

## Research Article

# Research on Multidomain Fault Diagnosis of Large Wind Turbines under Complex Environment

Rong Jia,<sup>1</sup> Fuqi Ma<sup>1</sup>,<sup>1</sup> Jian Dang,<sup>1</sup> Guangyi Liu,<sup>2</sup> and Huizhi Zhang<sup>1</sup>

<sup>1</sup>Key Laboratory of Smart Energy in Xi'an, Xi'an University of Technology, Xi'an 710048, China

<sup>2</sup>GEIRI North America, 250 W Tasman Dr., San Jose, CA 95134, USA

Correspondence should be addressed to Fuqi Ma; 18392647176@163.com

Received 2 April 2018; Accepted 6 June 2018; Published 17 July 2018

Academic Editor: Changzhi Wu

Copyright © 2018 Rong Jia et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Under the complicated environment of large wind turbines, the vibration signal of a wind turbine has the characteristics of coupling and nonlinearity. The traditional feature extraction method for the signal is hard to accurately extract fault information, and there is a serious problem of information redundancy in fault diagnosis. Therefore, this paper proposed a multidomain feature fault diagnosis method based on complex empirical mode decomposition (CEMD) and random forest theory (RF). Firstly, this paper proposes a novel method of complex empirical mode decomposition by using the correlation information between two-dimensional signals and utilizing the idea of ensemble empirical mode decomposition (EEMD) by adding white noise to suppress the problem mode mixing in empirical mode decomposition (EMD). Secondly, the collected vibration signals are decomposed into IMFs by CEMD. Then, calculate 11 time domain characteristic parameters and 13 frequency domain characteristic parameters of the vibration signal, and calculate the energy and energy entropy of each IMF components. Make all the characteristic parameters as the multidomain feature vectors of wind turbines. Finally, the redundant feature vectors are eliminated by the importance of each feature vector which has been calculated, and the feature vectors selected are input to the random forest classifier to achieve the fault diagnosis of large wind turbines. Simulation and experimental results show that this method can effectively extract the fault feature of the signal and achieve the fault diagnosis of wind turbines, which has a higher accuracy of fault diagnosis than the traditional classification methods.

## 1. Introduction

As a kind of abundant, renewable, and efficient clean energy, wind energy has developed rapidly in recent years. Currently, wind power generation technology has become an important research area for countries to compete and has been promoted to the height of national strategy [1–3]. As the installed capacity of wind turbines becomes larger and larger, the structure of the turbines becomes more and more complicated, and they work under harsh conditions for a long period of time. Therefore, higher requirements are put on the fault diagnosis technology of the wind turbines [4, 5]. It is of great significance for wind turbine condition monitoring and fault diagnosis accurately and comprehensively to extract the fault feature of vibration signals [6–8].

Since the wind turbine fault vibration signals with the characteristics of nonlinear and nonstationary [9], at present, many scholars have done some research on the fault feature extraction of a wind turbine. The main method uses vibration sensors which acquire the vibration signal of wind turbine, utilize some methods with strong applicability for feature extraction, and then use fault diagnosis methods to diagnose the fault by utilizing fault information extracted for wind turbine. The methods for signal processing include wavelet transform (WT) [10, 11], Hilbert-Huang transform, empirical mode decomposition (EMD) [12, 13], and variational mode decomposition (VMD) [14, 15]. For instance, Gao et al. [16] utilize load mean decomposition (LMD) decomposing the vibration signal into multiple product functions. The characteristic parameters were achieved by the

multiscale entropy method of processing the main product functions. The characteristic parameters were entered into the least square support vector machine (SVM) for fault diagnosis of the wind turbine.

Muralidharan and Sugumaran [17] compute the wavelet features by using discrete wavelet transform (DWT) from the vibration signals. And the rough sets are generated by wavelet features to classify using the fuzzy logic. Jiao et al. [18] use the EMD method to decompose the original vibration signals into finite intrinsic mode functions (IMFs) and a residual. And the energy of the first four IMFs is extracted as vibration signal fault feature. A probabilistic neural network (PNN) model is established to achieve the fault classification. However, these methods all use the signal processing method to extract the time-frequency characteristic information of the vibration signal, and the feature information extracted is often not comprehensive enough.

In order to comprehensively extract the fault feature information, many scholars have studied the method of multidomain feature fault diagnosis. Tang et al. [19] proposed a novel method for fault diagnosis based on manifold learning and Shannon wavelet support vector machine. And the Shannon wavelet support vector machine (SWSVM) is established to recognize faults by using the mixed-domain features extracted. Gan et al. [20] obtain the time domain and frequency domain characteristics of vibration signals by singular value decomposition (SVD) and utilize the multidomain manifold learning to achieve this method to realize the fault diagnosis of mechanical equipment. Shen et al. [21] decompose the vibration signal into IMFs by empirical mode decomposition (EMD). 13 time domain characteristic parameters and 16 frequency domain characteristic parameters were extracted, and the parameters into the support vector machine model for fault diagnosis were input. However, there are still some shortcomings in the current research of multidomain feature fault diagnosis. It includes that the effect of traditional time-frequency signal processing methods is often not ideal, and with the increase of feature vectors, it is more difficult for the wind turbine to diagnose and there will be redundant feature information in multidomain feature vectors.

Complex empirical mode decomposition (CEMD) is based on the principle of bivariate empirical mode decomposition, which uses the correlation information between two-dimensional signals to decompose synchronously and utilizes the principle of ensemble empirical mode decomposition (EEMD) through adding white noise to suppress mode mixing. This method can effectively improve the problem of mode mixing in EMD. As a classical algorithm in ensemble learning, random forest (RF) can not only effectively solve the problems of artificial neural network such as slow convergence and over-fitting but can also solve the shortcomings of the SVM algorithm's inability to process large sample data [22–24].

Considering the advantages of the two algorithms, this paper proposes a multidomain feature fault diagnosis method based on complex empirical mode decomposition and random forest theory and applies it to the fault diagnosis of wind turbines. The specific arrangement of this paper is

organized as follows. The view on the principle of complex empirical mode decomposition (BEMD) is illustrated in Section 2. Section 3 describes the method of multidomain feature vector extraction. Section 4 gives brief introductions of random forest theory. Section 5 describes the multidomain feature fault diagnosis method based on CEMD and random forest theory. Section 6 is the simulation verification of CEMD this paper proposed. Section 7 applies the proposed method to fault signals of rolling bearing. Conclusions come in Section 8.

## 2. The Principle of Complex Empirical Mode Decomposition

*2.1. The Basic Theory of CEMD.* At present, many scholars have done some research on the algorithm of CEMD. Tanaka and Mandic [25, 26] proposed a complex empirical mode decomposition to process two-dimensional signals, but the essence of this method is to perform empirical mode decomposition on the real and imaginary parts of the complex data composed of two-dimensional signals. But this method does not consider the correlation between real and imaginary parts in the decomposition process. Rilling et al. [27] proposed a new algorithm of bivariate empirical mode decomposition (BEMD) which fully considers the correlation between the real and imaginary parts, and unified decomposed complex data signals contained real and imaginary parts so that the decomposition results also have physical meanings. Therefore, this paper uses this method to perform complex data empirical mode decomposition. The main process is as follows [27]:

*Step 1.* Determine the projection direction  $\varphi_k = 2k\pi/N$ , where  $1 \leq k \leq N$ .

*Step 2.* The two-dimensional signal ( $t$ ) is projected onto the  $\phi_k$ .

$$p_{\varphi_k}(t) = \text{Re}(e^{-i\varphi_k}x(t)). \quad (1)$$

*Step 3.* Extract the corresponding moment for the local maximum of  $X\{t_j^k\}$ ; then, the set  $\{t_j^k, e^{i\varphi_k}p_j^k\}$  is interpolated. Get the maximum envelope  $e'_{\varphi_k}(t)$  in the direction  $\varphi_k$ .

