energies

MDPI

Article

Fault Diagnosis Method for Wind Turbine Gearboxes Based on IWOA-RF

Mingzhu Tang [1,†], Zixin Liang [1], Huawei Wu [2,*] and Zimin Wang [3,†]

[1] School of Energy and Power Engineering, Changsha University of Science & Technology, Changsha 410114, China; tmz@csust.edu.cn (M.T.); 20106011013@stu.csust.edu.cn (Z.L.)

[2] Hubei Key Laboratory of Power System Design and Test for Electrical Vehicle, Hubei University of Arts and Science, Xiangyang 441053, China

[3] School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541004, China; worthyman@guet.edu.cn

[*] Correspondence: whw_xy@hbuas.edu.cn

[†] Mingzhu Tang and Zimin Wang contributed equally to this work and share first authorship.

Abstract: A fault diagnosis method for wind turbine gearboxes based on undersampling, XGBoost feature selection, and improved whale optimization-random forest (IWOA-RF) was proposed for the problem of high false negative and false positive rates in wind turbine gearboxes. Normal samples of raw data were subjected to undersampling first, and various features and data labels in the raw data were provided with importance analysis by XGBoost feature selection to select features with higher label correlation. Two parameters of random forest algorithm were optimized via the whale optimization algorithm to create a fitness function with the false negative rate (*FNR*) and false positive rate (*FPR*) as evaluation indexes. Then, the minimum fitness function value within the given scope of parameters was found. The WOA was controlled by the hyper-parameter α to optimize the step size. This article uses the variant form of the sigmoid function to alter the change trend of the WOA hyper-parameter α from a linear decline to a rapid decline first and then a slow decline to allow the WOA to be optimized. In the initial stage, a larger step size and step size change rate can make the model progress to the optimization target faster, while in the later stage of optimization, a smaller step size and step size change rate allows the model to more accurately find the minimum value of the fitness function. Finally, two hyper-parameters, corresponding to the minimum fitness function value, were substituted into a random forest algorithm for model training. The results showed that the method proposed in this paper can significantly reduce the false negative and false positive rates compared with other optimization classification methods.

Keywords: wind turbine; gearbox fault; XGBoost feature selection; whale optimization; random forest

check for updates

Citation: Tang, M.; Liang, Z.; Wu, H.; Wang, Z. Fault Diagnosis Method for Wind Turbine Gearboxes Based on IWOA-RF. *Energies* **2021**, *14*, 6283. https://doi.org/10.3390/en14196283

Academic Editors: Xiandong Ma, Sinisa Durovic and Mohamed Benbouzid

Received: 4 August 2021
Accepted: 23 August 2021
Published: 2 October 2021

1. Introduction

Wind power generation [1] is of great importance in the clean energy power generation sector. In recent years, driven by policies of carbon neutrality and carbon emission peaks [2], China has focused on the development of clean energy, in which wind power is one of the most mature technologies and is expanding in scale [3]. However, while the use of wind turbines is expanding in scale, the number of wind turbine accidents is also increasing [4]. Wind turbine faults emerge endlessly due to changeful working conditions and exposure to the sun, rain, sandstorms, and other severe weather factors throughout the year [5]. As a result, faults such as turbine gearbox faults [6], main bearing faults [7], and generator faults [8] lead to wind turbine maintenance downtime. Wind turbine maintenance is difficult and expensive due to high-altitude maintenance operations, which results in the need for considerable manpower and material resources, and can incur huge economic losses. For this reason, monitoring and maintenance of wind turbines is of the utmost importance to avoid secondary damage and reduce the maintenance cost and difficulties [9].

The wind turbine gearbox [10] is an important transmission part, but its fault frequency is at the top of the list. Data [11] shows that gearbox faults lead to the longest maintenance downtime. Common wind turbine gearbox faults occur in gears and bearings [12]. A changeable running environment and complex operating conditions [13] lead to the necessary long-term adjustment of the load operation, thus causing excessive mechanical equipment wear, including wear and pitting corrosion of the gearbox gears, tooth surface deformation and tooth cracks, and the enhanced possibility of a crack on the inner ring, outer ring, and cage of the bearing [14].

Machine learning algorithms are widely used in wind turbine fault diagnosis. Xiang [15] used the method of convolutional neural network cascading to LSTM (long short-term memory) network to warn in the event of an abnormal state in wind turbines. Kordestani [16] combined the dynamic principal component analysis (DPCA) with the support vector machine (SVM) to identify dynamic fault states. Trizoglou [17] proposed an extreme gradient-reinforced model for fault diagnosis of wind turbine gearboxes with better prediction accuracy than LSTM and a lower calculation cost. Liu [18] proposed a hybrid model to predict the oil temperature of wind turbine gearboxes. Pan [19] utilized the deep belief network (DBN) and self-organizing map (SOM) to reduce the noise of the sensor's characteristic signal and then optimized and improved the particle filter (PF) algorithm via the fruit fly optimization algorithm (FOA) to predict the service life of the gearbox. The above methods improved the accuracy of the model, while simultaneously making the model more complex, increasing training time, and worsening real-time performance, thus making the practical application of some of these projects unfeasible.

The random forest (RF) [20] algorithm is an extended variant of bagging, and, belonging to the integration algorithm, it is simple, easy to implement, requires a lower computational expense, and has strong classification performance. The random forest algorithm is widely used in machine learning for fault diagnosis. In the field of medicine, Asadi [21] used the improved RF to diagnose patients with heart diseases and P. K. P [22] used the RF, subject to bayesian optimization, to realize an accurate classification of breast cancer. In traffic, a mixed method, combining random forest regression and maximum information coefficient, was used to predict whether a flight would be delayed [23]. In agriculture, Makungwe [24] combined the linear mixture model with RF to predict soil pH range. The RF algorithm, with strong computing power, was used in this paper for the fault diagnosis of wind turbine gearboxes.

