International Journal for Housing Science and Its Applications
Publish August 10, 2025. Volume 46, Issue 4 Pages 5568-5577

https://doi.org/10.70517/ijhsa464396

Research on Adaptive Random Forest Method for Fault Diagnosis of High-Noise Fan Gearboxes

Lipeng Cui[1,2], Yu Yu[3], Mingzhu Tang[3,*], Zhao Wang[4] and Jianyou Ouyang[4]
[1] School of Automation and Electrical Engineering, Tianjin University of Technology and Education, Tianjin, 300222, China
[2] School of Electronic Information and Automation, Tianjin Light Industry Vocational Technical College, Tianjin, 300350, China
[3] School of Energy and Power Engineering, Changsha University of Science & Technology, Changsha, Hunan, 410114, China
[4] Department of Energy Technology, Changsha Electric Power Technical College, Changsha, Hunan, 410131, China
Corresponding authors: (e-mail: tmz@csust.edu.cn).

Abstract Wind turbines operate in harsh environments for a long time, and the gearbox as a core transmission component faces severe reliability challenges. Aiming at the problem of low fault diagnosis accuracy of wind turbine gearbox in high noise environment, this study proposes a fault diagnosis method based on adaptive probabilistic random forest. The method firstly adopts the improved global projection algorithm for feature extraction and dimension reduction of gearbox operation data, which effectively retains the local structural information while taking into account the global features; then introduces the quantum wolf pack optimization algorithm to adaptively optimize the key parameters of the random forest, and constructs an adaptive probabilistic random forest classifier; and finally improves the fault identification capability through the multi-channel data fusion technology. The experimental results based on vibration data of 20 noisy wind turbine gearboxes show that the proposed method performs well in the identification of four states, namely, healthy state, secondary planetary gear ring wear, sun wheel crushing, and primary planetary gear ring wear. The fault identification accuracy after the fusion of both directions reaches 97.17%, which is significantly improved compared with 93.33% in the single X-direction and 95.17% in the Y-direction. Compared with the traditional method, the fault identification rate of this method reaches 92%, which is significantly better than the 84% of the support vector machine and the 89% of the traditional random forest, proving the effectiveness and superiority of the proposed method in the fault diagnosis of gearboxes of high-noise wind turbines.

Index Terms adaptive probabilistic random forest, localization projection algorithm, quantum wolf pack optimization, multi-channel data fusion, fault diagnosis, wind turbine gearboxes

I. Introduction

In recent years, with the construction of wind farms, a large number of wind turbines have been put into operation, and the related failure problems have emerged [1]. Wind turbines are often located in high mountains and oceans and other areas with abundant wind energy resources but inconvenient inspections, and failures cannot be found in time. Among them, the gearbox, as a key component of the wind turbine transmission system, will pose a serious threat to the normal operation of the wind turbine once a failure occurs, and may even lead to serious safety accidents and major economic losses [2]-[5]. Therefore, condition monitoring and fault diagnosis of wind turbine gearboxes have become important issues [6].

Due to the complexity of the wind turbine gearbox working environment and strong background noise, in order to extract as much effective information as possible and reduce the noise interference of wind turbine gearbox vibration signals, American scholars Dragomiretskiy and Zosso proposed a new adaptive signal variational modal decomposition model in 2013 [7]. The VMD model is a completely non-recursive signal decomposition algorithm, which, in the process of calculation The VMD model is a completely non-recursive signal decomposition algorithm, which abandons the constraints of recursive decomposition, effectively avoids the modal aliasing problem and the white noise residual problem, improves the speed of signal decomposition, and has a good effect on the processing of non-smooth and non-linear signals, and is therefore widely used in the fields of noise reduction, fault diagnosis and identification and classification [8]-[10]. The VMD model has a strong dependence on the number of decomposition layers (K) and the penalization factor (α), and a smaller K leads to the under-decomposition of vibration signals, resulting in modal aliasing problems and the creation of a modal aliasing problem. A small K will lead to under-decomposition of the vibration signal, resulting in modal aliasing, while a large K will lead to over-decomposition of the vibration signal, resulting in modal loss [11]. α is too large to lead to a decrease in the bandwidth of the eigenmode components obtained from decomposition, resulting in missing information, while α is too small to lead to an increase in the bandwidth of the eigenmode components obtained from decomposition, resulting in center-frequency overlap and modal confusions [12], [13]. In order to optimize the parameters K and α in the VMD model, Li et al. proposed a wind turbine gearbox fault type identification algorithm combining the VMD with an extreme learning machine to find the optimal K and perform the modal decomposition, which resulted in an effective improvement of the fault identification accuracy [14].