*Step 4.* Calculate the mean of the maximum envelope  $e'_{\varphi_k}(t)$  in each direction.

$$m(t) = \frac{2}{N} \sum_{k=1}^N e'_{\varphi_k}(t). \quad (2)$$

*Step 5.* Similar to the EMD decomposition process, calculate the residual component:

$$S(t) = x(t) - m(t), \quad (3)$$

whether the  $S(t)$  meets the requirements of IMF. If satisfied, proceed to Step 6. If not, repeat carried out Steps 2–6, until  $S(t)$  satisfies the conditions of the intrinsic mode function IMF.

*Step 6.* Record the resulting IMF, and remove it from the original signal. And obtain the IMF1 as  $c_1(t) = h(t)$ ; residual component can be expressed as

$$r_1(t) = x(t) - c_1(t). \quad (4)$$

*Step 7.* Repeat the above steps until you get all the IMFs. The original signal can be expressed as

$$x(t) = \sum_{k=1}^K c_k(t) + r_K(t), \quad (5)$$

where  $K$  represents the total number of IMFs.

**2.2. The Decomposition Principle of CEMD.** The decomposition principle of CEMD this paper proposed is based on the bivariate empirical mode decomposition proposed by Rilling et al. [27]. The specific construction ideas are as follows:

Let  $x_0(t)$  be the original vibration signal collected by the vibration sensor. Let  $x_n(t)$  be the white noise signal with a certain amplitude. Thus, a complex signal  $x_c(t)$  constitutes as follows:

$$x_c(t) = x_0(t) + ix_n(t). \quad (6)$$

Project the complex signal  $x_c(t)$  into all directions [19]:

$$p_{\varphi_k}(t) = \text{Re} \left( e^{-i\varphi_k} (x_0(t) + ix_n(t)) \right). \quad (7)$$

Substituting (7) into Euler's formula can simplify

$$p_{\varphi_k}(t) = x_0(t) \cos \varphi_k + x_n(t) \sin \varphi_k. \quad (8)$$

Formula (6) indicates, when  $\sin \varphi_k \neq 0$ , in other words,  $\varphi_k \neq k\pi (k = 1, 2, \dots)$ . The projection  $p_{\varphi_k}(t)$  is equivalent to adding white noise with limited amplitude to observation signals that scale at different scales. It can be seen that, in the given direction, the added noise has an effect on the selection extreme points for the signal. Then, the complex data can be obtained by again projecting the resulting project. That is to say, the data should be interpolated in Step 3 which can be expressed as  $e^{i\varphi_k} p_j^k$ :

$$\begin{aligned} e^{i\varphi_k} p_j^k &= e^{i\varphi_k} (x_0(t) \cos \varphi_k + x_n(t) \sin \varphi_k) \\ &= x_0(t) \cos^2 \varphi_k + x_n(t) \cos \varphi_k \sin \varphi_k \\ &\quad + i [x_0(t) \cos \varphi_k \sin \varphi_k + x_n(t) \sin^2 \varphi_k]. \end{aligned} \quad (9)$$

Then, the real and imaginary parts of the complex signal  $x_c(t)$  obtained are interpolated separately. Assume that the interpolated value of the real part  $x_0(t)$  of the complex signal is  $x'_0(t)$ . Assume that the interpolated value of the imaginary part  $x_n(t)$  of the complex signal is  $x'_n(t)$ . So the interpolation of the complex signal can be expressed as

$$\begin{aligned} e^{i\varphi_k} p_j^k &= x'_0(t) \cos^2 \varphi_k + x'_n(t) \cos \varphi_k \sin \varphi_k \\ &\quad + i [x'_0(t) \cos \varphi_k \sin \varphi_k + x'_n(t) \sin^2 \varphi_k]. \end{aligned} \quad (10)$$

After finding the envelope of the maximal values in each direction, we need to average the projections of the complex signal to obtain the centroids in all directions. When the number of projection directions selected approaches infinity, the idea of integration can be used. Considering that the processing object of this method is the real part of the complex data that collected original vibration signal, therefore, only the real part of the complex signal needs to be integrated. The result is as follows:

$$\begin{aligned} &\int_0^{2\pi} x'_0(t) \cos^2 \varphi_k + x'_n(t) \cos \varphi_k \sin \varphi_k d\varphi \\ &= \int_0^{2\pi} x'_0(t) \cos^2 \varphi_k d\varphi + \int_0^{2\pi} x'_n(t) \cos \varphi_k \sin \varphi_k d\varphi \quad (11) \\ &= x'_0(t) \int_0^{2\pi} \cos^2 \varphi_k d\varphi + x'_n(t) \int_0^{2\pi} \cos \varphi_k \sin \varphi_k d\varphi \\ &= \pi \cdot x'_0(t). \end{aligned}$$

It can be seen that white noise is added as the imaginary part of the complex data, and the projection that the decomposition of imaginary noise projects on the real part can assist in the selection of extreme points of the real part. So the phenomenon of mode mixing can be reduced. In addition, the added white noise is completely canceled when the average is calculated, and it does not affect the original signal.

### 3. Multidomain Feature Vector Extraction

In order to obtain comprehensive fault feature information, this paper uses parameter statistical analysis, Fourier transform, and complex empirical mode decomposition to extract multidomain feature vectors from fault diagnosis signals. There are 11 time domain feature vectors, 13 frequency domain feature vectors, and specific time-frequency characteristics [28]. The specific parameters are shown in Tables 1 and 2.

In addition, considering that when the drive system of the wind turbine fails, the energy in different frequency bands of the fault vibration signal will change, and the energy distribution of each frequency band will also change accordingly. Since the CEMD algorithm proposed in this paper can decompose the original vibration signal into stable IMF components in different frequency bands, on the basis of the decomposition of CEMD, the energy of each IMF component and the energy distribution of each frequency band are calculated to obtain the time-frequency domain characteristics of fault diagnosis. The specific time-frequency characteristic parameters are calculated as follows.

The fault vibration signals  $x(t)$  of wind turbines can be decomposed by CEMD to obtain the intrinsic mode functions (IMFs). And calculate the energy information  $E_1, E_2, \dots, E_n$  of each component to reflect the energy of every IMF.

It is known from the conservation of energy that the sum of the energy of  $n$  components should be equal to the total

TABLE 1: Time domain feature vectors of vibration signals.

Number	Characteristic expression
1	$T_1 = (1/N) \sum_{n=1}^N x(n)$
2	$T_2 = \sqrt{(1/N) \sum_{n=1}^N [x(n) - T_1]^2}$
3	$T_3 = \sqrt{(1/N) \sum_{n=1}^N x^2(n)}$
4	$T_4 = \sqrt{(1/N) \sum_{n=1}^N  x(n) ^2}$
5	$T_5 = \max  x(n) $
6	$T_6 = \left( \sum_{n=1}^N (x(n) - T_1)^3 \right) / ((n-1)T_2^3)$
7	$T_7 = \left( \sum_{n=1}^N (x(n) - T_1)^4 \right) / ((n-1)T_2^4)$
8	$T_8 = T_5/T_4$
9	$T_9 = T_5/T_3$
10	$T_{10} = T_4 / \left( (1/N) \sum_{n=1}^N  x(n)  \right)$
11	$T_{11} = T_5 / \left( (1/N) \sum_{n=1}^N  x(n)  \right)$

In the equations in Table 1,  $x(n)$ ,  $i = 1, 2, \dots, N$  represents the given signal and  $N$  is the data length.

energy of the original signal. And define the energy entropy to reflect the energy distribution of each IMF component. The energy entropy formula can be expressed as

$$H_{EN} = - \sum_{i=1}^n p_i \lg p_i, \quad (12)$$

where  $p_i$  represents the proportion of the energy of the  $i$ th intrinsic mode function IMF $_i$  in the total energy.  $p_i = E_i/E$  and  $E = \sum_{i=1}^n E_i$ .

#### 4. Random Forest Theory and Algorithm

The random forest [29] is based on the decision tree. It is composed of multiple decision trees, and the final result is decided by the voting principles of multiple decision trees. The basic process is as follows: firstly, the original samples were resampled by using the bootstrap method. Randomly select  $S$  samples from the original sample set to form the bootstrap sample set, and using different bootstrap sample sets to train, the decision tree was obtained. When constructing the decision tree, a subset of  $m$  attributes is randomly selected from the all feature attribute set, and the attributes of the subset are used to implement feature classification.