Feature selection [25] is necessary to find the optimal feature subset, as redundant features will increase model training time and irrelevant features will reduce the accuracy of the model in the dataset. Common feature selection methods include Pearson's correlation coefficient, chi-square test, XGBoost [26], variance selection, mutual information and maximum information coefficient, etc. In this study, XGBoost feature selection was used to process the data and identify the correlation between each feature and gearbox fault more conveniently and intuitively.

An appropriate classification algorithm selected for fault diagnosis may not achieve the expected results. For this reason, the hyper-parameters in the classification algorithm should be adjusted, namely, subject to optimization. Long [27] adopted the grey wolf optimization algorithm and proposed an exploration-enhanced grey wolf optimization algorithm (EEGWO) [28]. Long [29] also proposed an enhanced adoptive butterfly optimization algorithm for PV model parameter identification. Tang [30] proposed a cost-sensitive large margin distributor (CLDM) for fault diagnosis of wind turbine generators to reduce the influence of data category imbalance, and the cost-sensitive extreme random forest algorithm (CS-ERF) was proposed [31] to further study the fault of wind turbine generators.

In recent years, a large number of researchers have begun to use machine learning methods to realize the intelligent diagnosis of fan gearbox faults. McKinnon [32], using SCADA data, compared three algorithms of one-class support vector machine (OCSVM), isolation forest (IF), and elliptical envelope (EE) for the same fault. The average accuracy of OCSVM was 82%, which was better than IF and EE, but the accuracy of fault classification required improvement. Yang [33] proposed a method based on deep joint variational automatic encoder to detect wind turbine gearbox faults. This method reduced the *FNR*, but the optimization of the *FPR* was not obvious. Corley [34] combined thermal modeling and machine learning methods to detect gearbox faults. The method had less conclusive results, which may have affected the accuracy of fault detection. Tang [35] improved the lightGBM method, which solved the problem of low computational efficiency and poor real-time performance of the traditional GBDT algorithm, but there was also a risk that it would fall into local optimization.

However, there are still some problems to be solved in practical engineering applications. Overfitting and local optimum can easily occur due to the strong capability of the random forest algorithm. A hyper-parameter fault diagnosis model, based on a random forest algorithm optimized by WOA, was proposed to solve the problem of finding the global optimal parameters of the fault diagnosis model of wind turbine gearboxes, to allow for improved diagnosis performance of the fault diagnosis model.

The whale optimization algorithm (WOA) [36] is a bionic algorithm imitating whale hunting, which mainly includes three search mechanisms: (1) realizing the local search of the algorithm via shrinkage encirclement mechanism and (2) spiral mechanism, and (3) the global search of the algorithm via random learning strategy. With advantages of a simple process, a fast rate of convergence, and excellent performance in solving optimization problems, the WOA is widely used in various applications.

This article uses the advantages of WOA and flexibly controls the optimization speed by changing the hyper-parameters of WOA, and change the location update strategy to make it more random and reduce the risk of WOA falling into a local optimum. The hyper-parameters to be optimized in random forest algorithm were substituted into the improved whale optimization algorithm (IWOA). The optimal value was quickly found through three search methods, and then the optimal parameter value was returned to RF for training. Finally, the fault diagnosis was carried out by the trained model.

2. Materials and Methods

2.1. Random Forest Algorithm

After the bagging integration was created by RF with a decision tree as the base learner, random feature selection was further introduced into the RF on the basis of the training degree of the decision-making tree. For attribute selection and partition, the traditional decision tree selected an optimal attribute among the attributes of the current node, while the RF randomly selected a subset containing k attributes from the attribute set of the node on the base decision tree. It then selected an optimal attribute from the subset for the partition. The substitute of all training subsets was put into different base learners for training. Each learner′s classification results were voted for in a comprehensive way, and the result with the largest number of votes was the final prediction of the RF. Figure 1 showed the flow chart of the random forest algorithm.

The diversity of base learner in the random forest was achieved by sample disturbance and attribute disturbance, leading to further improvement of the finally-integrated generalization performance due to the increase in the difference degree among individual learners.

The hyper-parameters of the random forest included the number of the decision in the forest, the depth of the decision tree, the number of the optimal splitting point features, the criterion to measure the splitting, and the minimum number of samples on the leaf node, which directly affect the accuracy of the random forest classification and the required time [37]. The number of decision trees and the minimum number of samples on the leaf node were selected, in this study, via the empirical law for optimization.

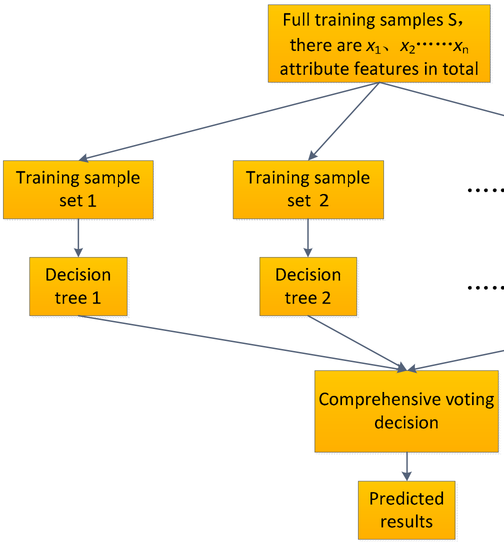


Figure 1. Flow chart of random forest algorithm.

