Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 2149048, 10 pages
https://doi.org/10.1155/2021/2149048

Research Article

Compound Fault Diagnosis of Stator Interturn Short Circuit and Air Gap Eccentricity Based on Random Forest and XGBoost

Rui Tian, Fuyang Chen, and Shiyi Dong

College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China

Correspondence should be addressed to Fuyang Chen; chenfuyang@nuaa.edu.cn

Received 18 June 2021; Revised 18 August 2021; Accepted 7 September 2021; Published 9 October 2021

Academic Editor: Juan P. Amezquita-Sanchez

Taking the traction motor of CRH2 high-speed train as the research object, this paper proposes a diagnosis method based on random forest and XGBoost for the compound fault resulting from stator interturn short circuit and air gap eccentricity. First, the U-phase and V-phase currents are used as fault diagnosis signal and then the Savitzky–Golay filtering method is used for the noise deduction from the signal. Second, the wavelet packet decomposition is used to extract the composite fault features and then the high-dimensional features are optimized by the principal component analysis (PCA) method. Finally, the random forest and XGBoost are combined to detect composite faults. Using the experimental data of CRH2 semiphysical simulation platform, the diagnosis of different fault modes is completed, and the high diagnosis accuracy is achieved, which verifies the validity of this method.

1. Introduction

Traction motor is one of the key components of the drive system in CRH2 high-speed train [1]. Because of the harsh working environment and its special structure, the traction motor is prone to failure. The short circuit between stator windings is a common fault in the motor [2]. Static air gap eccentricity exists more or less in the motors in engineering practice, so when the motor stator windings are short-circuited, it is equivalent to the compound fault of stator interturn short circuit and air gap eccentricity. The traction motor of high-speed railway is installed on the bogie of the train. Because of working for a long time in a harsh environment, the mechanical wear of the motor will destroy the symmetry of air gap magnetic field and cause air gap eccentricity [3]. The copper chip in the motor is easy to damage the insulation layer of stator windings, resulting in short connection between windings [4]. If not diagnosed in time, the fault will lead to the further expansion of insulation layer damage [5], which will sequentially increase the number of short circuit turns and cause more serious distortion of the air gap magnetic field [6].

In the domain of the engineering, the industries' circumstances varied. For the analysis of the complex mechanical components' fault under various nonlinear responses, Keshtegar et al. [7] proposed the Modified Response Surface basis Models for failure turbine blisk response which also appears to be multiphased, where it includes two regression processes for the purpose of regressing the input variables and calibration more precisely. Moreover, in the traction motor domain, the induction traction motor health diagnostics have been reviewed in [8]. Also, the new merged techniques such as deep learning [9] and transfer learning [10], where the detection knowledge could be learned from another domain, are quite useful when the data are deficient. Also, many genetic algorithms have also been studied and proposed where the method can handle various fault types [11, 12] using multiobjective optimization methods.

Motivated by these observations, a multiphase diagnosis method based on random forest and XGBoost for the compound fault of stator interturn short circuit and air gap eccentricity is proposed in this paper. Compared with the existing results, the main contributions of this paper are threefold:

(1) SG filtering is used for signal denoising pretreatment. The wavelet packet decomposition is used to extract the detailed information of each frequency band of the current signal to form the fault feature vector.

(2) Considering the problem that the dimension of feature vector is too high, the PCA method is used to reduce the dimension of feature signal and eliminate the unimportant features.

(3) The feature vectors after dimensionality reduction are used to train the random forest classifier, and the most important features are selected to train the XGBoost classifier, which improves the prediction accuracy and generalization performance of the classification model. The trained classification model is used to identify different fault modes, and the better result of compound fault diagnosis is obtained. The diagnostic flowchart is shown in Figure 1.

The remainder of this paper is summarized as follows. Section 2 proposes an SG filtering method to preprocess the three-phase current of the motor. Section 3 presents the signal feature extraction and optimization based on wavelet packet and PCA. Fault diagnosis based on random forest and XGBoost is provided in Section 4. Section 5 introduces CRH2 fault injection simulation platform. The experimental results and analysis are provided in Section 6. Section 7 concludes the paper.

2. Signal Noise Reduction Based on SG Filtering

The traction system of high-speed train works in tough environment, and the current signal is bound to be affected by the noise [12]. Savitzky–Golay filtering (SG) is one of the commonly used noise reduction pretreatment methods in signal analysis, which can improve the signal smoothness and reduce the impact of noise [13]. The SG filtering method is an improvement of the moving smoothing algorithm, which reduces the impact on useful information of the signal in the process [14]. According to the average trend in the signal time domain, the suitable filtering parameters are selected, and the least squares fitting in the set sliding time window is realized by polynomial [15]. By changing the window width, the SG filter is applied to the noise reduction smoothing of three-phase current signal to reduce the noise interference and facilitate the extraction of fault features in frequency domain [16, 17].

For a sequence of signals $x[n]$, set the size of SG filter window to $2M + 1$ while the center is $n = 0$. The following polynomial is used to fit this set of data points:

$$p(n) = \sum_{k=0}^{N} a_k n^k. \tag{1}$$

According to the fitting polynomial, the error between the fitting curve and the original data curve can be obtained as follows:

$$\varepsilon_N = \sum_{n=-M}^{M} (p(n) - x[n])^2 = \sum_{n=-M}^{M} \left(\sum_{k=0}^{N} a_k n^k - x[n] \right)^2. \tag{2}$$

When $N = 2$ and $M = 2$, the result of the filter is $y(0) = p(0) = a_0$. Constant terms of $p(n)$ are required. The SG filtering method uses convolution operation to perform an FIR filter on the original data to obtain the constant term, that is, to carry out weighted average operation on the data as shown in the following formula [18]:

$$y(n) = \sum_{m=-M}^{M} h[m]x[n-m] = \sum_{m=n-M}^{n+M} h[n-m]x[m]. \tag{3}$$

When the fitting residual of the least squares fitting curve is the smallest, its partial derivative with respect to each parameter is zero [19], as shown in the following formula:

$$\frac{\partial \varepsilon_N}{\partial a_i} = \sum_{n=-M}^{M} 2n^i \left(\sum_{k=0}^{N} a_k n^k - x[n] \right) = 0. \tag{4}$$