**4.1. The Classification Principles of Random Forest Theory.** Random forest (RF) is a collection classifier  $\{h(a, \theta_k, k = 1, 2, \dots, K)\}$  composed of multiple decision trees, where  $K$  represents the total number of decision trees,  $\{\theta_k\}$  represents independent and identical random vectors, and  $a$  represents the randomly input feature vector that is to

TABLE 2: Frequency domain characteristic parameters of vibration signal.

Number	Characteristic expression
1	$F_1 = (1/K) \sum_{k=1}^K s(k)$
2	$F_2 = (1/(K-1)) \sum_{k=1}^K [s(k) - F_1]^2$
3	$F_3 = \left( 1 / \left( K \left( (\sqrt{F_2})^3 \right) \right) \right) \sum_{k=1}^K [s(k) - F_1]^3$
4	$F_4 = (1/K F_2^2) \sum_{k=1}^K [s(k) - F_1]^4$
5	$F_5 = \left( \sum_{k=1}^K f_k s(k) \right) / \left( \sum_{k=1}^K s(k) \right)$
6	$F_6 = \sqrt{\left( \sum_{k=1}^K (f_k - F_5)^2 s(k) \right) / K}$
7	$F_7 = \sqrt{\left( \sum_{k=1}^K f_k^2 s(k) \right) / \left( \sum_{k=1}^K s(k) \right)}$
8	$F_8 = \sqrt{\left( \sum_{k=1}^K f_k^4 s(k) \right) / \left( \sum_{k=1}^K s^2(k) \right)}$
9	$F_9 = \left( \sum_{k=1}^K f_k^2 s(k) \right) \left( \sum_{k=1}^K s(k) \sum_{k=1}^K f_k^4 s(k) \right)^{-1/2}$
10	$F_{10} = F_6 / F_5$
11	$F_{11} = \left( \sum_{k=1}^K (f_k - F_5)^3 s(k) \right) / K F_6^3$
12	$F_{12} = \left( \sum_{k=1}^K (f_k - F_5)^4 s(k) \right) / K F_6^4$
13	$F_{13} = \left( \sum_{k=1}^K (f_k - F_5)^{1/2} s(k) \right) / K F_6^{1/2}$

In the equations in Table 2,  $s(k)$ ,  $i = 1, 2, \dots, K$  represents the spectral line of  $x(n)$ ,  $K$  is the total number of spectral lines, and  $f_k$  is the frequency value of the  $k$ th spectrum line.

say independent variables. The theory uses a simple majority voting method or takes the average of the output of each decision tree to determine the final classification result. Random forests effectively solve the problems of pan-performance difference and over-fitting in the decision tree by aggregating multiple CART trees. The specific algorithm is as follows [30]:

- (1) The Bagging method [31] is used to sample a given set of training samples  $X$  to obtain a self-help sample set  $\theta_k$ , that is, to use random sampling technology to perform random sampling on sample set  $X$ . The number of self-help sample set obtained is equal to the number of  $X$ .
- (2) Based on the CART algorithm, a binary tree corresponding to each self-help sample set  $\theta_k$  is generated. The specific process is as follows:
  - (a) Assuming that there are  $M$  feature attributes in total, when constructing the decision tree of each node,  $m$  feature attributes are randomly selected from  $M$  feature attributes as candidate attributes for sample classification. The empirical formula is given in [32], generally,  $m = \sqrt{M}$ , and round down.



- (b) According to the principle of minimum Gini impurity purity, a feature selected from the  $m$  feature vectors is regarded as the optimal classification attribute of the node for sample classification.
- (c) According to the characteristics, the nodes are divided into two branches, and the feature vectors with the best classification effect are searched from the remaining features, and it is ensured that each binary tree can fully grow without pruning.
- (3) Repeat Steps 1 and 2 until the tree can accurately classify the training set, or use all the attributes, and then, combine the  $K$  decision trees to generate a random forest classification model.
- (4) For a given unknown sample, the final output type result is generally obtained using the majority voting method [33]. The specific results are as follows:

$$c = \arg \max_c \left( \frac{1}{K} \sum_k I(h(a, \theta_k) = c) \right), \quad (13)$$

where  $I(\cdot)$  represents the pointer function and  $c$  represents the sample type with the most votes.

For all test samples, the mixed matrix CM is obtained after voting, where the elements in the mixing matrix  $CM(i, j)$  represent the number of times that the  $i$ th sample was divided into type  $j$ .

When  $i = j$  represents the correct classification, the correct rate CA of randomized forest classification can be expressed as

$$CA = \frac{\sum_{i=1}^M CM(i, i)}{\sum_{i,j=1}^M CM(i, j)}. \quad (14)$$

According to the above theory, the number  $K$  of decision trees that are set during the construction of the random forest classification model has an important influence on the accuracy of the model and the calculation efficiency. In general,  $K$  takes around 500 to 1000 [34].

This paper utilizes the mean decrease in accuracy and the mean decrease in the Gini index to measure the importance of feature vectors in random forests. The mean decrease in accuracy directly measures the effect of each feature on the accuracy of the random forest model. The main idea of this method is to disrupt the sequence of each feature vectors and measure the effect of sequence variation on the accuracy of the model. Obviously, the sequence of scrambling will have a much greater effect on the accuracy of the model if the feature vectors are more important. The mean decrease in the Gini index is another characteristic metric for the feature extract. The Gini index or information gain is usually used to measure the importance of feature vectors in the decision tree. Thus, the average number of decrease impurity in the Gini index of every feature vector can be used for the value of feature selection for random forests.

## 5. Multidomain Fault Diagnosis Based on CEMD-RF

In order to more comprehensively extract the fault feature information from the wind turbine vibration signal and solve the problem that the traditional fault diagnosis method has low recognition accuracy, this paper proposes a multidomain feature fault diagnosis method based on complex data empirical mode decomposition (CEMD) and random forest theory. The extraction process is shown in Figure 1, and the specific implementation steps are as follows:

*Step 1.* In signal acquisition. Utilize the vibration sensor to collect the fault vibration signal of the wind turbine operated.

*Step 2.* Calculate the time domain and frequency domain feature vectors of the original vibration signal. Statistical parameter analysis and Fourier transform were used to calculate the 11 time domain feature vectors and 13 frequency domain feature vectors of the wind turbine vibration signal collected.

*Step 3.* The fault vibration signal is decomposed by CEMD to obtain  $n$  intrinsic mode functions.

*Step 4.* Calculate the time-frequency characteristics of the IMFs. The energy and energy entropy of each IMF component are calculated.

*Step 5.* Build a training sample set  $X$ . 11 time domain feature vectors, 13 frequency domain feature vectors, and  $n + 1$  time-frequency domain feature vectors are calculated to form a training sample set  $X$ .

*Step 6.* Establish a random forest theory classifier. The training sample set  $X$  is trained to get the classifier of the corresponding random forest theory.

*Step 7.* Calculate the importance of each feature vector. Using the random forest classifier in Step 6, the out of bag (OOB) error of each feature vectors is calculated.

*Step 8.* Eliminate redundant feature vectors. According to the importance of each characteristic parameter obtained by Step 7, the less important characteristic parameters are eliminated to achieve redundant feature information eliminated.

*Step 9.* Build a new feature training set  $X'$ . The feature parameters after Step 8 are selected as the new training sample set  $X'$ .

*Step 10.* In fault pattern recognition, the random forest classifier based on the selected feature training set  $X'$  is established to realize the fault pattern recognition of the wind turbine.