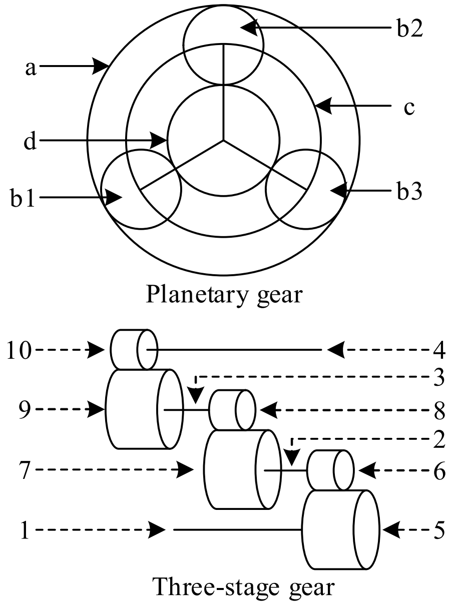
In order to classify the feature vectors so as to realize the fault diagnosis of wind turbine gearboxes, the use of random forests in machine learning is a more widely used method currently [15]. For example, Cabrera et al. developed a random forest classifier for multi-class fault diagnosis of gearboxes using wavelet packet decomposition technique, which achieved a classification accuracy of 98.68% and demonstrated high efficiency and reliability [16]. Chen et al. proposed an improved random forest algorithm, which combines graph-based semi-supervised learning and decision trees, aiming to improve the transmission fault diagnosis in terms of classification accuracy, especially for the case of insufficient number of labeled samples, to achieve the optimization of diagnostic efficiency [17]. Qin et al. combined the integrated empirical modal decomposition (EEMD) and random forest for an accurate diagnostic method of rolling bearing faults, which is more accurate than the wavelet method in terms of diagnostic efficiency [18]. Yuan et al. proposed a change-point detection based on SCADA data and a stacked model for wind turbine gearbox fault prediction, which was validated on historical data from five wind turbines to significantly improve fault detection and reduce downtime [19]. The above methods in the optimization algorithm often have the problem of uneven distribution of population initialization and easy to appear local optimum, which leads to the optimization algorithm is inaccurate for the random forest parameter optimization, which in turn affects the accuracy of the classification and identification of the random forest algorithm [20], [21]. The combination of VMD model and random forest is a new idea to solve the above problems, however, there are fewer related studies.

In this study, a fault diagnosis method for high-noise wind turbine gearboxes integrating multiple advanced algorithms is proposed. Firstly, for the problem of difficult extraction of signal features in high noise environment, an improved global projection algorithm is used to reduce and extract features from the original vibration data, which takes into account the global information while maintaining the local structure of the data to effectively improve the separability of the features. Secondly, in order to solve the problem of difficult parameter setting of the random forest algorithm, the quantum wolf pack optimization algorithm is introduced to realize the adaptive optimization of the parameters, and the convergence and robustness of the algorithm are improved by the quantum bit coding and dynamic adaptive rotation angle strategy. Finally, the multi-channel data fusion technology is used to integrate the vibration information in different directions to construct an adaptive probabilistic random forest classifier to realize the accurate recognition of multiple failure modes.

II. Composition and failure characteristics of fan gearboxes

II. A. Components of a wind turbine gearbox

The wind turbine gearbox is the most valuable component in a wind turbine, and its function is to convert the low-speed rotation of the blades into the high-speed rotation required by the wind turbine generator, and to transfer the kinetic energy of the blades to the generator. WTGs usually use a three-stage planetary gearbox. The three-stage planetary gearbox consists of planetary gears and three-stage gears, three-stage planetary gearbox composition is shown in Figure 1. The planetary gear consists of an internal toothed inner gear ring a, three planetary wheels b1, b2, b3, planetary gear rack c and sun wheel d. The inner gear ring a is fixed, and the planetary wheels are fixed on the planetary gear rack, which is connected to the main shaft and has the same rotational speed as that of the fan blades. Three planetary wheels rotate in the inner gear ring, which increases the rotational speed of the sun wheel. The three-stage gearing consists of gear sets of three classes: low-speed class 5, 6, medium-speed class 7, 8 and high-speed class 9, 10. The low-speed class large gear 5 is directly connected to the shaft 1, which is driven by the sun wheel, and the shaft 4, which is connected to the generator shaft. By accelerating the sun wheel in three stages, the shaft 4 reaches the required speed of the generator.



Planetary gear

Three-stage gear

Figure 1: The composition of the three-stage planetary gearbox

II. B.Wind turbine gearbox failure modes

Fan gearbox failure modes are mainly: insufficient lubrication, vibration and swaying, high temperature, aging failure of bearings and gears. Insufficient lubrication is mainly due to factors such as leakage of lubricant and excessive particulate content in the lubricant. The causes of vibration wobble are: design defects, gears and shafts not aligned, installation defects and foreign objects entering, etc., which can be judged by installing a vibration monitoring system. In order to avoid these failures, it is necessary to regularly overhaul the gearbox. Gears and bearings are in high speed rotation during operation, which are subject to severe wear and aging. Aging failures of gears, medium-speed bearings, and high-speed bearings will cause damage to the entire transmission, and the cost of failure is high. The strategy for overhauling the aging components is usually to replace them periodically, therefore, a method for troubleshooting the gearbox of a high-noise wind turbine is proposed.

III. Noisy wind turbine gearbox troubleshooting methods

In order to reduce the waste of human and material resources due to minor faults, and also to prevent secondary damage to wind turbines due to untimely handling of major faults, this paper is based on the adaptive probabilistic random forest algorithm for the fault diagnosis of gearboxes of high-noise wind turbines.