Auxiliary matrices A and B are introduced as follows:

$$\begin{aligned} &A = \left\{ a_{n,i} \right\}, a_{n,i} = n^i, \quad -M \le n \le M, 0 \le i \le N, \\ &B = A^T A. \end{aligned} \tag{5}$$

According to the above definition, the following formula can be deduced:

$$b_{i,k} = \sum_{n=-M}^{M} a_{i,k} a_{n,k} = \sum_{n=-M}^{M} n^{i+k} = b_{k,i}. \tag{6}$$

Then, matrices x and a are built as follows:

$$\begin{aligned} x &= \begin{bmatrix} x[-M] \\ \vdots \\ x[M] \end{bmatrix}, \\ a &= \begin{bmatrix} a[0] \\ \vdots \\ a[N] \end{bmatrix}. \end{aligned} \tag{7}$$

Therefore, we can obtain

$$\begin{aligned} Ba &= A^T A a = A^T x, \\ a &= \left(A^T A \right)^{-1} A^T x = Hx. \end{aligned} \tag{8}$$

In formula (8), the convolution coefficient of the first row of the vector in the H matrix is extremely required. According to the expression of the H matrix, it is only determined by the highest coefficient of the least square polynomial and the size of the filter window, which is independent of the original data.

The SG filtering method is used to preprocess the three-phase current of the motor [20], which can effectively reduce noise interference and enhance the discrimination between fault signals and normal signals in the frequency domain, so that the accuracy of feature extraction of compound fault can be improved.

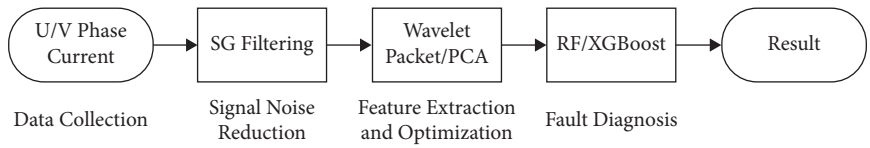

Figure 1: Fault diagnosis flowchart.

3. Signal Feature Extraction and Optimization Based on Wavelet Packet and PCA

In the process of wavelet packet transformation, a series of filters with the same bandwidth but different center frequencies are used to filter the signal, and the signal is decomposed into several layers, so as to analyze its details [21]. Wavelet packet transform can decompose the low-frequency and high-frequency components of the signal for many times at the same time step, which greatly improves the resolution and enables better local time-frequency analysis of the signal without redundancy and omission [22].

In this paper, the stator current after noise reduction is decomposed by wavelet packet, and the wavelet packet coefficient of the final layer is obtained. When the signal is reconstructed according to the wavelet packet coefficient, the energy of each node is defined as the two norm values of the wavelet packet coefficient of the node [23]:

$$E_i = \left|d_{mi}\right|^2, \quad i = 0, 1, \ldots, 2^m - 1, \tag{9}$$

where m is the number of the levels of decomposition and d_{mi} is the wavelet packet coefficient of the last layer.

The wavelet packet energy feature vector is extracted using the following formula. The elements represent the percentage value of the energy of each node and the total energy of all nodes, respectively.

$$X = \left[X_{m1}, X_{m2}, \ldots, X_{m(2^m-1)}\right]. \tag{10}$$

The compound fault feature vector obtained above has a high dimension. If it is directly used as the input sample of the classifier, the complexity of the classifier model will be increased and many abnormal features will be learned, resulting in the phenomenon of overfitting and the poor generalization performance of the new test data [24]. Therefore, the PCA is used to conduct dimensionality reduction preprocessing of feature vectors. On the basis of preserving the original information of data as much as possible, the correlation between features is analyzed and independent principal component features are selected, so that unimportant features such as noise are removed to enhance the generalization performance of the classifier.

Principle component analysis (PCA) is often used for dimensionality reduction compression of data [25]. Its core idea is to map the n-dimensional features to k-dimensional space [26]. The two orthogonal features of this dimension are called principal components, which are built on the basis of the original dimensional features [27]. In order to keep the information of the original data as much as possible in the new dataset, the PCA algorithm needs to find a set of dimensional basis vectors, so that the new data points generated when a feature of the original data is projected on the basis vector can be scattered as far as possible, that is, they have a large variance [28]. At the same time, the basis vectors must be orthogonal to each other to ensure a small degree of coupling [29]. In general, covariance is used to measure the linear correlation of feature's projection on different basis vectors.

In practice, the covariance matrix of dataset is the basis of PCA. For n-dimensional random variables $Q(x_1, x_2, \ldots, x_n)$, the sample collection is $x_{ij} = [x_{1j}, x_{2j}, \ldots, x_{nj}]\,(j = 1, 2, \ldots, m)$, where m is the number of samples. The covariance between the two characteristic dimensions of a and b can be expressed as follows:

$$\mathrm{Cov}\left(x_a, x_b\right) = \frac{\sum_{j=1}^{m}\left(x_{aj} - \overline{x}_a\right)\left(x_{bj} - \overline{x}_b\right)}{m - 1}, \tag{11}$$

where $\overline{x}_a$ and $\overline{x}_b$ are the mean values of the samples. Covariance is a measure of the degree of correlation between two variables.When the covariance is positive, it indicates the positive correlation between the two samples. When the covariance is negative, the two samples are negatively correlated. When the covariance is 0, the two samples are independent of each other. For n-dimensional samples, their covariance is actually a symmetric covariance matrix [30]. For example, the covariance matrix of 3-dimensional data $A = (x, y, z)$ can be expressed as

$$\mathrm{Cov}(A) = \begin{bmatrix} \mathrm{Cov}(x,x) & \mathrm{Cov}(x,y) & \mathrm{Cov}(x,z) \\ \mathrm{Cov}(y,x) & \mathrm{Cov}(y,y) & \mathrm{Cov}(y,z) \\ \mathrm{Cov}(z,x) & \mathrm{Cov}(z,y) & \mathrm{Cov}(z,z) \end{bmatrix}. \tag{12}$$

The elements on the main diagonal of the covariance matrix represent the variance on each feature dimension of the data, while the elements off the main diagonal represent the covariance between the two feature dimensions [31]. If the correlation between different features of the data is smaller, the value on the nonmain diagonal is smaller. When different features are not correlated with each other, the covariance matrix becomes a diagonal matrix. Therefore, the goal of PCA dimensionality reduction is to make the diagonal elements of the covariance matrix of the feature dataset as large as possible and the off-diagonal elements as small as possible.