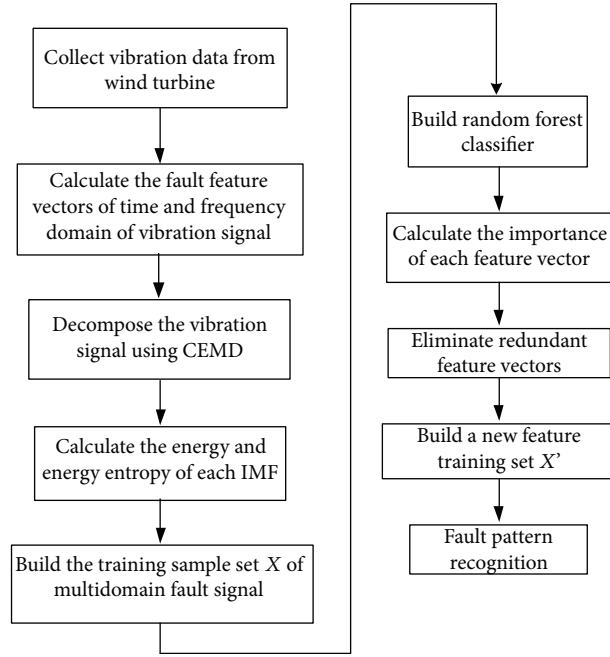


FIGURE 1: Flow chart of multidomain fault diagnosis based on CEMD-RF.

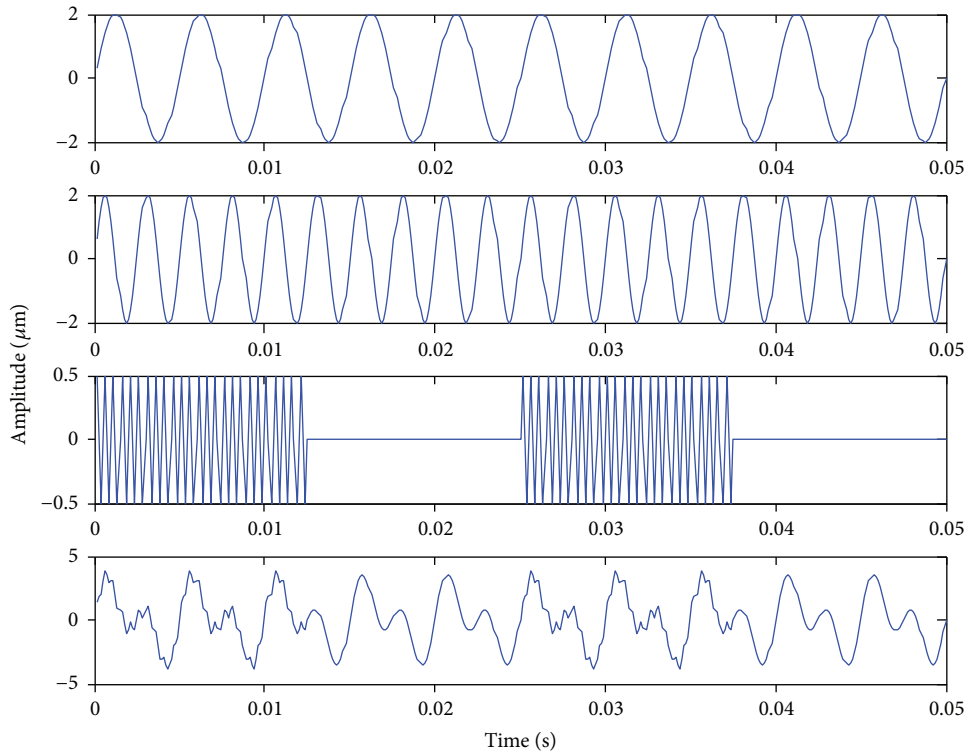


FIGURE 2: Simulation signal.

## 6. Simulation Verification

In order to verify the advantage of complex data empirical mode decomposition (CEMD) this paper proposed in dealing with mode mixing in EMD, simulations were

performed using simulation signals, considering that the mode mixing problem in EMD is usually caused by the presence of intermittent components or discontinuous components in the signal. Therefore, the two steady-state sine signals are superimposed with an intermittent signal

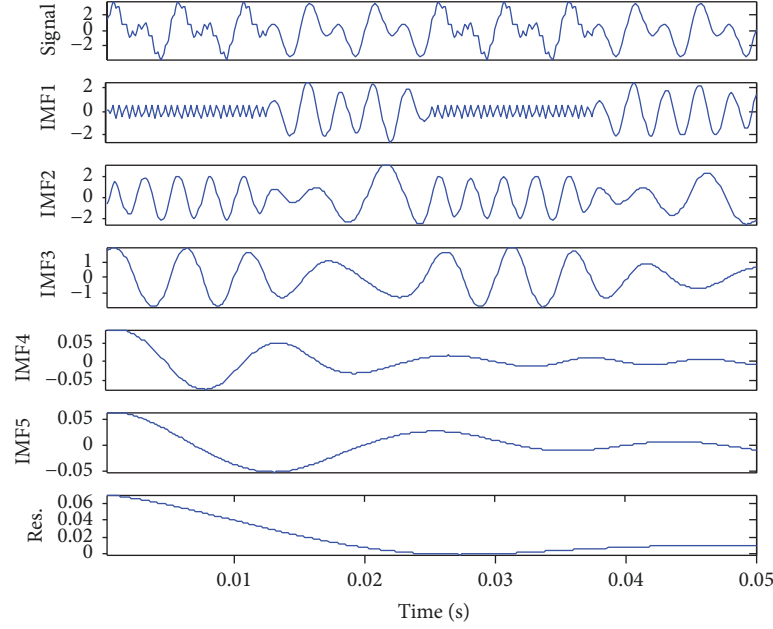


FIGURE 3: The decomposition results of EMD.

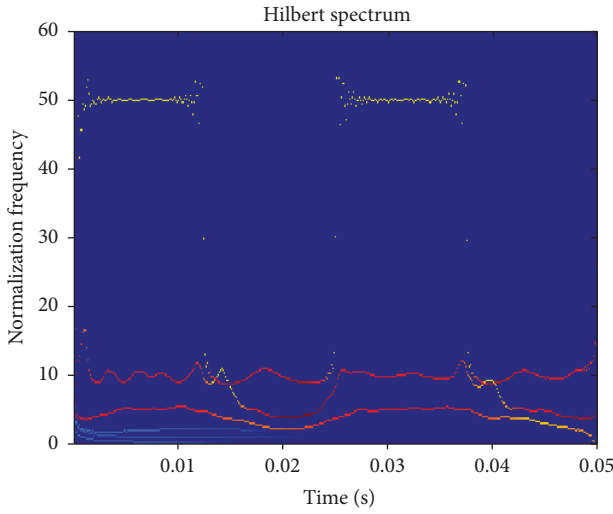


FIGURE 4: Hilbert spectrum of EMD.

to form a simulation signal. The simulation signal is shown in Figure 2, where the frequency of two steady-state sine signals is, respectively, 200 Hz and 400 Hz and the amplitude is 2. The frequency of the intermittent signal is 2000 Hz, and the amplitude is 0.5. The superimposed composite signal is shown in the lowermost layer of Figure 2.

First, the composite simulation signal is decomposed by the EMD method. The results of the decomposition are shown in Figure 3. The Hilbert spectrum of the decomposition results is shown in Figure 4. As Figure 3 has shown, the high-frequency components and low-frequency components of the original composite simulation signal

appear together in IMF1 so that the high-frequency intermittent signal is not effectively extracted. This indicates that mode mixing has occurred. Figure 4 reflects the mode mixing that exists in the EMD method.

In order to solve the problem of mode mixing in the process of EMD, this paper uses EEMD and CEMD, respectively, to decompose the composite simulation signal to compare and verify the advantages of CEMD in dealing with the problem of mode mixing. The decomposition results are shown in Figures 5–8.

