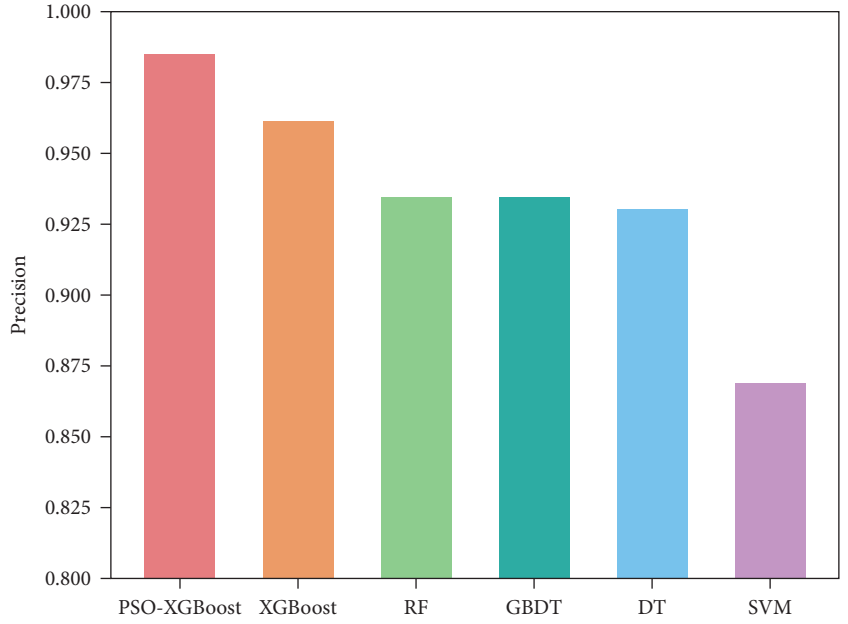

Figure 15: The precision of six algorithms.

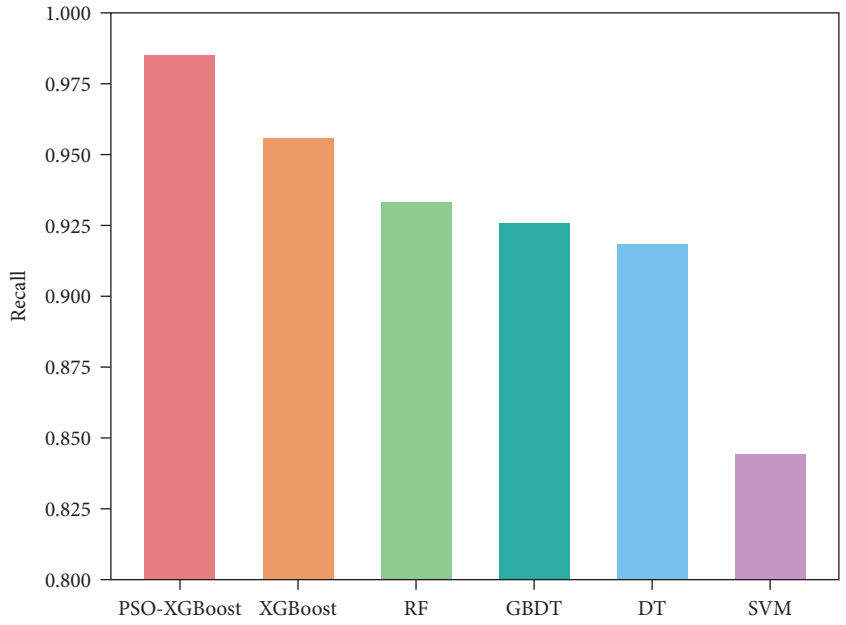

Figure 16: The recall of six algorithms.