For the original data feature set Q, the steps of PCA algorithm based on eigenvalue decomposition covariance matrix are as follows:

(1) Subtract the average of each dimension from its own.

(2) Calculate the covariance matrix QQ^T/n.

(3) Find the eigenvalue $\lambda_i (i = 1, 2, \ldots, n)$ of covariance matrix with the eigenvalue decomposition method and the corresponding eigenvectors.

(4) The eigenvalues are arranged from large to small, and the total contribution rate of the eigenvalues is calculated to select the feature that retains the most original data information. The total contribution rate of the first k eigenvalues can be expressed as

$$G_k = \frac{\sum_{j=1}^{k} \lambda_j}{\sum_{i=1}^{n} \lambda_i}, \quad 1 \le k \le n. \tag{13}$$

In general applications, the feature vectors corresponding to k eigenvalues with a total contribution rate of more than 85% are selected to form the dimensionality reduction matrix p, and then the principal component dataset Q' can be obtained as $Q' = PQ$.

4. Fault Diagnosis Based on Random Forest and XGBoost

4.1. The Principle of Random Forest. Random forest (RF) is a classification model based on decision tree and Bagging algorithm [32]. Random forest can be regarded as a set of decision tree classifiers, which is a kind of classifier with high accuracy, capable of processing a large number of input features, and able to evaluate the importance of features when determining categories. It has a high fault tolerance for abnormal feature values and is widely used in gene classification, image recognition, and other classification problems. The classification model of random forest is composed of n decision trees, in which each decision tree will classify the input samples. Finally. the random forest will vote on the classification results of each decision tree, and the classification results with the most votes will be counted as the final classification results of random forest [33].

In this paper, the characteristic optimization algorithm used in the decision tree model of random forest is CART algorithm, and every decision tree is binary tree. Random forest is an integration model of multiple decision trees, and Bagging algorithm is a common integration method of random forest model [34]. This method firstly trains several classifiers with training samples and finally sets up multiple classifiers through the clustering method to obtain the final classification results. Compared with a single classifier, the integrated classifier has higher accuracy and generalization [35]. Bagging algorithm from the given training focuses back on the extraction of the training sample, respectively, in each sample drawn to generate corresponding decision tree; then, all decision tree models were tested by testing the sample set and vote on several classification results. Finally, the decision tree with the most votes in the category of the classification result is the test sample. Bagging algorithm is a simple but stable integrated learning algorithm, which can effectively improve the generalization ability of decision tree classification model.

Random forest classification algorithm adds random thought in the steps of Bagging algorithm, which is reflected in two aspects:

(1) For the original dataset, a random sampling method is adopted, which is put back. The data in different subdatasets can be repeated, but the amount of data is consistent with the original dataset. Multiple output results can be obtained by training decision tree classifier according to the subdataset. When the new data are used to test the classification effect, the final classification result of random forest can be obtained by voting the classification results of multiple decision trees.

(2) The random forest randomly extracts a certain amount of feature attributes from the feature set as the node split attribute set and then selects the optimal feature attributes among the features, so that the classification results of each decision tree are not the same, and the performance of the classifier is improved.

The classification process of random forest is as follows:

(1) The Bagging algorithm is used to randomly sample the training dataset with return to obtain the training sample set θ_k

(2) Based on CART algorithm, a binary classification tree is generated for each random training sample, and the nodes of the classification tree select the classification attribute from the random feature attribute subset [36].

(3) Repeat Steps 1 and 2 until all the decision trees can classify all the samples in the training set or all the features are used. Then, the k decision trees are generated, that is, the diagnostic model of random forest is generated.

(4) The test samples are used to test the classification accuracy of random forest. The classification results of samples are determined by the classification results of k decision trees by voting, that is, the classification results of all decision trees with the most number of occurrence are the final results of random forest.

4.2. RF Feature Importance Assessment. In the process of random forest construction, the importance of the features used in each tree is different, so the importance of each feature needs to be evaluated as the basis for feature selection. In this paper, Gini coefficient is used to rank the importance of feature sets and the features with the highest importance are selected as training samples of the XGBoost classifier in the next section.

The feature importance score is represented by VIM to calculate the average change of the information impurity of upper and lower nodes split with feature x_j in all decision trees in the random forest model, that is, the Gini index of this feature. The Gini index of node m can be defined as

$$G_m = 1 - \sum p_k^2. \tag{14}$$

In the above formula, k is the number of representative categories and p_k represents the weight of the k-th sample. The significance of x_j in node m can be defined as the change of Gini index before and after the node splitting, that is,

$$\mathrm{VIM}_{jm}^{(\mathrm{Gint})} = G_m - G_l - G_r, \tag{15}$$

where G_l and G_r, respectively, represent the Gini index of the two children nodes obtained after the node m split to the lower layer of the decision tree. If all nodes x_j of the i-th decision tree of random forest fall into set M, the importance of x_j in the tree can be obtained as follows:

$$\mathrm{VIM}_{ij}^{(\mathrm{Gini})} = \sum_{meM} VIM_{jm}^{(\mathrm{Gin})}. \tag{16}$$

If there are n trees in the random forest, the total importance score of x_j can be further obtained:

$$\mathrm{VIM}_{j}^{(\mathrm{Gini})} = \sum_{i=1}^{n} \mathrm{VIM}_{ij}^{(\mathrm{Gini})}. \tag{17}$$

Finally, the importance score of all features can be normalized as follows:

$$\mathrm{VIM}_j = \frac{\mathrm{VIM}_j^{(\mathrm{Gini})}}{\sum_{i=1}^{c} \mathrm{VIM}_i}. \tag{18}$$

The denominator part in the formula is the sum of all the characteristic gains. In this way, the importance of the features used in the random forest model training can be obtained.