III. A. Feature extraction of operational data

The local projection-preserving (LPP) algorithm is a linear approximation of the Laplace feature mapping (LE) algorithm to preserve local information in the data, which allows the data to maintain an internally fixed local structure while decreasing the dimensionality. Compared with the global information retained by principal component analysis, the local information can better highlight the feature differences between different fault data. Meanwhile the algorithm obtains the locally preserved projection by finding the optimal approximation of the eigenfunction of the Laplace-Beltrami operator on the manifold. Therefore, the algorithm also has the data representation properties of nonlinear techniques.

The operating data of noisy wind turbine gearboxes has the characteristics of nonlinearity as well as strong information coupling, high noise content, and many irrelevant parameters, which makes the fault information of the data difficult to utilize in fault diagnosis. In order to improve this situation, this paper introduces a localization projection algorithm to reduce the dimensionality of the gearbox operation data and extract the fault information before carrying out the fault diagnosis of the gearbox, so as to improve the accuracy of the fault diagnosis.

In the given initial data space, the data set $X=[x_1,x_2,\cdots x_n]$ has m parameter dimensions, and the data set $Y=[y_1,y_2,\cdots y_n]$ in the dimensionalized space is obtained by transforming the matrix A and calculating $Y=A^TX$, so that the dimensionality of its data, m', is much smaller than m to achieve the dimensionality reduction purpose. Considering that the original algorithm only focuses on the local information of the data and does not take into account the global structure. The objective function of the improved algorithm introduces the global concept to minimize the following objective function under certain constraints:

$$\sum_{ij}(y_i - y_j)^2 W_{ij} - \frac{1}{n}(y_i - \overline{y})^2 \tag{1}$$

where i, j represent two different data in the initial space X, W_{ij} represents the matrix consisting of the distance weight coefficients between the two data points i, j, and at the same time the computational process to prevent the elimination of arbitrary scaling factors, the following restriction formulae are added to the computational process:

$$A^T XDX^T A = 1 \tag{2}$$

At this point minimization becomes a problem of finding eigenvalues:

$$(B - XLX^T)A = \lambda XDX^T A \tag{3}$$

where B is $(\sum_{i=1}^{N}(x_i - \overline{x})(x_i - \overline{x})^T)/N$, L is the Laplacian matrix, with $L = D - W$, and D is the diagonal matrix, which is used to provide the natural measure of the data points, and the D in which the The y_i corresponding to the maximum value D_{ii} is the most important data value. $D_{ii} = \sum_j W_{ji}$ This solves the difficult problem of finding generalized eigenvalues by solving for the solution of the smallest eigenvalue. The steps of the algorithm are:

(1) Construct the adjacency graph, construct the data adjacency graph by the $k-$ nearest neighbor algorithm, if the data x_j is in the range of x_i's k nearest neighbor data, or x_i is in the range of x_j's k nearest neighbor data, then connect a relation line at the data points x_i and x_j.

(2) Select the weight of the relationship line, the relationship line connected between the nodes of the adjacency graph is weighted, generally through the heat kernel function to calculate the weighted value W_{ji} of the connection line between the nodes, the result of the calculation so that the larger the distance between the data of x_i and x_j, the smaller the corresponding weighted value W_{ji}. Conversely the smaller the distance between the data, the larger the corresponding weighted value W_{ji}, the heat kernel function equation is defined as:

$$W_{ij} = e^{-\frac{\|x_i - x_j\|^2}{t}} \tag{4}$$

where t is the heat kernel constant, which usually defaults to 1.

(3) Calculate the projection matrix, according to Eq. (3), the obtained eigenvector solutions are arranged as $a_0, a_1, a_2, \cdots a_{n-1}$ in ascending order, the eigenvectors correspond to the positional arrangement of the eigensolutions, and finally, the desired reduced dimensional column vectors are selected to form the transformation matrix A.

III. B. Adaptive Probabilistic Random Forest Approach

Random forest is the most important method in integrated learning. The parameters of the number of decision trees constructed in the random forest algorithm t and the proportion of node feature set selection f need to be set values at the beginning of the algorithm, and different values of the parameters have a significant effect on the classification results of the random forest. This algorithm introduces the quantum wolf pack optimization algorithm to perform adaptive optimization of the parameters of the random forest algorithm to obtain the best values for parameter initialization.

III. B. 1) Random Forest Algorithm

In this paper, a random forest model is used for failure analysis of gearboxes of high noise wind turbines. Random forest is a learning method that utilizes an ensemble of multiple decision trees to improve the accuracy and robustness of predictions by aggregating their outputs. A decision tree is a tree structure where each node represents a test condition for an attribute and each leaf node represents a category or a value. The learning process focuses on the splitting of nodes with the aim of dividing the dataset into purer subsets by attribute selection. When there is a new sample to be predicted, Random Forest will let all the decision trees judge it and then determine the final result based on majority vote or average. There are two main principles of random forests: self-sampling and random feature selection. Random forests use self-sampling to take some samples from the original training set with put-back, so that each tree has a different training set, increasing the diversity of trees and reducing the risk of overfitting. At the same time, the unsampled samples are used as a test set for the tree to evaluate the tree's generalization ability, which is called out-of-bag data. Random forests randomize some of the features from all the features while constructing the decision tree and then select the optimal features from them for splitting, which is called random subspace method. This reduces the correlation of features, increases the independence of the tree and improves the integration.