Figures 5 and 7 show the decomposition results of EEMD. As shown in Figure 5, IMF1 is the high-frequency intermittent signal given. IMF2 and IMF3 are the steady-state sine signals with the frequency of 400 Hz. IMF4 is the steady-state sine signal with the frequency of 200 Hz. From the decomposition results, EEMD can solve the problem of mode mixing in EMD to some extent. However, the white noise added in the decomposition process of EEMD will also have a certain influence on the original signal. The Hilbert spectrum in Figure 7 can clearly reflect this phenomenon. In addition, the EEMD decomposition also has a serious problem of excess decomposition, such as the IMF5, IMF6, and IMF7 shown in Figure 5. The decomposition results of CEMD are shown in Figures 6 and 8. Compared with EEMD, in the decomposition process of CEMD, because the white noise is regarded as an imaginary part; it only influences the selection of the extreme point and does not add white noise to the original signal. Therefore, the method has almost no effect on the original signal. As shown in Figure 8, the decomposition result by CEMD is very clearly. This method effectively solves the problem of mode mixing existing in EMD and does not affect the original signal. In addition, as shown in Figure 6, except for the residual part, the decomposition

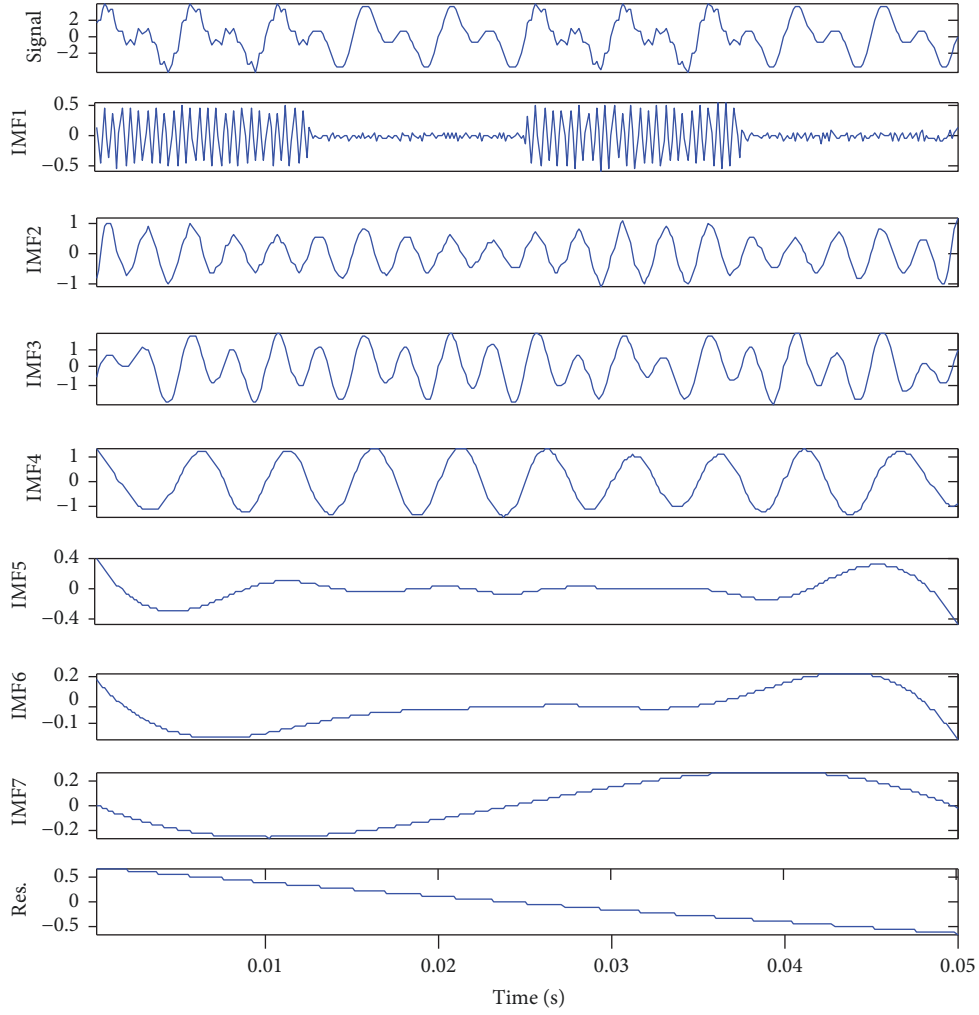


FIGURE 5: The decomposition results of EEMD.

result only remains the IMF5 component that can be negligible. Therefore, CEMD can effectively solve the problem of excessive decomposition existing in EEMD. However, this method has some shortcomings in handling high-frequency components. As shown in Figure 6, CEMD decomposes the set high-frequency signal into two components which are IMF1 and IMF2 as shown in Figure 6. But the problem can be solved by superposing the frequency components. It can be seen that the complex data empirical mode decomposition this paper proposed has a great advantage in signal processing.

## 7. Analysis of Rolling Bearing Faults

To verify the effectiveness and efficiency of the proposed method in practical applications, this paper uses the data from the Case Western Reserve University (CWRU) Bearing Data Center website [35] to conduct experimental analysis. The motor speed is  $f_r = 1750$  rpm (29.17 Hz), and the sampling frequency is 12 kHz.

According to the multidomain fault diagnosis method proposed in this paper based on CEMD and RF, 11 time domain feature vectors, 13 frequency domain feature vectors, and time-frequency feature vectors obtained by the CEMD of the collected vibration signals are calculated. Combine all above feature vectors into a vector set  $\mathbf{X}$ ,  $\mathbf{X} = [T_1, \dots, T_{11}, F_1, \dots, F_{13}, E_1, \dots, E_n, H_{EN}]$ . And mark the fault status of the corresponding bearing  $C = [0, 1, 2, 3]$  as the input of classifier. A total of 480 sets of data were selected, 400 sets of data were trained, and 80 sets of data were used as out-of-pocket data (OOB) to verify the classification accuracy.

First, the bearing fault classification is classified. The four states of health, inner ring fault, outer ring fault, and ball fault of the corresponding bearing are, respectively, regarded as the input of the classifier  $C = [0, 1, 2, 3]$ . When only 24 feature vectors of time domain and frequency domain are used to classify the failure, the results of the classification are shown in Figure 9. Figure 9 shows that the classification accuracy of the method this paper proposed can reach 100% when less than 20 decision trees are generated.