4.3. The Principle of XGBoost. XGBoost (Extreme Gradient Boosting) algorithm is an implementation of the traditional boosting tree algorithm [37]. XGBoost algorithm overrides the target function and defines the tree complexity on the basis of the traditional gradient lifting decision tree. The traditional lifting tree model reduces the loss error through the iteration of several decision trees and finally obtains the classification model. During the construction of decision tree, node splitting is carried out according to the splitting criteria of regression tree, including the least square loss and logarithmic loss.

In XGBoost, the tree structure in the boosting model is decomposed into structure part and leaf weight part, and the complexity of the tree is defined as follows:

$$\Omega(f_t) = rT + \frac{1}{2}\sum_{j=1}^{T} w_j^2, \tag{19}$$

where T is the number of leaf nodes, r is the control parameter of T, w_j is the weight of leaf node, and f_t is the mode function. The calculation method of lifting tree model is shown in formula (20), where L is the loss function and $\widehat{y}_i'$ and y_i, respectively, represent the predicted value and the actual value of the model:

$$L = \sum_i l(\widehat{y}_i' - y_i) + \sum_k \Omega(f_k). \tag{20}$$

Considering the lifting model loss function of n trees, the objective function of tree model learning can be defined as shown in the following equation, where c is a constant:

$$\mathrm{Obj}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \widehat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) + c. \tag{21}$$

The objective function in the above equation is relatively simple to find the optimal solution for the least square loss, but the solution for other loss functions is more complex. On this basis, the XGBoost algorithm is optimized by the second-order Taylor expansion. The second-order Taylor expansion is shown in the following equation:

$$f(x + \Delta x) \approx f(x) + f'(x)\Delta x + \frac{1}{2}f''(x)\Delta x^2. \tag{22}$$

For the original objective function, two variables are defined so that they can be expanded:

$$\begin{aligned} g_i &= \frac{\partial l\left(y_i, \widehat{y}^{(t-1)}\right)}{\partial \widehat{y}^{(t-1)}}, \\ h_i &= \frac{\partial^2 l\left(y_i, \widehat{y}^{(t-1)}\right)}{\partial \left(\widehat{y}^{(t-1)}\right)^2}. \end{aligned} \tag{23}$$

Then, the original objective function can be rewritten in the following form:

$$\mathrm{Obj}^{(t)} \approx \sum_{i=1}^{n} Y_i + \Omega(f_t) + c, \tag{24}$$

where $Y_i = l(y_i, \widehat{y}_i^{(t-1)}) + g_i f_t(x_i) + (1/2)h_i f_t^2(x_i)$.

Define the candidate feature set for each tree when nodes split as $I_j = \{i \mid q(x_i) = j\}$. Then, when the model is trained, the objective function can be expressed as

$$\mathrm{Obj}^{(t)} = \sum_{j=1}^{T}\left[G_j w_j + \frac{1}{2}(H_i + \lambda)w_j^2\right] + \gamma T, \tag{25}$$

where $G_i = \sum_{i \in I_j} g_i$ and $H_i = \sum_{i \in I_j} h_i$.

The value of w_j can be obtained by taking the derivative of the above equation, and the optimal solution can be obtained by substituting it into the function, as shown in equations (26) and (27):

$$w_j^* = -\frac{G_j}{H_j + \lambda}, \tag{26}$$

$$\mathrm{Obj} = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j + \lambda} + \gamma T. \tag{27}$$

In this paper, some feature sets of high importance of the random forest model are selected to train the XGBoost model to complete the classification and recognition of different fault modes, and the classification accuracy of the random forest model (RF) and the classification model

combining random forest and XGBoost algorithm (RF + XGBoost) is compared.

5. CRH2 Fault Injection Simulation Platform

The experimental data used in this paper come from the CRH2 experimental simulation platform of Electric Locomotive Research Institute in Hunan Province. The platform consists of upper computer, dSPACE, traction control unit (DCU), and signal regulator. Its internal control strategy and relevant parameters are consistent with the traction drive system of CRH2 high-speed train. The platform appearance is shown in Figure 2.

The main circuit simulation module is built based on MATLAB and controlled by physical DCU, mainly including traction transformer, rectifier, intermediate DC part, inverter, and 4 asynchronous motors. MATLAB is used to generate code in the upper PC, and it is downloaded to dSPACE real-time simulator to run.

The upper PC can control dSPACE real-time simulation module of the simulator using the software compiled in MATLAB.

DSPACE real-time simulator is mainly used to control the corresponding real-time computing simulation module.

DCU functions as the main circuit control system. It sends control instructions to the simulation module and receives the feedback value calculated by DSPACE to control the main breaker and the contactor in the main circuit.

DCU connects the debugging PC through its own Ethernet port. Thus, the controlling parameters as well as the data can be dynamically displayed and recorded; moreover, DCU provides the function of online disposal of the parameters and program download.

6. Experimental Results and Analysis

In this section, the experimental results of three parts in the process of compound fault diagnosis are introduced. The data used are from the semiphysical simulation experimental platform of Zhuzhou Electric Locomotive Research Institute. The first part is the signal noise deduction based on the SG filter. The second part is fault feature extraction based on wavelet packet and PCA. The third part is fault diagnosis based on random forest and XGBoost.

The U-phase and V-phase current data of 300 sets under three fault modes were collected with 100 sets per fault mode. The three fault modes are as follows: Mode I represents the data of a normal situation; Mode II represents the data of 2-turn stator short circuit and 5% air gap eccentricity; Mode III represents the data of 4-turn stator short-circuit and 10% air gap eccentricity. The percentage of air gap eccentricity represents the ratio of the offset distance between stator and rotor center to the radial distance of the original air gap. In this paper, the data of three fault modes are detected and classified.

6.1. Signal Noise Reduction. The U-phase current signal is collected when the current sensor is normal and the sampling frequency is 2500 Hz. Because of the high sampling frequency, the data segment of 0.1 second is captured for smoothing and noise reduction pretreatment, which is convenient to observe the filtering and smoothing effect. By experimental comparison, the least square polynomial with the order of 4 and window length of 13 is selected to obtain a better smoothing effect. The experimental results are shown in Figure 3.