Let the training set be $P=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$, where x_i is an M -dimensional feature vector, and y_i is the corresponding category or value. The random forest contains K decision trees, and each tree is generated as follows: first, n samples are randomly drawn from P in a relaxed manner as the training set P_k for the k th tree. Then, m features are randomly drawn from M features in a relaxed manner as the candidate feature set F_k for the k th tree. Finally, P_k and F_k are used to construct the k th tree T_k. Each time a node is split, a feature from F_k is randomly selected and the best split point is determined based on some criterion (e.g., information gain, Gini index, etc.). The above steps are repeated until K trees are generated.

In case of a classification problem, the formula is:

$$y=\arg\max\sum_{k=1}^{K}\sum_{i=1}^{N}I(T_k(x)=c_i) \tag{5}$$

where K is the number of decision trees, I is the indicator function, $T_k(x)$ is the classification result of the k th tree on the input x, and N is the number of categories c_i is the category label. The meaning of the indicator function is that it returns 1 if the condition in the parentheses holds, and 0 otherwise. e.g., $I(T_1(x)=c_1)$ is 1 if the 1st tree classifies x into category c_1, and 0 otherwise. for a new input x, the Random Forest will let all the K decision trees categorize it, and then tally up the number of votes for each category c. That is, how many trees classify x into c_i categories. Finally, the random forest chooses the category with the most votes as the final prediction y.

If it is a regression problem, the formula is:

$$y=\frac{1}{k}\sum_{k=1}^{K}T_k(x) \tag{6}$$

where $T_k(x)$ is the regression result of the k th tree on x. For a new input x, the random forest will let all the K decision trees regress on it, and each tree will give a result with the value $T_k(x)$. The random forest then computes the average of the results from all the trees as the final prediction y.

In this paper, we optimize the random forest algorithm using an improved quantum wolf pack optimization algorithm to obtain an adaptive probabilistic random forest algorithm.

III. B. 2) Quantum Wolfpack Algorithm Improvement

The basic idea of the wolf pack algorithm is derived from the behavior of wolf pack groups, the algorithm has good search and development ability, but its existence is due to the randomness of the update mechanism leads to an increase in the uncertainty of the wolf pack population, which reduces the algorithm's generalization ability and robustness. The introduction of quantum computing increases the optimization performance of the wolf pack algorithm, controls the convergence speed of the algorithm, and strives to achieve better optimization results.

The core idea of the quantum wolf pack algorithm is to represent the artificial wolf in the form of multiple quantum bits, and use the quantum revolving door to change the positional transformation of the wolf pack and adjust its hunting behavior, and the optimization-seeking performance is further improved. In the population updating behavior of the wolf pack algorithm, the way of eliminating some poorly evaluated artificial wolves and then randomly producing the same number of artificial wolves is used, which is easy to make the algorithm fall into the local optimal solution and lack of robustness. This study improves on this foundation, when the wolves reach the step of population updating behavior, evaluate and rank the position of the wolves, eliminate the γ part of the artificial wolves ranked at the back, take the same number of artificial wolves ranked at the front as the basic samples, and update these samples using quantum revolving door, which increases the diversity of the algorithm while keeping the number of wolves unchanged, and improves the algorithm's Robustness.

Quantum computation consists of quantum bit encoding and quantum revolving door updating, which utilizes the uncertainty of quantum states for computation and ensures the diversity of quantum computation. Therefore, combining quantum computing with optimization algorithms can improve the diversity, convergence, generalization ability and robustness of optimization algorithms.

Quantum bit encoding contains information about the independent variables in the algorithm, and quantum bit encoding differs from normal optimization algorithms in that it can contain information about multiple quantum states simultaneously. Usually quantum bits are represented as $|\varphi\rangle = \alpha|0\rangle + \beta|1\rangle$, where (α, β) is the probability amplitude of $|0\rangle$ and $|1\rangle$, respectively, and satisfies $|\alpha|^2 + |beta|^2 = 1$, and $|0\rangle$ and $|1\rangle$ denote spin states. So a quantum state contains both $|0\rangle$ and $|1\rangle$ different information, and the bit encoding formula is:

$$q_j^t = \begin{bmatrix} \alpha_{11}^t & \alpha_{12}^t & \cdots & \alpha_{1k}^t & \alpha_{21}^t & \alpha_{22}^t & \cdots & \alpha_{2k}^t & \alpha_{m1}^t & \alpha_{m2}^t & \cdots & \alpha_{mk}^t \\ \beta_{11}^t & \beta_{12}^t & \cdots & \beta_{1k}^t & \beta_{21}^t & \beta_{22}^t & \cdots & \beta_{2k}^t & \beta_{m1}^t & \beta_{m2}^t & \cdots & \beta_{mk}^t \end{bmatrix} \tag{7}$$

where, q_j^t is the t th generation of the population, the artificial wolf of the j th individual, k represents the number of quantum bits encoding a gene, and m represents the number of chromosomal genes.