2.2. Whale Optimization Algorithm (WOA)

Derived through the imitation of whale hunting, WOA [38] is a meta-heuristic algorithm based on swarm intelligence, which mainly includes three search mechanisms: realizing the local search of the algorithm via shrinkage encirclement mechanism and spiral mechanism, and the global search of the algorithm via random learning strategy. It has the advantage of being a simple process with a fast rate of convergence.

2.2.1. Search Prey Model

The mathematical model when the whale searches for prey is as follows:

$$D = |CX_{rand} - X(t)|, \tag{1}$$

$$X(t+1) = X_{rand} - AD, \tag{2}$$

t is the current number of training, $t \geq 1$; and X_{rand} is the randomly selected whale position vector. When $A \geq 1$ in the algorithm, a search location was randomly selected to update the position of the other whales to force the whale to deviate from the prey, thereby finding a more appropriate prey. In this case, the exploration ability of the algorithm was strengthened so that the WOA could perform a global search.

2.2.2. Mathematical Model

WOA's mathematic model to encircle preys is as follows:

$$D = |CX^*(t) - X(t)|, \tag{3}$$

$$X(t+1) = X^*(t) - AD, \tag{4}$$

In the above equations, t is the number of current iterations; A and C are coefficient vectors, $X^*(t)$ represents the best position for whale hunting, namely the current optimal solution, $X(t)$ represents the current position vector of the whale, and D represents the distance between the front and rear positions of the whale. Coefficient vectors A and C can be represented in the following form:

$$A = 2\alpha r_1 - \alpha, \tag{5}$$

$$C = 2r_2, \tag{6}$$

$$\alpha = 4 - \frac{4}{1+e^{-t}}, \tag{7}$$

where r_1 and r_2 are two random numbers in (0, 1) and α is the hyper-parameter that controls the whale's moving step length. The value of α decreases curvilinearly from 2 to 0.

2.2.3. Hunting Mathematical Model

In order to reduce the possibility of the model falling into the local optimum, this article will use the randomness of the parameters in the hunting model to obtain three random moving positions of the whale, and use the average of these three moving positions as the next moving position. According to the whale's spiral movement for hunting, the mathematic model of the whale's movement location is as follows:

$$X_1 = X^*(t) + D_{p_1} e^{bl_1} cos(2\pi l_1)\ , \tag{8}$$

$$X_2 = X^*(t) + D_{p_2} e^{bl_2} cos(2\pi l_2)\ , \tag{9}$$

$$X_3 = X^*(t) + D_{p_3} e^{bl_3} cos(2\pi l_3)\ , \tag{10}$$

$$X(t+1) = \frac{X_1 + X_2 + X_3}{3} \tag{11}$$

where $D_p = |X^*(t) - X(t)|$ represents the distance between the whale and the prey; b is a constant to define the shape of the helix; and l is a random number in $(-1, 1)$. The whale swims to the prey in a helix with increased curvature. Suppose that it is of probability σ to select the shrinkage encirclement mechanism and 1 − σ to select the spiral mechanism for updating the whale's location. The mathematical model is as follows:

$$X(t+1) = \begin{cases} X^*(t) - AD, p < \sigma \\ X(t+1) = X^*(t) + D_p e^{bl} cos(2\pi l), p \geq \sigma \end{cases}, \tag{12}$$

When the whale was going to attack the prey, α was set to decrease when the whale got closer to the prey in the mathematical model. The fluctuation range also decreased over α's decrease. When α decreased from 2 to 0 during iteration, A was a random value within $[-\alpha, \alpha]$. When the value of A was with [−1, 1], the whale may appear at any position between its current position and the position of the prey. When A < 1 was set in the algorithm, the whale attacks the prey.

2.3. Optimization Algorithm Flow

Table 1 presents the functions and value ranges of the number τ of decision trees in the IWOA-RF, the main optimization model used in this experiment, and the minimum number δ of samples on the leaf node.

Table 1. Tuning parameters adjusted for optimization.

Parameter	Function	Value Range
τ	Number of decision trees in random forest	[10, 200]
δ	Minimum number of samples on the leaf node of random forest	[1, 300]

2.3.1. WOA Pseudo-Code

Algorithm 1 is the original algorithm of WOA. The paper combined improved WOA with RF to form IWOA-RF. The specific details of IWOA-RF improvement can be reflected in its pseudo code.

The optimization process of WOA can be seen as follows:

Algorithm 1: The steps of WOA optimization parameters.

Input: Number of iterations: *t*, The maximum number of iterations: *max_iter*, Population size: *SN_num*, dimension: *dim*, parameter vector with dimension *dim*: *X**;

```
1: Initialize the whale population $SN\_num(i = 1,2, \ldots n)$, $t = 1$;
2: Initialize the position of the whale population;
3: Calculate the fitness value corresponding to each whale and rank fitness values to select
   $SN\_num$ whales as the initial population;
4: Calculate the fitness value of the $SN\_num$ individual and find the position of the individual
   with the smallest fitness value as the optimal position;
5: Update the location of the next generation;
6: While $t = max\_iter$, output the optimal individual, namely the optimal solution found by the
   algorithm. Otherwise $t < max\_iter$,$t = t + 1$, return to step (4);
```

Output: Get the best parameter vector *X** as the optimal position of dimension *dim*.

2.3.2. IWOA-RF Pseudo-Code

Implementation of the proposed RF hyper-parameters optimization can be detailed as follows (Algorithm 2):

Algorithm 2: Implementation of IWOA-RF fault detection method.