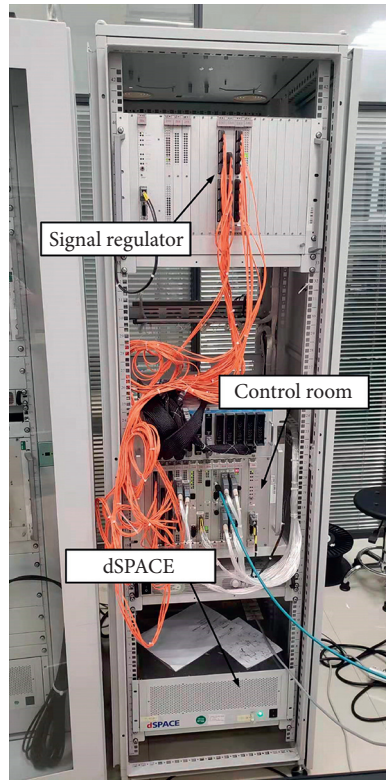

Figure 2: CRH2 experimental simulation platform.

It can be seen from the above figure that the SG filter can effectively reduce the influence of noise and smooth the signal by performing the least-squares fitting filter on the signal in each time window, which lays a foundation for frequency domain feature extraction of composite faults.

6.2. Fault Feature Extraction and Optimization. According to the wavelet packet decomposition principle, the dbN wavelet is used to decompose the U-phase and V-phase currents of three fault modes in three layers. 8 energy values are obtained in its bottom layer to form a 16-dimensional eigenvector.

Three 100×16 matrices are obtained, respectively, by decomposing 300 sets of data of three fault modes, and matrix eigenvalue decomposition approach is used to dimension the matrices. The obtained eigenvalues were arranged from large to small, and the increasing curve of the total contribution rate was obtained as shown in Figures 4–6.

As can be seen from the three pictures above, the total contribution rate of top 9 in the eigenvalues obtained from mode I data and top 8 in the eigenvalues obtained from mode III data is over 90%, while the total contribution rate of top 6 in the eigenvalues obtained from mode II data is over

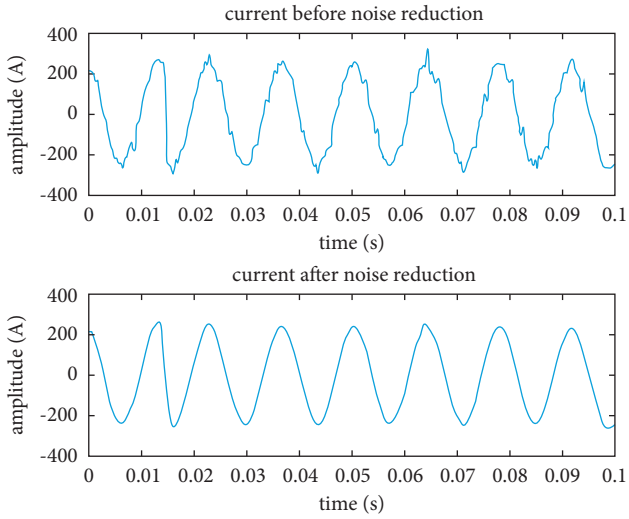

Figure 3: U-phase current before and after noise reduction.

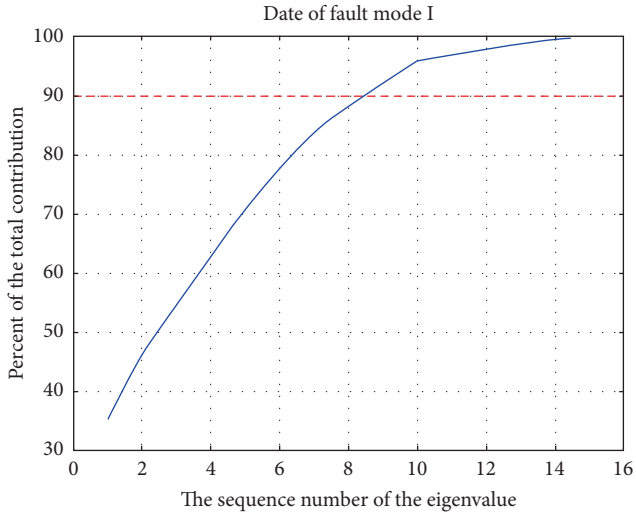

Figure 4: The percent of the total contribution of the eigenvalues obtained from data of mode I.

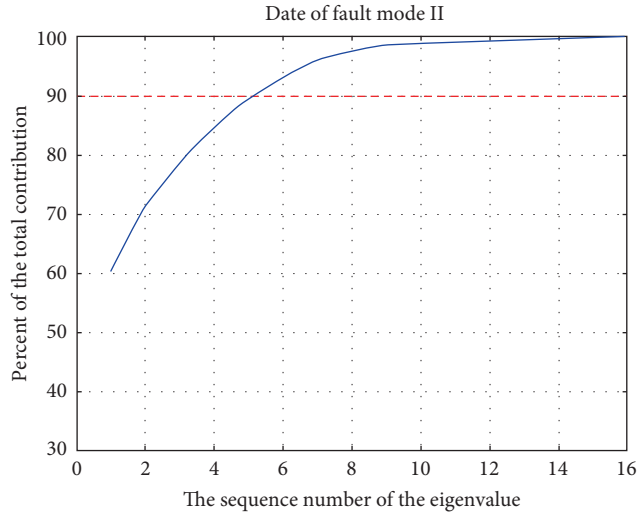

Figure 5: The percent of the total contribution of the eigenvalues obtained from data of mode II.