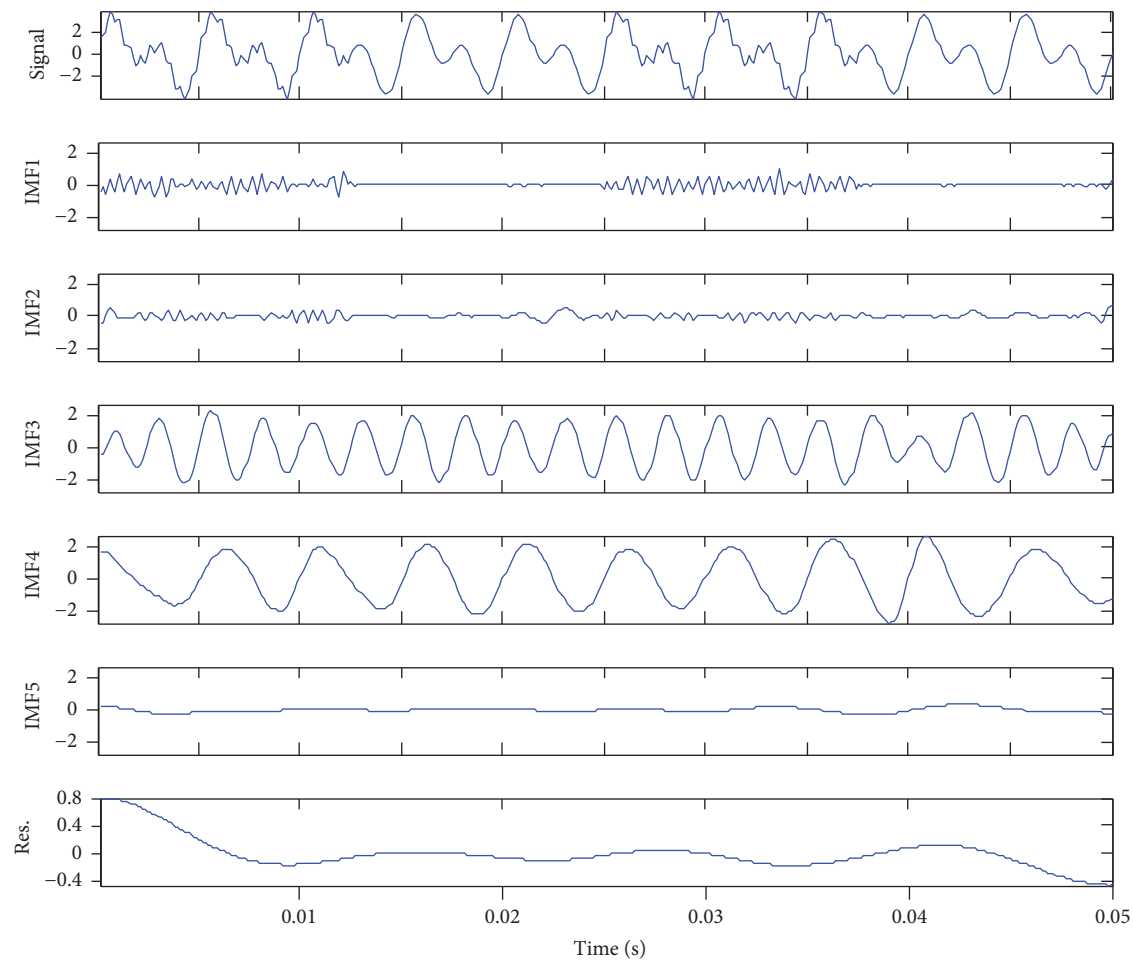


FIGURE 6: The decomposition results of CEMD.

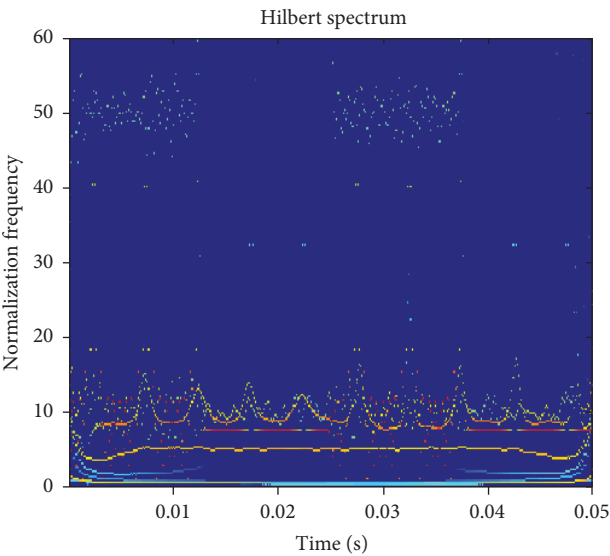


FIGURE 7: Hilbert spectrum of EEMD.

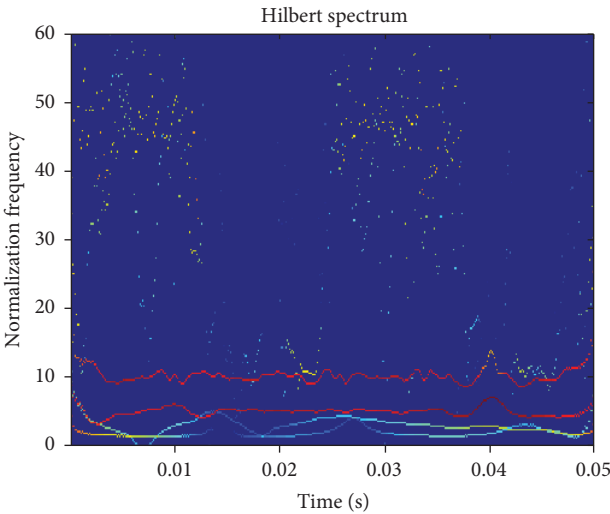


FIGURE 8: Hilbert spectrum of CEMD.

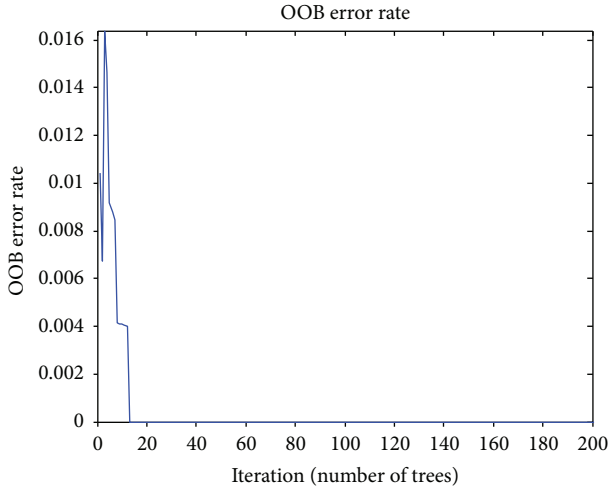


FIGURE 9: Diagnosis results of different fault categories.

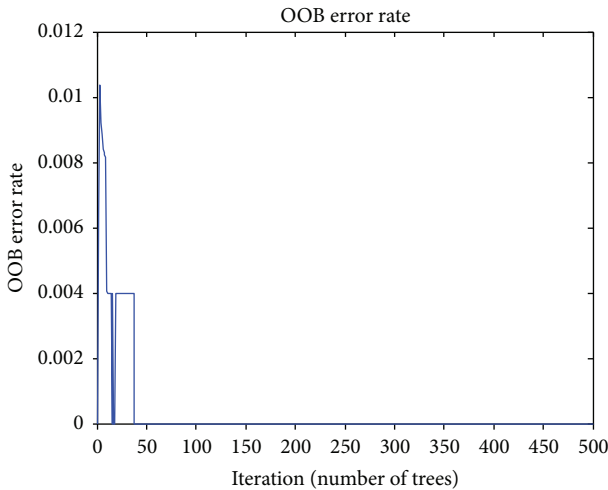


FIGURE 10: Diagnostic results at different levels of the same fault.

Then, classify the different fault levels of the same fault. This paper takes the inner ring fault of the rolling bearing as an example. The four states of health, fault with the size of 0.1778 mm, fault with the size of 0.3556 mm, and fault with the size of 0.5334 mm of the corresponding bearing are regarded as the input of the classifier  $C = [0, 1, 2, 3]$ . When only 24 feature vectors of time domain and frequency domain are used to classify the failure, the results of the classification are shown in Figure 10. Figure 10 shows that the classification accuracy of the method this paper proposed can reach 100% when less than 50 decision trees are generated.

Finally, classify the different load conditions of the same fault. This paper takes the inner ring fault of rolling bearing as an example. The ball bearings of the type 6205-2RS SKF on the motor shaft end were artificially made a fault point with the size of 0.1778 mm. The four load statuses of load status 0 (1797 r/min), load status 1 (1772 r/min), load status 2 (1750 r/min), and

load status 3 (1730 r/min) are regarded as the input of the classifier. The results of the classification are shown in Figure 11.

Figure 11(a) shows the diagnostic results by 24 feature vectors of time domain and frequency domain of vibration signal. Figure 11(b) shows the diagnostic results by multidomain feature vectors of the vibration signal. In comparison, the misjudgment rate of diagnosis results by 24 feature vectors of time domain and frequency domain is basically stable at 0.25. That is to say, the accuracy of diagnosis is stable at about 75%. However, the misjudgment rate of multidomain feature diagnosis can be reduced to less than 0.2; that is, the accuracy of the diagnostic results is stable at more than 80%. Therefore, the accuracy of fault diagnosis can be improved by using the CEMD method and taking the feature vectors of the time-frequency domain into account in fault diagnosis.

However, there is some information redundancy in the feature vectors including time domain, frequency domain, and time-frequency domain. In order to solve the problem of information redundancy caused by increasing the feature vectors, this paper calculates the importance of all the feature vectors based on the random forest theory. The results are shown in Figure 12. This paper uses the mean decrease in accuracy and the mean decrease in the Gini index to measure the importance of feature vectors as shown in Figure 12. Taking into account the size of the two indicators of each feature vector, the feature vectors which two indicators are all smaller are eliminated. Then, the random forest classification model is reconstructed by the feature vectors that have been selected. The finally fault diagnosis results are shown in Figure 13. The misjudgment rate of diagnosis results can be reduced to less than 0.15 by removing redundant feature vectors. That is to say, the accuracy of diagnosis is stable at more than 80%.