Input: WOA parameters($lb(\tau\_min, \delta\_min)$, $ub(\tau\_max, \delta\_max)$, dimension, and maximum number of iterations), *SN_num*, φ, σ, μ, *t*, and RF parameters (τ, δ);

```
x_train,y_train,x_test,y_test→RF($\tau$, $\delta$);
Construct fitness function : $fitness = FPR + \varphi * FNR$;
Initialize the whale population $SN\_num(i = 1,2, \ldots n)$;
Calculate the fitness of each search agent;
$X^*(\tau, \delta)$ = the best search agent;
while ($t$ < maximum number of iterations);
    for each search agent;
        if_1($p < \sigma$);
          if_2($|A| < \mu$);
                update the parameter vector by the Equation (11);
          else if_2($|A| \geq \mu$);
                select a random search agent($X\_rand$);
                update the parameter vector by the Equation (4);
          end if_2;
        else if_1($p \geq \sigma$);
        update the parameter vector by the Equation (12);
        end if_1;
    end for;
    check if any search agent goes beyond the search space and amend it;
    calculate the fitness of each search agent;
    find out the three positions corresponding to the minimum three fitness values and
calculate the average position;
    update $X^*$ if there is a better solution;
    $t = t + 1$;
end while;
return $X^*(\tau\_optimal,\delta\_optimal)$;
$X^*(\tau\_optimal, \delta\_optimal) \rightarrow$ RF;
x_test,y_test→RF($\tau\_optimal$, $\delta\_optimal$);
```

Output: *FNR* and *FPR*.

3. Results

The wind turbine gearbox was taken as the experimental object. Due to the large number of sensors in the turbine, which receive data, the complex relationship among data features, and the different size of the relationship between features and faults, the parameters of the data and model should be adjusted to clean out redundant and irrelevant features. The hyper-parameters in the diagnostic model should be considered to optimize the whole model and prevent the model from local optimum and overfitting. The flow chart of fault diagnosis for wind turbine gearboxes by Algorithm 2 is shown in Figure 2.

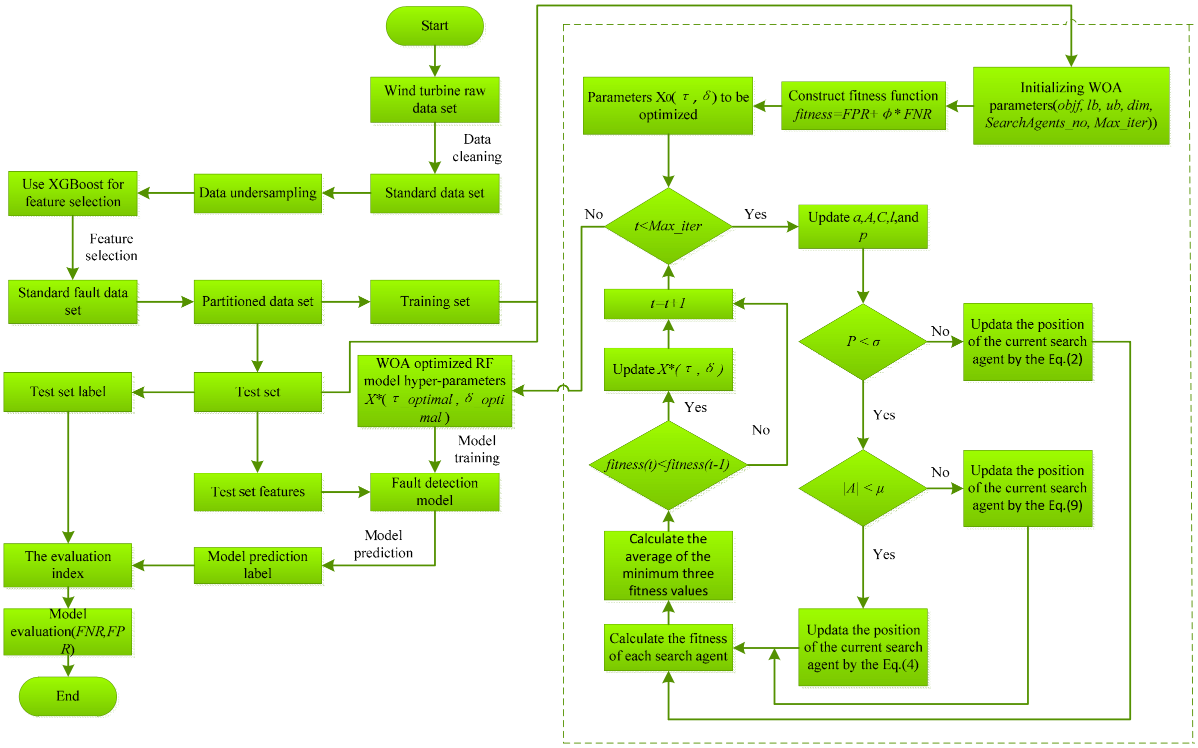

Start
Wind turbine raw data set
Data cleaning
Standard data set
Data undersampling
Use XGBoost for feature selection
Feature selection
Standard fault data set
Partitioned data set
Training set
Test set
Test set label
WOA optimized RF model hyper-parameters $X^*(\tau\_optimal, \delta\_optimal)$
Model training
Test set features
Fault detection model
Model prediction
Model prediction label
The evaluation index
Model evaluation(*FNR,FPR*)
End
Initializing WOA parameters(*objf, lb, ub, dim, SearchAgents_no, Max_iter*))
Construct fitness function *fitness=FPR+ ϕ* FNR*
Parameters X0(τ , δ) to be optimized
No
t<*Max_iter*
Yes
Update *a,A,C,l,*and *p*
$P < \sigma$
No
Updata the position of the current search agent by the Eq.(2)
Yes
$|A| < \mu$
No
Updata the position of the current search agent by the Eq.(9)
Yes
Updata the position of the current search agent by the Eq.(4)
Calculate the fitness of each search agent
Calculate the average of the minimum three fitness values
fitness(t)<fitness(t-1)
No
Yes
Update $X^*(\tau, \delta)$
t=t+1

Figure 2. Gearbox fault diagnosis flowchart.