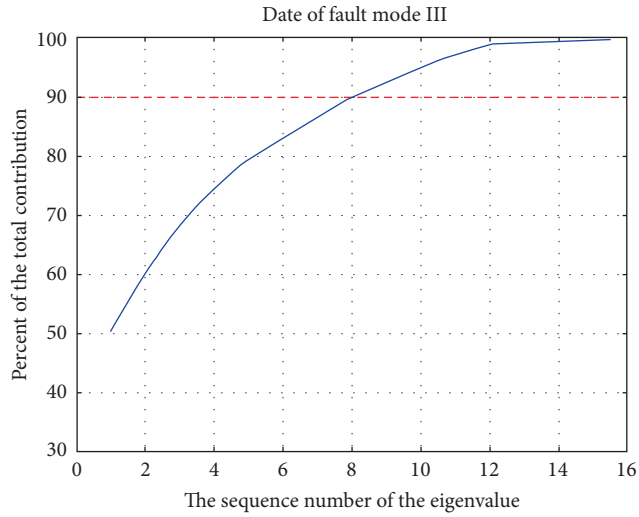

Figure 6: The percent of the total contribution of the eigenvalues obtained from data of mode III.

90%. In order to retain the original information of the signal to the greatest extent, the original 16-dimensional feature data samples of the three fault modes are dimensionalized to a new space composed of feature vectors corresponding to the first nine eigenvalues, and the new 9-dimensional principal component feature dataset $X = \{x_0, x_1 \ldots x_8\}$ is used as the input sample of the classifier in the next section.

6.3. Fault Diagnosis. The 300 groups of 9-dimensional principal component samples are divided into training sample sets and test sample sets in a ratio of 3 : 1. Because of the randomness of sampling in the process of random forest training, for each decision tree, there is a part of data that is not involved in its generation process. This part of data is called out-of-bag (OOB) data of the tree. For each sample in the total sample set, the classification results of all the trees using it as OOB data are calculated, and the final classification results of the sample are obtained by voting. Finally, the ratio of the number of misclassified samples to the total number of samples is used as the OOB error rate of random forest. The OOB error rate is an unbiased estimator of random forest generalization errors [38].

In this section, the random forest classifier of the scikit-learn machine learning module in the Python library is applied to diagnose composite faults. The feature attribute function of decision tree node partition selects the default CART algorithm. The mesh search method was used to adjust other parameters of the model, and the optimal value of parameters was selected by comparing the out-of-pocket error rate of the random forest model. The original sample and the principal component sample are trained, respectively, to compare the performance of the random forest model.

The number of decision trees in the random forest is set as 200, and other parameters are adjusted. The training is conducted on the basis of 300 groups of original feature samples and principal component feature samples, respectively. The change curves of the out-of-bag error rate of the two models with the increase of the number of decision trees are obtained as shown in Figure 7.

The n features and n estimators in the above figure, respectively, represent the dimension of the input feature dataset and the number of the decision tree, while the blue line and the yellow line represent the OOB error rate of the random forest model derived from the 9-dimensional principal component feature dataset and the 16-dimensional original feature dataset, respectively. It can be seen that, after the dimensionality reduction optimization of the input feature dataset, the OOB error rate of the random forest is significantly reduced from 8% to less than 5%, that is, the prediction accuracy of the random forest reaches more than 95%.

The function in the RF model is called feature_importance, which ranks the importance of training features according to the Gini index principle, so as to obtain the normalized score of the feature importance of 9-dimensional principal component feature sample for training the RF model, as shown in Figure 8.

As can be seen from the above figure, after ranking the features according to the importance score, the total importance score of the features ranked in the top 6 reach over 90%. Therefore, it is considered to remove the features ranked in the bottom 3 and train the XGBoost model with the collection of the top 6 features as the input sample.

The XGBClassifier function module in the XGBoost toolkit is used, and 300 groups of 6-bit solicitation are input for training. The proportion of the test dataset is set as 0.5, the training set is used to train the model, and the test set is used to test the accuracy of the model. Since XGBoost's base learner also selects the CART decision tree model, most of the tunable parameters are similar to those in the random forest model. In this section, the web search method is used to adjust parameters, the function is called GridSearchCV, and the optimal parameter combination is determined according to the classification accuracy of the model. The values of some important parameters after parameter adjustment of the XGBoost model are listed in Table 1.

With the exception of the remaining parameters listed in the above table using the default values in the XGBoost toolkit, the final model has an accuracy value of 97.4%. The classification performance of RF model and RF + XGBoost model was tested by reselecting 80 groups of data in each of the three fault modes. The results are shown in Table 2.

It can be seen from the above table that both the RF model and the RF + XGBoost model can effectively diagnose the stator interturn short circuit and air gap eccentricity compound fault compared with other methods such as SVM and ANN. At the same time, the RF + XGBoost model has certain improvement in detection accuracy compared with the RF model alone.

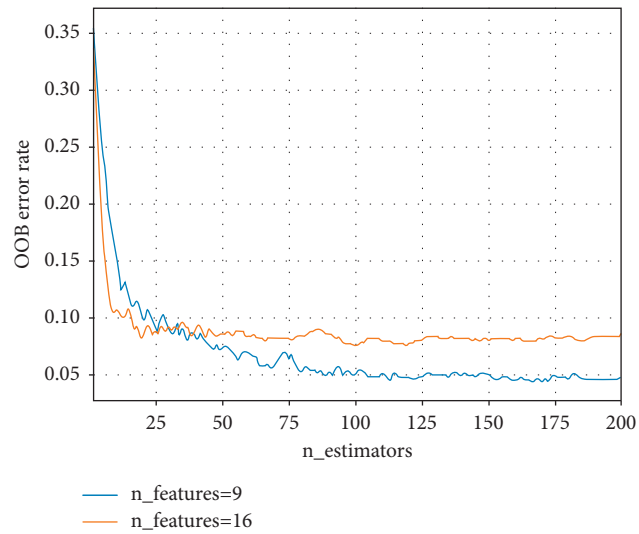

Figure 7: The OOB error rate curve of the RF model.