The use of quantum bit encoding allows a quantum state to represent a superposition of multiple states, which allows for better diversity and also better convergence of optimization algorithms introduced into quantum computing.

The quantum revolving door is used to change the artificial wolf position, which has different effects on the performance of the algorithm when different rotation angles are chosen. How to choose the rotation angle of the quantum revolving door is an extremely important issue, and setting the rotation angle too large is prone to produce a local optimal solution, and too small will lead to a longer running time of the algorithm. This study proposes a dynamic adaptive rotation angle to solve this problem, namely:

$$\Delta\theta_i = -\mathrm{sgn}(A_i)\theta_i \tag{8}$$

where, $-\mathrm{sgn}(A_i)$ denotes the positive or negative direction of the rotation angle. $A_i = \begin{bmatrix} \alpha_h & \alpha_i \\ \beta_h & \beta_i \end{bmatrix}$, $\begin{bmatrix} \alpha_h \\ \beta_h \end{bmatrix}$ denotes the probability amplitude corresponding to the current optimal quantum bit, $\begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix}$ denotes the probability amplitude corresponding to the current quantum bit, and θ denotes the magnitude of the rotation angle, which is calculated as

$$\theta_i = \left|\frac{\theta_b - \theta}{M}\right|^{\lambda} \Delta\theta + 0.08\pi e^{-\frac{t}{T}} \tag{9}$$

where, θ_b denotes the current optimal rotation angle, θ denotes the current rotation angle, M denotes a fixed value, generally taken as π, λ denotes a dynamically adjusted nonlinear exponent with a range of $(1.0, 2.0)$, $\Delta\theta$ denotes a dynamic rotation angle with a range of $(0, 0.5\pi)$, t denotes the current number of iterations, and T denotes the maximum number of iterations. Controlling the size of λ and $\Delta\theta$ allows dynamic adaptive control of the size of the rotation angle. The quantum revolving door transformation formula is:

$$U(\Delta\theta_i) = \begin{bmatrix} \cos(\Delta\theta_i) & -\sin(\Delta\theta_i) \\ \sin(\Delta\theta_i) & \cos(\Delta\theta_i) \end{bmatrix} \tag{10}$$

Its transformation process is given by Eq:

$$\begin{bmatrix} \alpha_i' \\ \beta_i' \end{bmatrix} = U(\theta_i)\begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} = \begin{bmatrix} \cos(\Delta\theta_i) & -\sin(\Delta\theta_i) \\ \sin(\Delta\theta_i) & \cos(\Delta\theta_i) \end{bmatrix}\begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} \tag{11}$$

where, $(\alpha_i, \beta_i)^T$ and $(\alpha_i', \beta_i')^T$ are the probability amplitudes of the i th quantum bit before and after it is updated by the revolving door, respectively.

The dynamic adaptive rotation angle of the quantum rotating gate update is used, which makes the optimization algorithm introduced into quantum computing have better practicality and also improves the generalization ability and robustness of the algorithm.

III. B. 3) Adaptive probabilistic random forests

In order to ensure that the improved quantum wolf pack algorithm always runs efficiently when optimizing the two parameters in random forests, it is necessary to find the appropriate λ and $\Delta\theta$ parameters in the dynamic adaptive rotation angle function first. In this study, the Rastrigin function is utilized to carry out the search for the optimal values of the λ and $\Delta\theta$ parameters in the dynamic adaptive rotation angle function, which is computed as:

$$\min f(x) = 20 + x_1^2 + x_2^2 - 10(\cos 2\pi x_1 + \cos 2\pi x_2) \tag{12}$$

Its minimum value is taken to be 0 at point $(0,0)$.

IV. Troubleshooting analysis of the model

IV. A. Model test data

In order to verify the effectiveness of the PE-VMD method in signal denoising, a simple simulation signal of a localized fault occurring in the gear running state is constructed. Based on the vibration data of 20 noisy wind turbine gearboxes collected from this wind farm, four states are classified as healthy state (N), secondary planetary gear ring wear (SPG), sun wheel crush (SWC), and primary planetary gear ring wear (PPG), and the time-domain waveform data in the X and Y directions of each state are obtained.

For the four gearbox states mentioned above, this experiment intercepts 3500 data points as a sample in a randomized manner at the starting point in sequence, generates 1500 samples for the time domain signals of each state, and randomly selects 80% as the training set and 20% as the test set.

IV. B. Analysis of test results

The collected training samples are used as the dataset for this experiment, which are fed into the network for training and testing, and in order to ensure the accuracy of the test results, the ten-fold cross-validation method is used to reduce the chance of random assignment of the training and testing sets. The optimizer is Adam, the Dropout rate is 0.3, the learning rate is 0.002, and the number of iterations is set to 100.

Using the method of this paper in the adaptive probabilistic random forest fault recognition model for 100 iterations, the fault recognition accuracy curve obtained is shown in Figure 2. By observing the accuracy curve, it can be found that the training set reaches convergence when it is iterated in the model up to 36 times, and at this time, the fault recognition accuracy of the training set and the test set is 96.35% and 92.56%, which verifies the validity of the adaptive probabilistic random forest model established in this paper in the application of high-noise wind turbine gearbox fault diagnosis.