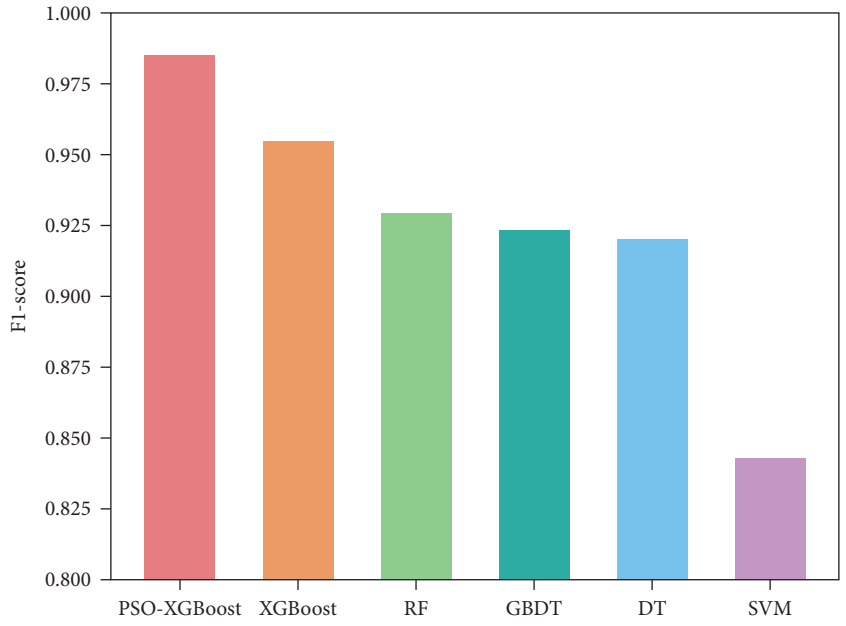

Figure 17: The F1-score of six algorithms.

diagnosis model constructed using XGBoost superior to the other four algorithms in accuracy, precision, recall, and F1-score because XGBoost uses second-order derivatives and regularization terms, which improves the accuracy and is not affected by the size of the dataset. After optimizing the parameter tuning process during training an XGBoost model by using PSO, the PSO-XGBoost model's accuracy, precision, recall, and F1-score are the highest of the five algorithms. Evidently, PSO can effectively optimize the parameters of XGBoost, thereby improving the classification performance on the dataset. From the model's comprehensive classification performance perspective, choosing the PSO-XGBoost model for diagnosing rotor fault causes is more reasonable than other algorithms.

The comparison of different algorithms in the iterative process is shown in Figure 18.

From Figure 18, in the initial iteration stage, for the proposed method, the iterative curve shows a rapid downward trend; then, the iterative process in the proposed method is easily converged. Obviously, other five methods

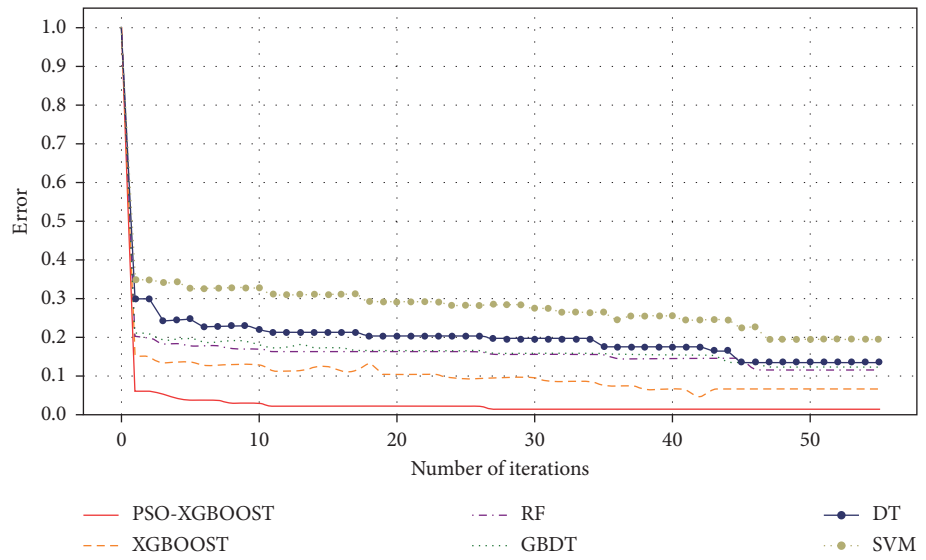

Figure 18: The comparison of different algorithms in the iterative process.

Table 7: Knowledge base for rubbing fault, mass imbalance fault, and self-excited oscillation of a high-pressure steam turbine rotor.

Fault type	Fault cause		Fault measure	
	Fault cause label	Description	Fault measure label	Solution
Rubbing fault	C1	Rubbing at shaft seal caused by cylinder deformation	M1	Adjust the clearance of vertical pin and thrust bearing and tighten the bolt of valve screw
	C2	Rubbing at shaft seal caused by the fast rate of pupinization	M2	Reduce the rate of pupinization
	C3	Rubbing at shaft seal caused by the long time of low load remaining	M3	Reduce residence time under low load and increase load as soon as possible
	C4	Rotor rubbing with oil baffle	M4	Adjust the dynamic and static clearance and control the thermal parameters in the start-up operation
Mass imbalance fault	C5	Poor stiffness of bearing pedestal	M5	Reduce the excitation force by rotor dynamic balance
	C6	Fracture and falling off of rotating parts (blades and coupling wind shields)	M6	Deal with the wind deflector or replace high quality hexagon bolts
	C7	Other reasons	M7	Carry out first- or second-order dynamic balance
Self-excited oscillation	C8	Poor stability of bearing	M8	Use bearings with good stability such as tilting pad and elliptical bush Increase the bearing specific pressure such as reducing the bearing length and adjusting the bearing height Increase the temperature of lubricating oil and reduce the viscosity of lubricating oil Reduce the top clearance of the fixed pad bearing and improve the bearing preload
	C9	Excessive journal disturbance	M9	Reduce the vibration of the shaft and the disturbing force of the journal

still need several iterations to achieve the final convergence result. Therefore, the proposed method is faster in convergence rate and has higher efficiency in practice.