To further illustrate the effectiveness of the proposed method, this paper makes a comparison of diagnosis results among this paper proposed and genetic algorithm back propagation (GA-BP) neural network, support vector machines (SVM), and extreme learning machine (ELM). The diagnosis accuracy of the three methods is obtained by using time domain and frequency domain feature vectors and multidomain feature vectors. For comparison, EEMD is utilized to decompose the original signal into IMFs. This paper separately classifies vibration signals including the difference bearing fault classification, the difference fault degree, and the difference load condition. The diagnosis accuracy of every method is shown in Table 3.

As shown in Table 3, not the three traditional methods but random forest, the diagnosis accuracy by using multidomain feature vectors is much higher than the diagnosis accuracy by using time domain and frequency domain feature vectors. Therefore, utilizing multidomain feature vector can improve the accuracy of fault diagnosis for the wind turbine. And compared with the three traditional methods the GA-BP, SVM, and

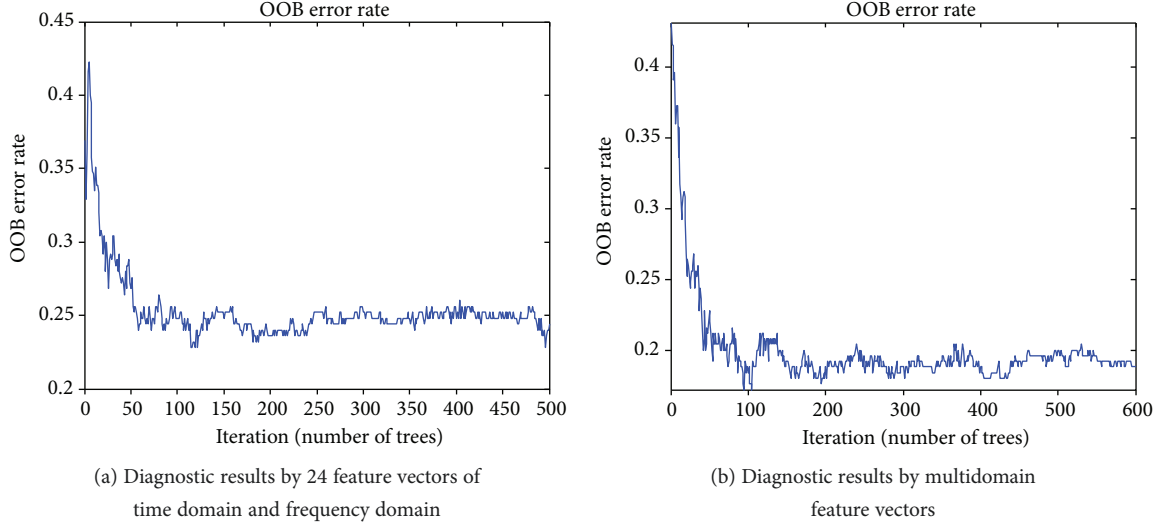


FIGURE 11: Diagnostic results under different load conditions of the same fault.

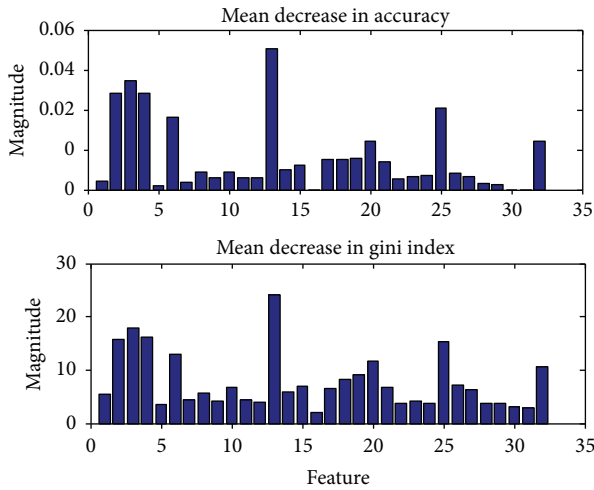


FIGURE 12: Estimation of the importance of all characteristic parameters.

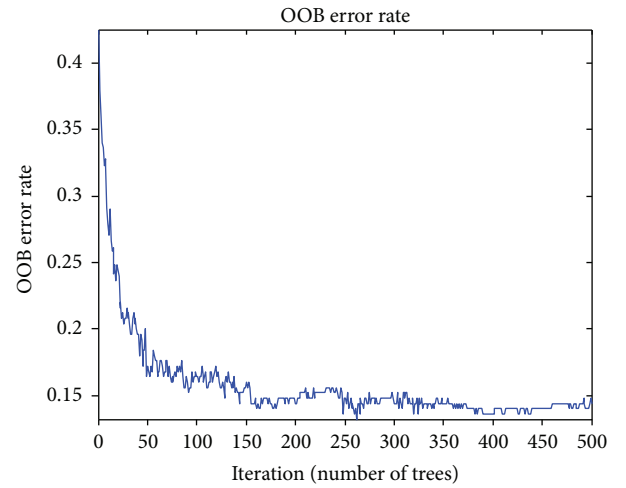


FIGURE 13: Fault diagnosis results after eliminating redundant information.

ELM methods, the multidomain feature fault diagnosis method based on CEMD-RF proposed in this paper has a higher accuracy of diagnosis for the wind turbine. In addition, the diagnosis accuracy of CEMD-RF is higher than CEMD. Thus, the CEMD method that this paper proposed can improve the diagnosis accuracy.

## 8. Conclusions

This paper proposes a multidomain feature based on complex empirical mode decomposition (CEMD) and random forest theory (RF) for the problems of mode mixing existing in EMD methods and feature information redundancy in the multidomain fault diagnosis method. The specific conclusions are as follows:

- (1) Complex empirical mode decomposition (CEMD) eliminates the mode mixing in the integration average process of EEMD caused by adding white noise. Therefore, compared with the result of EEMD, the decomposition results of CEMD have a clearer IMF spectrum and have no effect on the original signal.
- (2) Increasing the number of feature vectors can effectively improve the fault diagnosis accuracy of the wind turbine. By utilizing time domain, frequency domain, and time-frequency domain feature vectors of the wind turbine vibration signal, the method this paper proposed can comprehensively extract the fault information of the unit and effectively improve the fault diagnosis accuracy of the wind turbine.

TABLE 3: The diagnosis accuracy of every method.

Classification method	GA-BP	SVM	ELM	EEMD-GABP	EEMD-SVM	EEMD-ELM	RF	EEMD-RF	CEMD-RF
Fault classification	0.953	0.975	0.952	0.961	0.984	0.973	1	1	1
Fault degree	0.932	0.944	0.913	0.938	0.962	0.956	1	1	1
Load condition	0.678	0.773	0.731	0.721	0.804	0.791	0.752	0.826	0.855

- (3) The multifault diagnosis method based on the CEMD-RF proposed in this paper can effectively solve the problem of information redundancy in multifeature fault diagnosis. By eliminating redundant feature vectors based on the calculated importance of each feature vector, this method can effectively improve the fault diagnosis accuracy.

## Data Availability

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

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (no. 51779206) and the Science and Technology Projects of Water Resources Department of Shaanxi Province (no. 2015slkj-04). The authors thank the support of the National Natural Science Foundation of China. And the authors would like to thank CWRU for providing the free bearing data download.