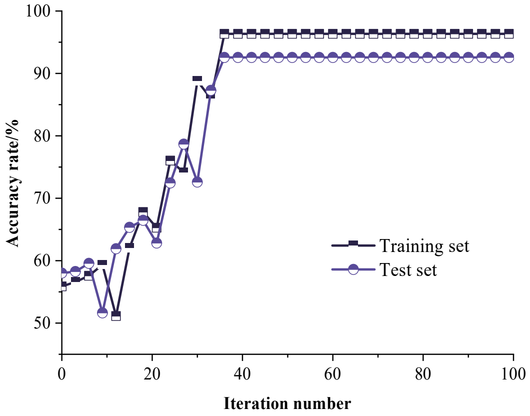
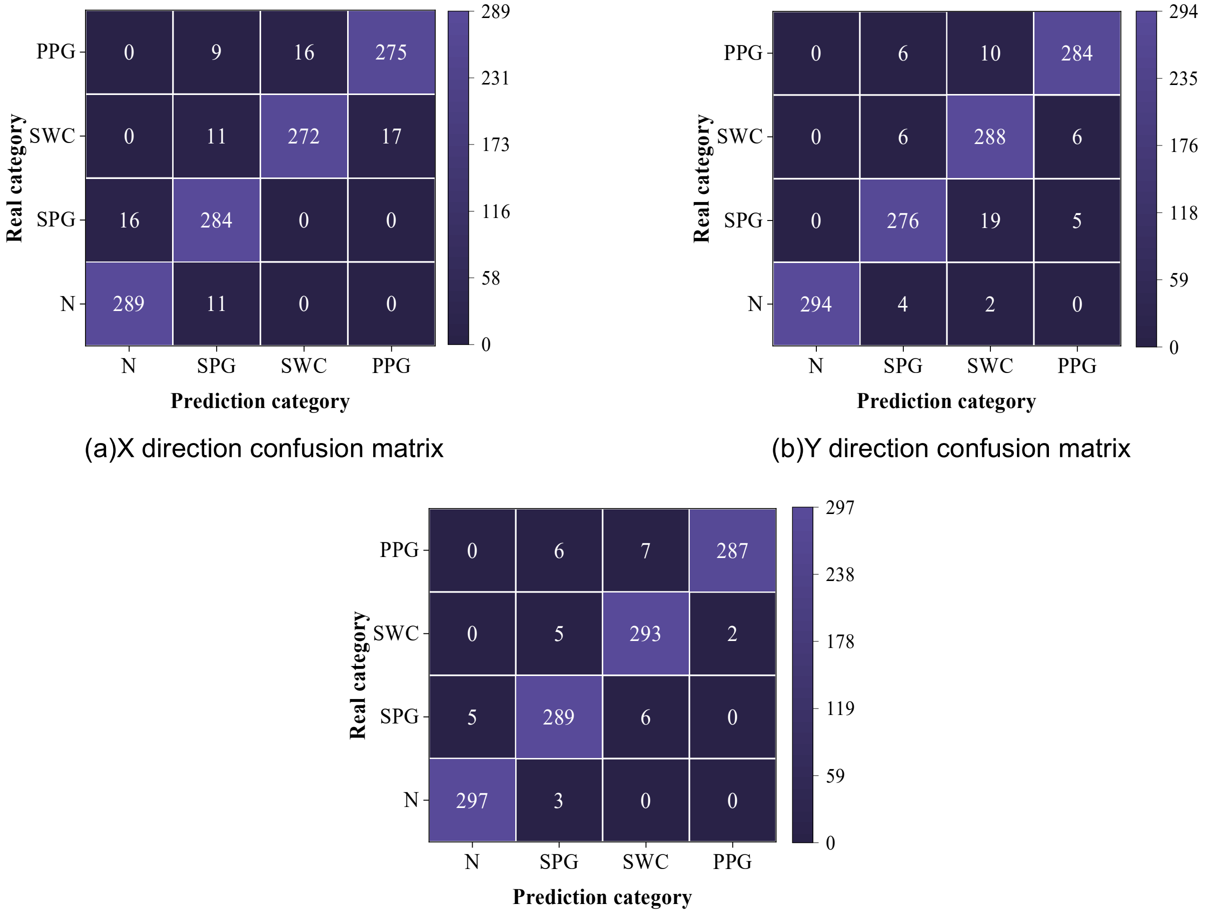


Figure 2: Fault identification accuracy curve

In order to further study the necessity of multi-channel data fusion, the parameters of the adaptive probabilistic random forest network structure are now kept unchanged, and the X-direction and Y-direction data in the training samples are inputted into the adaptive probabilistic random forest network for training respectively, and the confusion matrices for wind turbine gearbox fault classification in the X-direction, Y-direction, as well as the fusion of the X- and Y-direction data are obtained, and the confusion matrices for gearbox fault classification are shown in Fig. 3 shown in Fig. 3.