3.5. Maintenance Strategy according to Fault Causes. For nine different rotor fault causes, we build a knowledge base, mapping each rotor fault cause to a specific solution, in order

to achieve the purpose of intelligent operation and maintenance. For example, when we diagnose the rotor fault cause C1, the computer will automatically link to the solution M1. Other details in the knowledge base are given in Table 7.

4. Conclusions

On basis of fault categories detection, the diagnosis of rotor fault causes is proposed, which has great contributions to the field of intelligent operation and maintenance. This study proposes a hybrid model for diagnosing rotor fault causes using the PSO-XGBoost algorithm. Aiming at the problems of low accuracy and low efficiency in using empirical methods to adjust parameters of the XGBoost model, PSO is used to solve the difficulty of parameter adjustment when using the XGBoost model to diagnose rotor fault causes and improve the diagnostic accuracy at the same time. The experimental results show that

(1) Compared with the direct construction of the XGBoost model to diagnose rotor fault causes, the hybrid model can achieve higher diagnostic accuracy and practical efficiency

(2) The hybrid model can effectively identify nine different failure causes under three types of failures, and the classification accuracy, precision, recall, and F1-score are all above 98%. Compared with XGBoost, RF, GBDT, DT, and SVM, from the perspective of the PSO-XGBoost's comprehensive classification performance, choosing the PSO-XGBoost model in diagnosing rotor fault causes is more effective than other algorithms.

Data Availability

The csv data used to support the findings of this study have been deposited in the Baidu Netdisk repository (https://pan.baidu.com/s/1A8jqMmykRYbOxqJwYPC3fw; password: suep).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by Shanghai 2019 "Science and Technology Innovation Action Plan'' High-tech Field Project" (19511103700).

References

[1] H. Shao, H. Jiang, F. Wang, and H. Zhao, "An enhancement deep feature fusion method for rotating machinery fault diagnosis," *Knowledge-based Systems*, vol. 119, pp. 200–220, 2017.

[2] H. Zhao, J. Zheng, J. Xu, and W. Deng, "fault diagnosis method based on principal component analysis and broad learning system," *IEEE Access*, vol. 7, pp. 99263–99272, 2019.

[3] X. Wu, Z. Peng, J. Ren, C. Cheng, W. Zhang, and D. Wang, "Rub-impact fault diagnosis of rotating machinery based on 1-D convolutional neural networks," *IEEE Sensors Journal*, vol. 20, no. 15, pp. 8349–8363, 2020.

[4] J. Zheng, H. Pan, S. Yang, and J. Cheng, "Adaptive parameterless empirical wavelet transform based time-frequency analysis method and its application to rotor rubbing fault diagnosis," *Signal Processing*, vol. 130, pp. 305–314, 2017.

[5] X. Zhu, D. Hou, P. Zhou et al., "Rotor fault diagnosis using a convolutional neural network with symmetrized dot pattern images," *Measurement*, vol. 138, pp. 526–535, 2019.

[6] S. Pang, X. Yang, X. Zhang, and X. Lin, "Fault diagnosis of rotating machinery with ensemble kernel extreme learning machine based on fused multi-domain features," *ISA Transactions*, vol. 98, pp. 320–337, 2020.

[7] J.-D. Wu, M. R. Bai, F.-C. Su, and C.-W. Huang, "An expert system for the diagnosis of faults in rotating machinery using adaptive order-tracking algorithm," *Expert Systems with Applications*, vol. 36, no. 3, pp. 5424–5431, 2009.

[8] Z. Sun and H. Sun, "Health status assessment for wind turbine with recurrent neural networks," *Mathematical Problems in Engineering*, vol. 2018, Article ID 6972481, 16 pages, 2018.

[9] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[10] S. U. Jan, Y.-D. Lee, J. Shin, and I. Koo, "Sensor fault classification based on support vector machine and statistical time-domain features," *IEEE Access*, vol. 5, pp. 8682–8690, 2017.

[11] T. H. G. Lobato, R. R. Da Silva, E. S. Da Costa, and A. L. A. Mesquita, "An integrated approach to rotating machinery fault diagnosis using, EEMD, SVM, and augmented data," *Journal of Vibration Engineering & Technologies*, vol. 8, no. 3, pp. 403–408, 2020.