Figure 2 is the overall flow chart of IWOA-RF. The original data was cleaned and the data was under-sampled to obtain a standard data set. The standard data set was divided into training set and test set with a split function, and the training set and test set were substituted into IWOA The optimal parameter vector of the current RF was obtained, and the optimal parameter vector is obtained after maximum number of cycles. Finally, the test set is substituted into the RF to obtain the FNR and FPR.

3.1. Data Description

The operation data of a 1.5 MW wind turbine in a wind farm in Inner Mongolia was used in this experiment. This wind farm contains 33 wind turbines. The data of a wind turbine with a gearbox fault, chosen through random selection, was used. The data in-terval was 1 min. For the rigor of the experiment, this experiment selected the data 5 h before the failure, the data when the failure occurs, and the data 5 h after the failure. The gearbox is an important variable-speed transmission in wind turbines, mainly composed of gears, bearings, and transmission shafts. Most faults occur on the gearbox's gears and bearings. The gearbox and other components were monitored by measuring the bearing temperature, gear speed, generated power, and other parameters of the turbine by sensors. Figure 3 presents the overall system diagram of the wind turbine.

The original data of the turbine shown in Table 2 includes a portion of the data of the No. 8 wind turbine collected in January.

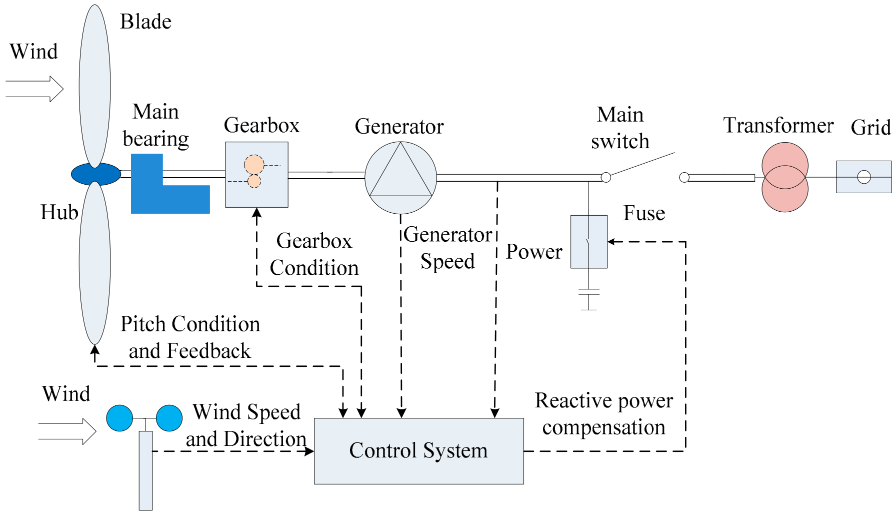


Figure 3. Wind turbine system diagram.

Table 2. Original data of the wind turbine.

Feature Parameters	**Time**								
	01:45	**01:46**	**01:47**	**01:48**	**01:49**	**01:50**	...	**22:58**	**22:59**
rotor_speed	17.4	17.39	17.41	17.37	17.41	17.47	...	17.45	17.41
converter_motor_speed	1749.2	1747.4	1746.1	1745.4	1749.4	1747.7	...	1748.1	1749
...	...	...	...	...	...	...	...	...	...
converter_power	875	845.9	758.3	797.1	745	789.5	...	789.5	734.5

3.2. Data Cleaning and Preprocessing

With a large amount of data, incomplete, null, and other values in the original data, due to sensors and for other reasons, may be directly discarded to obtain a non-null dataset. The wind turbine runs normally the majority of the time, leading to an excess of normal samples and unbalanced categories. In this paper, the number of normal samples in the original data was reduced and the number of normal samples and fault samples in the data was balanced by undersampling. The dataset also contained pitch fault, yaw fault, and other faults at a certain point in time. This experiment aimed to study the gearbox fault, so other disturbing fault data was filtered out to obtain a standard single fault and normal datasets. The data, after the cleaning, was preprocessed for standardization.

Each feature has different sensitivity to different faults of wind turbines. XGBoost was used to screen all features once in this experiment to select the more important features more accurately and reduce the complexity of the model.

3.3. Feature Selection

More features does not lead to a higher accuracy of the training in machine learning [39]. On the contrary, a large number of redundant features in the data will not only affect the classification speed of the model, but also affect the classification accuracy of the model and increase the risk of missing and false alarms. XGBoost was used for feature selection and calculating the features to be split through parallel features. Multiple threads were used in an attempt to take each feature as the feature to be split and find the optimal splitting point of each feature. The feature with the largest gain was selected as the feature to be split, after calculation of the gain was generated after splitting.

Table 3 shows the feature importance of each feature obtained after XGBoost feature selection.

Table 3. Feature importance.