When only the X-direction is used as input to the model, the fault identification accuracy is 93.33%, when only the Y-direction is used as input to the model, the fault identification accuracy is 95.17%, and after the fusion of the two dual directions, the fault identification accuracy can reach 97.17%. By comparison, it can be seen that referring to the vibration information of both directions can significantly improve the accuracy of fault recognition. Studying the above three confusion matrices, it can be found that the X-direction data is more sensitive to the secondary ring wear fault SPG, and its fault identification accuracy is 2.67% higher than that of the Y-direction fault identification accuracy. On the other hand, the Y-direction vibration data is more sensitive to the sun wheel crush SWC and the primary planetary gear ring wear fault PPG, and its recognition accuracy for these two fault types is 5.33% and 3.00% higher than that of the X-direction, respectively. This analysis of the confusion matrix further validates the effectiveness of the adaptive probabilistic random forest model developed in this paper in the application of high-noise wind turbine gearbox fault diagnosis.



(a)X direction confusion matrix

(b)Y direction confusion matrix

(c)The confusion matrix after fusion

Figure 3: Fault classification confusion matrix of gearbox

IV. C. Comparative analysis of methods

The SVM model and the unimproved random forest model are trained with the same data to compare the advantages and disadvantages between the proposed adaptive probabilistic random forest model. The article uses MATLAB program to record the fault diagnosis effect of the improved adaptive probability random forest model. The test results of the adaptive probabilistic random forest model are shown in Fig. 4, with labels 1 to 4 denoting N, SPG, SWC, and PPG, respectively.The proposed model in this article is able to efficiently identify different faults in the high-noise wind turbine gearbox model. The adaptive probabilistic random forest model proposed in the article has only 4 wrong diagnoses in the fault diagnosis of noisy wind turbine and all others are predicted correctly with a fault identification rate of 92%. The test results of the same SVM model and the unimproved random forest model are shown in Figures 5 and 6. The support vector machine is obviously not able to meet the requirements in the classification of transmission faults, and there are many recognition errors, the fault recognition rate is only 84%, while the random forest model for different faults diagnosis recognition rate is 89%. The adaptive probabilistic random forest model proposed in the article is used in the identification of different faults in the gearbox of a high-noise wind turbine, which can improve the recognition rate of fault diagnosis.

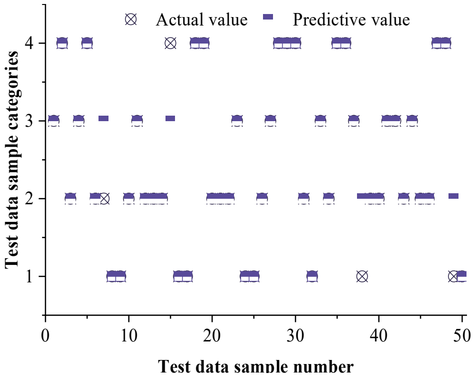


Figure 4: Test results of Adaptive probability random forest

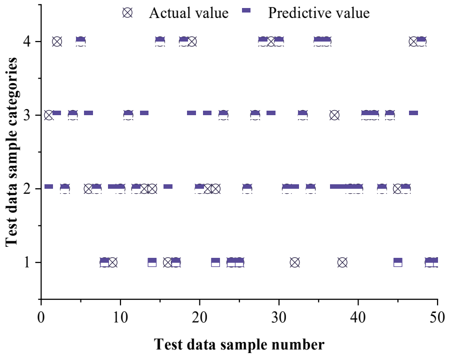


Figure 5: Test results of SVM model

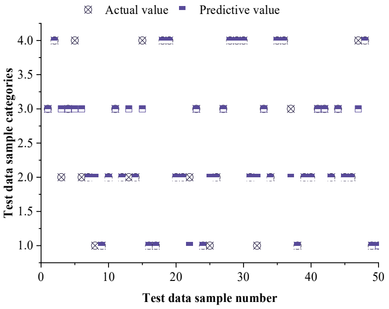


Figure 6: Test results of Random Forest model

V. Conclusion

In this study, by constructing an adaptive probabilistic random forest-based fault diagnosis model for high-noise wind turbine gearboxes, we effectively solve the problems of feature extraction difficulties and low recognition accuracy of traditional methods in strong noise environments. The improved localization-preserving projection algorithm successfully realizes the effective dimensionality reduction of high-dimensional vibration data, which retains the key fault information while eliminating the redundant features. The introduction of quantum wolf pack optimization algorithm significantly improves the adaptive optimization ability of random forest parameters, and the dynamic adaptive rotation angle strategy enhances the convergence and stability of the algorithm.

The experimental validation shows that the method performs well in the identification of four typical fault states of the gearbox. The training process reaches convergence at the 36th iteration, and the fault recognition accuracies of the training and test sets are 96.35% and 92.56%, respectively, which proves the good generalization ability of the model. The effectiveness of the multi-channel data fusion strategy is fully verified, and the fault recognition accuracy after the fusion of X- and Y-direction data reaches 97.17%, which is significantly improved compared with single-direction data. Comparative analysis results show that the comprehensive performance of the proposed method is significantly better than the traditional algorithm, and there are only 4 misdiagnoses in noisy wind turbine gearbox fault diagnosis, and the overall recognition accuracy rate reaches 92%, which provides a reliable technical support for the intelligent operation and maintenance of wind turbines. The method not only improves the accuracy and real-time of fault diagnosis, but also provides new ideas and methods for the development of predictive maintenance technology in the wind power industry.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62173050); the Natural Science Foundation of Hunan Province, China (Grant No. 2023JJ30051); the Energy Conservation and Emission Reduction Hunan University Student Innovation and Entrepreneurship Education Center (Grant No. 2020-19); the Major Scientific and Technological Innovation Platform Project of Hunan Province (2024JC1003); and the Postgraduate Scientific Research Innovation Project of Hunan Province (Grant No. QL20230214).