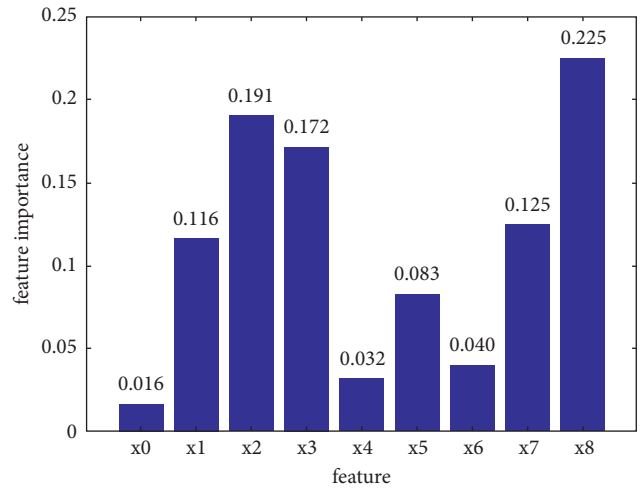

Figure 8: The normalized score of features used in the RF model.

Table 1: Important parameters of the XGBoost model.

Parameter	Value
max_depth	500
max_depth	3
min_weight	1
gamma	0.5
colsample bytree	0.6
learning_rate	0.08
subsample	0.6
reg_alpha	1e-05
reg_lambda	1

7. Results and Discussion

Aiming at the compound fault of stator interturn short circuit and air gap eccentricity of high-speed traction motor, this paper studies a diagnosis algorithm based on random forest and XGBoost. The detailed information of each

Table 2: Test results of different methods.

Method	Fault type	Accuracy	Recall	Precision
SVM	I	0.850	0.88	0.8725
	II	0.8625	0.875	0.86
	III	0.845	0.8575	0.86
ANN	I	0.895	0.87	0.875
	II	0.885	0.8	0.8
	III	0.875	0.85	0.8475
RF	I	0.875	0.8860	0.875
	II	0.8625	0.8414	0.8625
	III	0.8375	0.8481	0.8375
RF + XGB	I	0.975	0.974	0.975
	II	0.9875	0.9634	0.9875
	III	0.975	1	0.975

frequency band of current signal is extracted by wavelet packet decomposition to form fault features. Considering the problem of too high dimension of feature vectors, principal component analysis is adopted to eliminate unimportant features. Finally, the dimensionality reduced preprocessed feature vectors are used to train the random forest and XGBoost classifier, which improves the prediction accuracy and generalization performance of the model. Based on the onboard experimental data of CRH2 train motor provided by Zhuzhou, the effectiveness of the diagnosis method is proved. In the future, further research will be made on the composite faults under no-load and half load conditions of the motor and multiple diagnosis classifiers will be designed for different working conditions to improve the application scope of the diagnosis method.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This project was financially supported by the National Natural Science Foundation of China (61490703) and the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

[1] X. Lu, X. Wang, and Q. Guo, "Research on application of equivalent spwm modulation strategy in high speed train traction drive system," *Electrotehnica, Electronica, Automatica*, vol. 65, no. 2, 2017.

[2] A. Muxiri, F. Bento, D. Fonseca, and A. Cardoso, "Thermal analysis of an induction motor subjected to inter-turn sort-circuit failures in the stator windings," in *Proceedings of the International Conference on Industrial Engeneering, Applications and Manufacturing*, Sochi, Russia, March 2019.

[3] S. Imajo, S. Yamashita, H. Akutsu, H. Kumagai, and Y. Nakazawa, "Gap symmetry of the organic superconductor -(bets) 2 gacl 4 determined by magnetic-field-angle-resolved heat capacity," *Journal of the Physical Society of Japan*, vol. 88, no. 2, Article ID 023702, 2019.

[4] H. Nakamura, "Proposal of method for identifying short-circuit fault points in concentrated windings of motor stators," *IEEE Transactions on Industry Applications*, vol. 138, no. 7, pp. 623–629, 2018.

[5] "Composite fault diagnosis of rotor broken bar and air gap eccentricity based on park vector module and decision tree algorithm," in *Proceedings of the 2019 CAA Symposium on Fault Detection, Supervision and Safety for Technical Processes (SAFEPROCESS)*, Xiamen, China, July 2019.

[6] H. Qiu, W. Yu, S. Yuan, B. Tang, and C. Yang, "The influence of inter-turn short circuit fault considering loop current on the electromagnetic field of permanent magnet synchronous motor," *COMPEL: The International Journal for Computation & Mathematics in Electrical & Electronic Engineering*, vol. 36, no. 4, pp. 1028–1042, 2017.

[7] J. Haddad, S. Cattari, and S. Lagomarsino, "Use of the model parameter sensitivity analysis for the probabilistic-based seismic assessment of existing buildings," *Bulletin of Earthquake Engineering*, vol. 17, no. 4, pp. 1983–2009, 2021.

[8] Y. Tian, D. Guo, K. Zhang, L. Jia, H. Qiao, and H. Tang, "A review of fault diagnosis for traction induction motor," in *Proceedings of the 2018 37th Chinese Control Conference (CCC)*, pp. 5763–5768, Wuhan, China, July 2018.

[9] Y. Zou, Y. Zhang, and H. Mao, "Fault diagnosis on the bearing of traction motor in high-speed trains based on deep learning," *Alexandria Engineering Journal*, vol. 60, no. 1, pp. 1209–1219, 2021.

[10] S. H. Cho, S. Kim, and J.-H. Choi, "Transfer learning-based fault diagnosis under data deficiency," *Applied Sciences*, vol. 10, no. 21, 2020.

[11] J. Sun, Z. Miao, D. Gong, X.-J. Zeng, J. Li, and G. Wang, "Interval multiobjective optimization with memetic algorithms," *IEEE Transactions on Cybernetics*, vol. 50, no. 8, pp. 3444–3457, 2020.

[12] T. Xu, X. Wang, and Z. Li, "Fault diagnosis of rolling element bearing for the traction system of high-speed train based on wavelet segmented threshold de-noising and HHT," in *Proceedings of the 4th International Conference on Electrical and Information Technologies for Rail Transportation (EITRT)*, Qingdao, China, October 2019.

[13] Yuangui, Y. Yang, Z. Zhao, Y. Zuo, and Wang, "Determination of total flavonoids for paris polyphylla var. yunnanensis in different geographical origins using uv and ft-ir spectroscopy," *Journal of Aoac International*, vol. 101, no. 3, pp. 732–738, 2018.