Feature	Importance	Feature	Importance
converter_power	0.141234	generator_winding_temperature_u1	0.055195
converter_generator_torque	0.092532	pitch_ssb_motor_temperature_1	0.053571
...	...	...	...
gearbox_oil_temperature_oil_inlet	0.066558	pitch_position_2	0.048701
pitch_ssb_motor_temperature_3	0.056818	main_bearing_rotor_side_temperature	0.045454
pitch_position_3	0.056818	generator_winding_temperature_v1	0.042207

Figure 4 presents the partial feature importance after visualization to more clearly distinguish the importance of each feature to the tag.

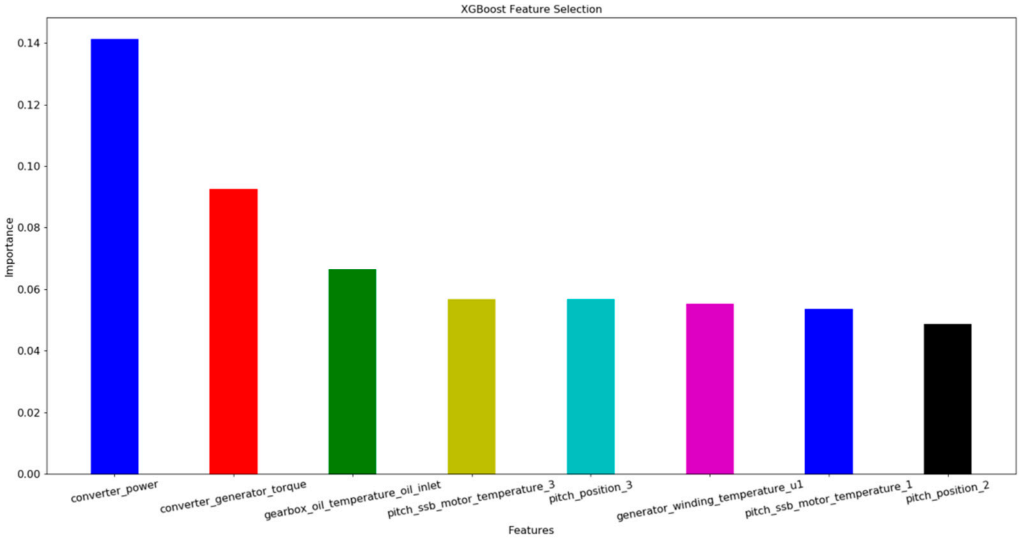


Figure 4. XGBoost partial feature importance.

From Figure 4, it is evident that the generator power, generator shaft torque, and generator inlet oil temperature are the main factors affecting gearbox faults. Feature screening was performed by calculating the importance of each feature. Features with low importance affected the accuracy of the model, but also consumed a considerable amount of the computing power, causing poor real-time performance of the model. For this reason, the features that fell below the mean value of feature importance were discarded and those above the mean value were retained. The mean feature importance was 0.026316, and, thus, 12 feature data were retained.

3.4. Performance Evaluation Index of Fault Diagnosis

The unimproved whale optimization algorithm (WOA), particle swarm optimization (PSO) [40], and salp swarm algorithm (SSA) were compared with the IWOA-RF to verify the effectiveness and superiority of the IWOA-RF in the diagnosis of wind turbine gearbox faults. Three random forest models, after optimization and improved whale optimization algorithm (IWOA), were substituted into three new data to test the reliability of the model.

The false negative rate (*FNR*) and false positive rate (*FPR*) were taken as performance evaluation indexes in this paper. *FNR* and *FPR* can be represented in the following form:

$$FNR = \frac{FN}{TP + FN}, \tag{13}$$

$$FPR = \frac{FP}{TN + FP}, \tag{14}$$

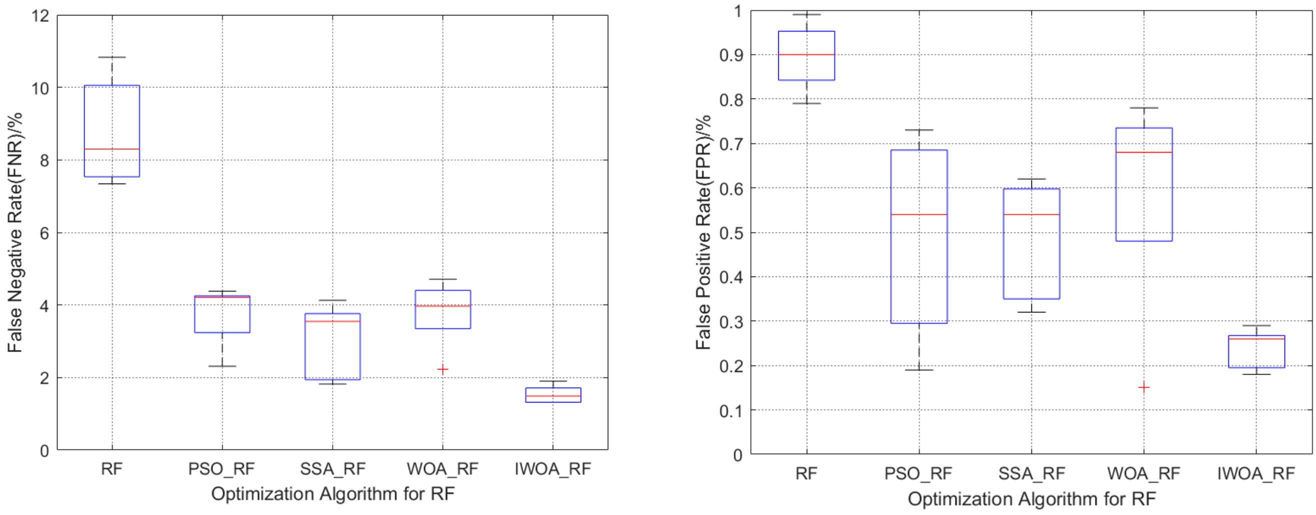
where *TP*, *TN*, *FP*, and *FN* were confusion matrix as shown in Table 4.