References

[1] Reder, M. D., Gonzalez, E., & Melero, J. J. (2016, September). Wind turbine failures-tackling current problems in failure data analysis. In Journal of Physics: Conference Series (Vol. 753, No. 7, p. 072027). IOP Publishing.

[2] Chen, J., Pan, J., Li, Z., Zi, Y., & Chen, X. (2016). Generator bearing fault diagnosis for wind turbine via empirical wavelet transform using measured vibration signals. Renewable Energy, 89, 80-92.

[3] Liu, H., Wang, Y., Zeng, T., Wang, H., Chan, S. C., & Ran, L. (2024). Wind turbine generator failure analysis and fault diagnosis: A review. IET Renewable Power Generation, 18(15), 3127-3148.

[4] Mojallal, A., & Lotfifard, S. (2017). DFIG wind generators fault diagnosis considering parameter and measurement uncertainties. IEEE Transactions on Sustainable Energy, 9(2), 792-804.

[5] Gao, Z., & Liu, X. (2021). An overview on fault diagnosis, prognosis and resilient control for wind turbine systems. Processes, 9(2), 300.

[6] Dameshghi, A., & Refan, M. H. (2019). Wind turbine gearbox condition monitoring and fault diagnosis based on multi-sensor information fusion of SCADA and DSER-PSO-WRVM method. International Journal of Modelling and Simulation, 39(1), 48-72.

[7] Dragomiretskiy, K., & Zosso, D. (2013). Variational mode decomposition. IEEE transactions on signal processing, 62(3), 531-544.

[8] Zhang, F., Sun, W., Wang, H., & Xu, T. (2021). Fault diagnosis of a wind turbine gearbox based on improved variational mode algorithm and information entropy. Entropy, 23(7), 794.

[9] Zhang, B. C., Sun, S. Q., Yin, X. J., He, W. D., & Gao, Z. (2023). Research on gearbox fault diagnosis method based on vmd and optimized LSTM. Applied Sciences, 13(21), 11637.

[10] Liu, R., Ding, X., Zhang, Y., Zhang, M., & Shao, Y. (2023). Variable-scale evolutionary adaptive mode denoising in the application of gearbox early fault diagnosis. Mechanical Systems and Signal Processing, 185, 109773.

[11] Wang, Z., He, G., Du, W., Zhou, J., Han, X., Wang, J., ... & Kou, Y. (2019). Application of parameter optimized variational mode decomposition method in fault diagnosis of gearbox. Ieee Access, 7, 44871-44882.

[12] Liu, T., Wang, Y., Cui, L., & Zhang, C. (2022, August). Study on fault diagnosis method of key components of the gearbox under variable working conditions based on improved VMD algorithm. In International conference on the Efficiency and Performance Engineering Network (pp. 74-89). Cham: Springer Nature Switzerland.

[13] Sharma, V. (2021). A review on vibration-based fault diagnosis techniques for wind turbine gearboxes operating under nonstationary conditions. Journal of The Institution of Engineers (India): Series C, 102(2), 507-523.

[14] Li, H., Fan, B., Jia, R., Zhai, F., Bai, L., & Luo, X. (2020). Research on multi-domain fault diagnosis of gearbox of wind turbine based on adaptive variational mode decomposition and extreme learning machine algorithms. Energies, 13(6), 1375.

[15] Muralidharan, A., Sugumaran, V., Soman, K. P., & Amarnath, M. (2014). Fault diagnosis of helical gear box using variational mode decomposition and random forest algorithm. Structural Durability & Health Monitoring, 10(1), 55.

[16] Cabrera, D., Sancho, F., Sánchez, R. V., Zurita, G., Cerrada, M., Li, C., & Vásquez, R. E. (2015). Fault diagnosis of spur gearbox based on random forest and wavelet packet decomposition. Frontiers of Mechanical Engineering, 10, 277-286.

[17] Chen, S., Yang, R., & Zhong, M. (2021). Graph-based semi-supervised random forest for rotating machinery gearbox fault diagnosis. Control Engineering Practice, 117, 104952.

[18] Qin, X., Li, Q., Dong, X., & Lv, S. (2017). The fault diagnosis of rolling bearing based on ensemble empirical mode decomposition and random forest. Shock and Vibration, 2017(1), 2623081.

[19] Yuan, T., Sun, Z., & Ma, S. (2019). Gearbox fault prediction of wind turbines based on a stacking model and change-point detection. Energies, 12(22), 4224.

[20] Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: data mining and knowledge discovery, 9(3), e1301.

[21] Han, T., & Jiang, D. (2016). Rolling Bearing Fault Diagnostic Method Based on VMD-AR Model and Random Forest Classifier. Shock and Vibration, 2016(1), 5132046.