[12] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[13] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[14] Z. Wang, Q. Zhang, J. Xiong, M. Xiao, G. Sun, and J. He, "fault diagnosis of a rolling bearing using wavelet packet denoising and random forests," *IEEE Sensors Journal*, vol. 17, no. 17, pp. 5581–5588, 2017.

[15] J. C. Quiroz, N. Mariun, M. R. Mehrjou, M. Izadi, N. Misron, and M. A. Mohd Radzi, "Fault detection of broken rotor bar in LS-PMSM using random forests," *Measurement*, vol. 116, pp. 273–280, 2018.

[16] K. Zhu, S. Ying, N. Zhang et al., "A performance fault diagnosis method for SaaS software based on GBDT algorithm," *Computers, Materials & Continua*, vol. 62, no. 3, pp. 1161–1185, 2020.

[17] H. M. Zhong, W. L. Zhang, Y. R. Li et al., "GBDT based railway accident type prediction and cause analysis," *Acta Automatica Sinica*, vol. 45, pp. 1–9, 2020.

[18] TQ. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system,," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, San Francisco, CA, August 2016.

[19] D. Zhang, L. Qian, B. Mao, C. Huang, B. Huang, and Y. Si, "A data-driven design for fault detection of wind turbines using random forests and XGboost," *IEEE Access*, vol. 6, pp. 21020–21031, 2018.

[20] Y. Lei, W. Jiang, A. Jiang, Y. Zhu, H. Niu, and S. Zhang, "Fault diagnosis method for hydraulic directional valves integrating PCA and XGBoost," *Processes*, vol. 7, no. 9, p. 589, 2019.

[21] Z. Wu, X. Wang, and B. Jiang, "fault diagnosis for wind turbines based on ReliefF and eXtreme gradient boosting," *Applied Sciences*, vol. 10, no. 9, p. 3258, 2020.
[22] Y. Jiang, G. Tong, H. Yin, and N. Xiong, "A pedestrian detection method based on genetic algorithm for optimize XGBoost training parameters," *IEEE Access*, vol. 7, pp. 118310–118321, 2019.
[23] X. Cai, H. Zhao, S. Shang et al., "An improved quantum-inspired cooperative co-evolution algorithm with muli-strategy and its application," *Expert Systems with Applications*, vol. 171, Article ID 114629, 2021.
[24] 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, 2021.
[25] W. Deng, S. Shang, X. Cai, H. Zhao, Y. Song, and J. Xu, "An improved differential evolution algorithm and its application in optimization problem," *Soft Computing*, vol. 25, no. 7, pp. 5277–5298, 2021.
[26] W. Deng, J. Xu, H. Zhao, and Y. Song, "A novel gate resource allocation method using improved PSO-based QEA," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, pp. 1–9, 2020.
[27] W. Deng, J. Xu, X.-Z. Gao, and H. Zhao, "An enhanced MSIQDE algorithm with novel multiple strategies for global optimization problems," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 22, pp. 1–10, 2020.
[28] Y. Song, D. Wu, A. Zhou, B. Zhang, and W. Deng, "Enhanced success history adaptive DE for parameter optimization of photovoltaic models," *Complexity*, vol. 2021, Article ID 6660115, 22 pages, 2021.
[29] B. Wang, Y. Sun, B. Xue et al., "Evolving deep convolutional neural networks by variable-length particle swarm optimization for image classification," in *Proceedings of IEEE Congress on Evolutionary Computation*, pp. 1514–1521, Rio de Janeiro, Brazil, July 2018.
[30] X. Li, Y. Guo, and Y. Li, "Particle swarm optimization-based SVM for classification of cable surface defects of the cable-stayed bridges," *IEEE Access*, vol. 8, pp. 44485–44492, 2020.
[31] M. Zhang, Z. Liu, and X. Dang, "fault diagnosis on train brake system based on multi-dimensional feature fusion and GBDT enhanced classification," in *Proceedings Of the International Conference On Intelligent Rail Transportation,*, Singapore, December 2018.
[32] I. I. E. Amarouayache, M. N. Saadi, N. Guersi et al., "Bearing fault diagnostics using EEMD processing and convolutional neural network methods," *International Journal Of Advanced Manufacturing Technology*, vol. 107, no. 9-10, pp. 4077–4095, 2020.
[33] XF. Wang, XB. Yan, and YC. Ma, "Research on user consumption behavior prediction based on improved XGBoost algorithm," in *Proceedings Of IEEE International Conference On Big Data*, pp. 4169–4175, Seattle, WA, December 2018.