## References

- [1] Z. Hameed, Y. S. Hong, Y. M. Cho, S. H. Ahn, and C. K. Song, "Condition monitoring and fault detection of wind turbines and related algorithms: a review," *Renewable and Sustainable Energy Reviews*, vol. 13, no. 1, pp. 1–39, 2009.
- [2] B. Lu, Y. Li, X. Wu, and Z. Yang, "A review of recent advances in wind turbine condition monitoring and fault diagnosis," in *2009 IEEE Power Electronics and Machines in Wind Applications*, pp. 1–7, Lincoln, NE, USA, June 2009.
- [3] D. Yu, Z. M. Chen, K. S. Xiahou, M. S. Li, T. Y. Ji, and Q. H. Wu, "A radically data-driven method for fault detection and diagnosis in wind turbines," *International Journal of Electrical Power & Energy Systems*, vol. 99, pp. 577–584, 2018.
- [4] W. Teng, X. Ding, Y. Zhang, Y. Liu, Z. Ma, and A. Kusiak, "Application of cyclic coherence function to bearing fault detection in a wind turbine generator under electromagnetic vibration," *Mechanical Systems and Signal Processing*, vol. 87, pp. 279–293, 2017.
- [5] T. Wang, F. Chu, and Q. Han, "Fault diagnosis for wind turbine planetary ring gear via a meshing resonance based filtering algorithm," *ISA Transactions*, vol. 67, pp. 173–182, 2017.
- [6] N. A. Ahmed and M. Cameron, "The challenges and possible solutions of horizontal axis wind turbines as a clean energy solution for the future," *Renewable and Sustainable Energy Reviews*, vol. 38, pp. 439–460, 2014.
- [7] R. Bi, C. Zhou, and D. M. Hepburn, "Detection and classification of faults in pitch-regulated wind turbine generators using normal behaviour models based on performance curves," *Renewable Energy*, vol. 105, pp. 674–688, 2017.
- [8] X. Liu, X. Xu, Z. Jiang, G. Wu, and Y. Zuo, "Application of the state deterioration evolution based on bi-spectrum entropy and HMM in wind turbine," *Chaos, Solitons & Fractals*, vol. 89, pp. 160–168, 2016.
- [9] Z. Feng, S. Qin, and M. Liang, "Time-frequency analysis based on Vold-Kalman filter and higher order energy separation for fault diagnosis of wind turbine planetary gearbox under non-stationary conditions," *Renewable Energy*, vol. 85, pp. 45–56, 2016.
- [10] Y. Feng, J. R. Cao, B. C. Lu, and D. F. Zhang, "Fault feature extraction method for rolling bearing based on wavelet transform optimized by continuous kurtosis," *Journal of Vibration and Shock*, vol. 34, no. 14, pp. 27–32, 2015.
- [11] W. Teng, X. Ding, X. Zhang, Y. Liu, and Z. Ma, "Multi-fault detection and failure analysis of wind turbine gearbox using complex wavelet transform," *Renewable Energy*, vol. 93, pp. 591–598, 2016.
- [12] L. Saidi, J. B. Ali, and F. Fnaiech, "Bi-spectrum based-EMD applied to the non-stationary vibration signals for bearing faults diagnosis," *ISA Transactions*, vol. 53, no. 5, pp. 1650–1660, 2014.
- [13] I. Antoniadou, G. Manson, W. J. Staszewski, T. Barszcz, and K. Worden, "A time-frequency analysis approach for condition monitoring of a wind turbine gearbox under varying load conditions," *Mechanical Systems and Signal Processing*, vol. 64–65, pp. 188–216, 2015.
- [14] M. Zhang, Z. Jiang, and K. Feng, "Research on variational mode decomposition in rolling bearings fault diagnosis of the multistage centrifugal pump," *Mechanical Systems and Signal Processing*, vol. 93, no. 460, pp. 460–493, 2017.
- [15] S. Liu, G. Tang, X. Wang, and Y. He, "Time-frequency analysis based on improved variational mode decomposition and Teager energy operator for rotor system fault diagnosis," *Mathematical Problems in Engineering*, vol. 2016, Article ID 1713046, 9 pages, 2016.
- [16] Q. W. Gao, W. Y. Liu, B. P. Tang, and G. J. Li, "A novel wind turbine fault diagnosis method based on intergral extension load mean decomposition multiscale entropy and least squares support vector machine," *Renewable Energy*, vol. 116, pp. 169–175, 2018.
- [17] V. Muralidharan and V. Sugumaran, "Rough set based rule learning and fuzzy classification of wavelet features for fault diagnosis of monoblock centrifugal pump," *Measurement*, vol. 46, no. 9, pp. 3057–3063, 2013.

- [18] X. Jiao, B. Jing, Y. Huang, J. Li, and G. Xu, "Research on fault diagnosis of airborne fuel pump based on EMD and probabilistic neural networks," *Microelectronics Reliability*, vol. 75, pp. 296–308, 2017.
- [19] B. Tang, T. Song, F. Li, and L. Deng, "Fault diagnosis for a wind turbine transmission system based on manifold learning and Shannon wavelet support vector machine," *Renewable Energy*, vol. 62, pp. 1–9, 2014.
- [20] M. Gan, C. Wang, and C. Zhu, "Multiple-domain manifold for feature extraction in machinery fault diagnosis," *Measurement*, vol. 75, pp. 76–91, 2015.
- [21] Z. Shen, X. Chen, X. Zhang, and Z. He, "A novel intelligent gear fault diagnosis model based on EMD and multi-class TSVM," *Measurement*, vol. 45, no. 1, pp. 30–40, 2012.
- [22] C. Li, R. V. Sanchez, G. Zurita, M. Cerrada, D. Cabrera, and R. E. Vásquez, "Gearbox fault diagnosis based on deep random forest fusion of acoustic and vibratory signals," *Mechanical Systems and Signal Processing*, vol. 76–77, pp. 283–293, 2016.
- [23] M. Cerrada, G. Zurita, D. Cabrera, R. V. Sánchez, M. Artés, and C. Li, "Fault diagnosis in spur gears based on genetic algorithm and random forest," *Mechanical Systems & Signal Processing*, vol. 70–71, pp. 87–103, 2016.
- [24] M. W. Ahmad, M. Mourshed, and Y. Rezgui, "Trees vs neurons: comparison between random forest and ANN for high-resolution prediction of building energy consumption," *Energy and Buildings*, vol. 147, pp. 77–89, 2017.
- [25] T. Tanaka and D. P. Mandic, "Complex empirical mode decomposition," *IEEE Signal Processing Letters*, vol. 14, no. 2, pp. 101–104, 2007.
- [26] M. U. B. Altaf, T. Gautama, T. Tanaka, and D. P. Mandic, "Rotation invariant complex empirical mode decomposition," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, pp. 1009–1012, Honolulu, HI, USA, April 2007.
- [27] G. Rilling, P. Flandrin, P. Goncalves, and J. M. Lilly, "Bivariate empirical mode decomposition," *IEEE Signal Processing Letters*, vol. 14, no. 12, pp. 936–939, 2007.
- [28] Y. Lei, Z. He, and Y. Zi, "Application of an intelligent classification method to mechanical fault diagnosis," *Expert Systems with Applications*, vol. 36, no. 6, pp. 9941–9948, 2009.
- [29] L. Breiman, "Random forest," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [30] T. Han and D. Jiang, "Rolling bearing fault diagnostic method based on VMD-AR model and random forest classifier," *Shock and Vibration*, vol. 2016, Article ID 5132046, 11 pages, 2016.
- [31] L. Breiman, "Using iterated bagging to Debias regressions," *Machine Learning*, vol. 45, no. 3, pp. 261–277, 2001.
- [32] A. Cutler, D. R. Cutler, and J. R. Stevens, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 157–175, 2012.
- [33] N. Morizet, N. Godin, J. Tang, E. Maillet, M. Fregonese, and B. Normand, "Classification of acoustic emission signals using wavelets and random forests: application to localized corrosion," *Mechanical Systems and Signal Processing*, vol. 70–71, pp. 1026–1037, 2016.
- [34] E. E. Tripoliti, D. I. Fotiadis, and G. Manis, "Automated diagnosis of diseases based on classification:dynamic determination of the number of trees in random forests algorithm," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 4, pp. 615–622, 2012.
- [35] <http://csegroups.case.edu/bearingdatacenter/pages/12k-drive-end-bearing-fault-data>.