Table 4. Confusion matrix of classification results.

The Actual Situation	Forecast Classification	
	Predicted Failure	**Predicted Normal**
The actual fault	*TP*	*FN*
The actual normal	*FP*	*TN*

3.5. Experimental Results

The *FNR* and *FPR* of dataset 1 were obtained after multiple iterations of the model, as shown in Figure 5.



(**a**) Boxplot of *FNR* for dataset 1

(**b**) Boxplot of *FPR* for dataset 1

Figure 5. Boxplot of *FNR* and *FPR* for dataset 1.

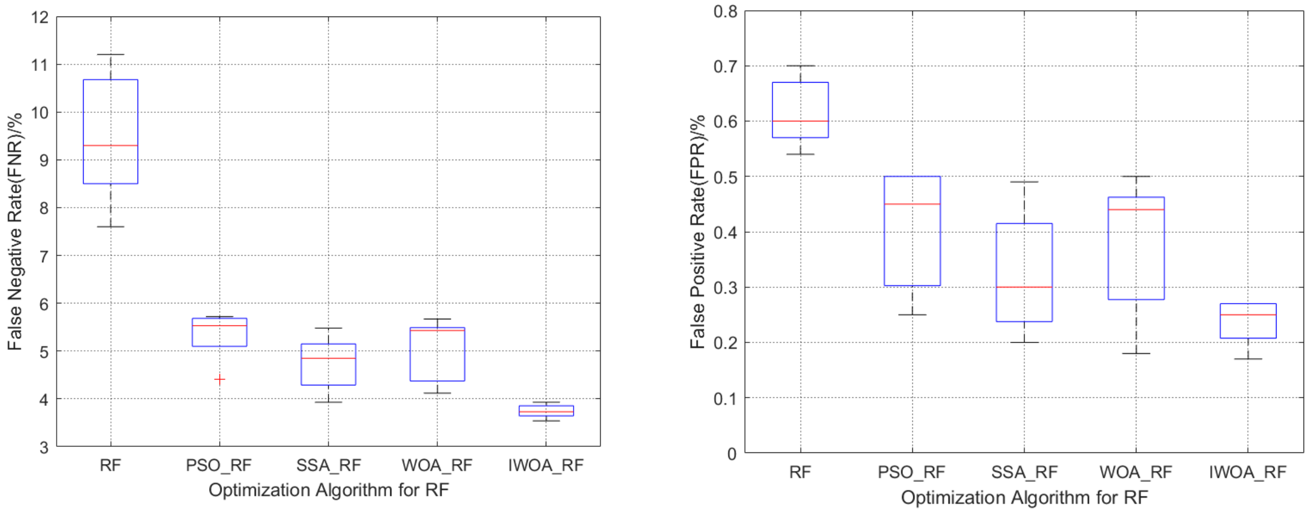
Under the same amount of training, it can be seen from Figure 5a that the average *FNR* of the unimproved random forest algorithm model exceeds 8%, while the average *FNR* of PSO-RF, SSA-RF and WOA-RF is about 4%, and the optimization effect is obvious; but The *FNR* of IWOA-RF is controlled below 2%, which is better than the other three optimization methods. It can be seen from Figure 5b that the average *FPR* of the five models are less than 1%, and the average *FPR* of PSO-RF, SSA-RF and WOA-RF are all higher than 0.5%, and the classification results fluctuate greatly. The performance of the optimization of the *FPR* is unstable, which may be that the algorithm has fallen into a local optimum; while the *FPR* of IWOA-RF is controlled between 0.1% and 0.3%, the *FPR* is low, the fluctuation is small, and the classification effect is better than others Four algorithms.

The *FNR* and *FPR* of dataset 2 are shown in Figure 6.

Comparing Figures 5 and 6, it is clear the general tendency of *FNR* and *FPR* of dataset 2 was similar to that of dataset 1. The *FNR* of IWOA-RF is lower than that of the other four algorithms. The *FNR* is controlled below 4%. In Figure 6b, although the average *FPR* of SSA-RF is slightly higher than that of IWOA-RF, the fluctuation of *FPR* of SSA-RF range is too large. Overall, the classification performance of IWOA-RF is better than the other three optimizations.

The *FNR* and *FPR* of dataset 3 are shown in Figure 7.