[14] H. Zheng-Hai and S. Jie, "A smoothing Newton algorithm for mathematical programs with complementarity constraints," *Journal of Industrial and Management Optimization*, vol. 1, no. 2, pp. 153–170, 2017.

[15] Y. V. García-Tejeda and V. Barrera-Figueroa, "Least squares fitting-polynomials for determining inflection points in adsorption isotherms of spray-dried açaí juice (Euterpe oleracea Mart.) and soy sauce powders," *Powder Technology*, vol. 342, pp. 829–839, 2019.

[16] Y. Liu, B. Dang, Y. Li, H. Lin, and H. Ma, "Applications of savitzky-golay filter for seismic random noise reduction," *Acta Geophysica*, vol. 64, no. 1, pp. 101–124, 2016.

[17] K. S. Sim, M. A. Kiani, M. E. Nia, and C. P. Tso, "Signal-to-noise ratio estimation on SEM images using cubic spline interpolation with Savitzky-Golay smoothing," *Journal of Microscopy*, vol. 253, no. 1, pp. 1–11, 2013.

[18] X. Li and H. Ning, "Chinese text classification based on hybrid model of cnn and lstm," in *Proceedings of the DSIT 2020: 2020 3rd International Conference on Data Science and Information Technology*, Xiamen, China, July 2020.

[19] A. Panuszko, J. Stangret, B. Nowosielski, and P. Bruździak, "Interactions between hydration spheres of two different solutes in solution: the least squares fitting with constraints as a tool to determine water properties in ternary systems," *Journal of Molecular Liquids*, vol. 310, p. 113181, 2020.

[20] Y. K. Lee, "A new method to minimize overall torque ripple in the presence of phase current shift error for three-phase BLDC motor drive," *Canadian Journal of Electrical and Computer Engineering*, vol. 42, 2019.

[21] F. Chris, *Audio Signal Bandwidth to Codec Bandwidth Analysis and Response*, My Science Work, San Francisco, CA, USA, 2020.

[22] B. Bessam, A. Menacer, M. Boumehraz, and H. Cherif, "Wavelet transform and neural network techniques for inter-turn short circuit diagnosis and location in induction motor," *International Journal of System Assurance Engineering and Management*, vol. 8, no. 1, pp. 478–488, 2017.

[23] T. Hu, J. Zhao, S. Yan, and W. Zhang, "Performance analysis of a wavelet packet transform applied to concrete ultrasonic detection signals," *Journal of Physics: Conference Series*, vol. 1894, no. 1, p. 9, Article ID 012062, 2021.

[24] J. Köhler, M. Autenrieth, and W. H. Beluch, "Uncertainty based detection and relabeling of noisy image labels," 2019, https://arxiv.org/abs/1906.11876.

[25] J. Li and X. Cai, "Summation pollution of principal component analysis and an improved algorithm for location sensitive data," *Numerical Linear Algebra with Applications*, vol. 1, 2021.

[26] H. Hu, Y. Yan, Q. Zhu, and G. Zheng, "Generation and frame characteristics of predefined evenly-distributed class centroids for pattern classification," 2021, https://arxiv.org/abs/2105.00401.

[27] E. Banguero, A. Correcher, Á. Pérez-Navarro, E. García, and A. Aristizabal, "Diagnosis of a battery energy storage system based on principal component analysis," *Renewable Energy*, vol. 146, no. 2, pp. 2438–2449, 2020.

[28] S. Gorfman, "Algorithms for target transformations of lattice basis vectors," *Acta crystallographica. Section A, Foundations and advances*, vol. 76, no. 6, pp. 713–718, 2020.

[29] B. Kska, "Parameterization of the crystal structure of garnet in terms of symmetry-adapted basis-vectors of the ideal tetrahedron and octahedron: application to the pressure-dependence of the crystal structure of y 3 al 5 o 12 between 0 and 126gpa," *Materials Chemistry and Physics*, vol. 227, pp. 72–82, 2019.

[30] A. Ishii, K. Yata, and M. Aoshima, "Hypothesis tests for high-dimensional covariance structures," *Annals of the Institute of Statistical Mathematics*, vol. 73, 2020.

[31] S. Zachary, D. J. Eisenstein, B. Florian et al., "The large-scale three-point correlation function of the sdss boss dr12 cmass galaxies," *Monthly Notices of the Royal Astronomical Society*, vol. 1, p. 1, 2017.

[32] J. W. Zhang, W. Shen, L. F. Liu, and Z. D. Wu, "Face recognition model based on privacy protection and random forest algorithm," in *Proceedings of the 2018 27th Wireless and Optical Communication Conference (WOCC)*, Hualien, Taiwan, April-May 2018.

[33] J. H. Kim, H. K. Kim, K. H. Jang, J. M. Lee, and Y. S. Moon, "Object classification method using dynamic random forests and genetic optimization," *Journal of the Korea Society of Computer and Information*, vol. 21, no. 5, pp. 79–89, 2016.

[34] A. Dogan and D. Birant, "A two-level approach based on integration of bagging and voting for outlier detection," *Journal of Data and Information Science*, vol. v.5, no. 2, pp. 113–137, 2020.

[35] G. Xu, M. Liu, Z. Jiang, D. Söffker, and W. Shen, "Bearing fault diagnosis method based on deep convolutional neural network and random forest ensemble learning," *Sensors*, vol. 19, no. 5, 2019.

[36] O. Rahmati, F. Falah, S. A. Naghibi, T. Biggs, and D. T. Bui, "Land subsidence modelling using tree-based machine learning algorithms," *The Science of the Total Environment*, vol. 672, 2019.

[37] C. Fang, Z. Shao, and C. Wu, "A low-cost method for designing and updating a drgs classifier based on machine learning," in *Proceedings of the ICMHI 2020: 2020 4th International Conference on Medical and Health Informatics*, Kamakura City, Japan, August 2020.

[38] B. An and Y. Suh, "Identifying financial statement fraud with decision rules obtained from modified random forest," *Data Technologies and Applications*, vol. 54, no. 2, pp. 235–255, 2020.