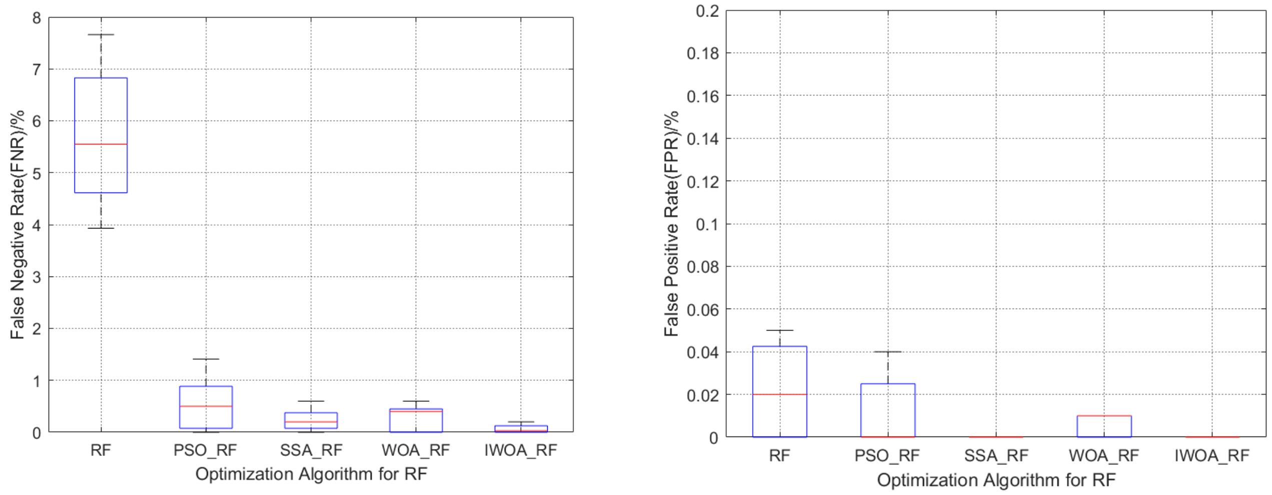
The *FPR* of all five models was almost zero, which may result from the fact that the data quality of dataset 3 was relatively better, with less fault data, leading to a greater impact on the model. It can be seen from Figure 7a,b that the *FPR* and *FNR* of IWOA-RF are compared with the other three types. There is little difference between the *FNR* and the *FPR* of the optimized model, especially the *FPR*. The *FPR* of the five models all appear zero, and the overall average *FPR* is close to zero; The average *FNR* of the four optimization algorithms is less than 1%, and the *FNR* of IWOA-RF is even closer to zero. It can be proved that the diagnostic effect of IWOA-RF is better than the other three optimization algorithms.



(**a**) Boxplot of *FNR* for dataset 2

(**b**) Boxplot of *FPR* for dataset 2

Figure 6. Boxplot of *FNR* and *FPR* for dataset 2.



(**a**) Boxplot of *FNR* for dataset 3

(**b**) Boxplot of *FPR* for dataset 3

Figure 7. Boxplot of *FNR* and *FPR* for dataset 3.

4. Conclusions

A fault diagnosis method for wind turbine gearboxes based on undersampling, XGBoost feature selection, and IWOA-RF was proposed in this paper to solve the problem of high *FNR* and *FPR* of wind turbine gearboxes. The main contributions of this paper are as follows: (1) undersampling and XGBoost feature selection were adopted to reduce the dimension of the original data and eliminate the negative impact of data category imbalance and redundant features on the model; and (2) an optimized Whale optimization algorithm was used to optimize the number of classifiers and the minimum number of samples on leaf nodes in the random forest. The IWOA-RF was compared with the WOA_RF, PSO-RF, and SSA-RF using three datasets of *FNR* and *FPR*. The results showed that the proposed method can effectively reduce the *FNR* and *FPR* during fault diagnosis under dataset imbalance and many redundancy features. The method proposed in this paper was more stable than other optimization methods after the comparison of *FNR* and *FPR*.

Author Contributions: Conceptualization and methodology, M.T.; software, validation, and original draft, Z.L.; data curation and visualization, H.W.; writing—review and editing, Z.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China (grant nos. 62173050 and 61403046), the Natural Science Foundation of Hunan Province, China (grant no. 2019JJ40304), Changsha University of Science and Technology's "The Double First Class University Plan" International Cooperation and Development Project in Scientific Research in 2018 (Grant No. 2018IC14), the Research Foundation of the Education Bureau of Hunan Province (grant no.19K007), the Hunan Provincial Department of Transportation 2018 Science and Technology Progress and Innovation Plan Project (grant no. 201843), the Energy Conservation and Emission Reduction Hunan University Student Innovation and Entrepreneurship Education Center, the Innovative Team of Key Technologies of Energy Conservation, the Emission Reduction and Intelligent Control for Power-Generating Equipment and System, CSUST, the Hubei Superior and Distinctive Discipline Group of Mechatronics and Automobiles (XKQ2021003 and XKQ2021010), the Major Fund Project of Technical Innovation in Hubei (grant no. 2017AAA133), the Guangxi Key Laboratory of Trusted Software (no. 201728), and the Graduate Scientific Research Innovation Project of Changsha University of Science &Technology (No. 2021-89).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author, upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Xi, L.; Wu, J.; Xu, Y.; Sun, H. Automatic Generation Control Based on Multiple Neural Networks With Actor-Critic Strategy. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 2483–2493. [CrossRef]
2. Li, Z.; Li, R.Y.M.; Malik, M.Y.; Murshed, M.; Khan, Z.; Umar, M. Determinants of Carbon Emission in China: How Good is Green Investment? *Sustain. Prod. Consum.* **2021**, *27*, 392–401. [CrossRef]
3. Song, D.; Chang, Q.; Zheng, S.; Yang, S.; Yang, J.; Joo, Y.H. Adaptive Model Predictive Control for Yaw System of Variable-speed Wind Turbines. *J. Mod. Power Syst. Clean Energy* **2021**, *9*, 219–224. [CrossRef]
4. Zhang, Z.Y.; Wang, K.S. Wind turbine fault detection based on SCADA data analysis using ANN. *Adv. Manuf.* **2014**, *2*, 70–78. [CrossRef]
5. Yang, J.; Fang, L.; Song, D.; Su, M.; Yang, X.; Huang, L.; Joo, Y.H. Review of control strategy of large horizontal-axis wind turbines yaw system. *Wind Energy* **2021**, *24*, 97–115. [CrossRef]
6. Zhang, K.; Tang, B.; Deng, L.; Liu, X. A hybrid attention improved ResNet based fault diagnosis method of wind turbines gearbox. *Measurement* **2021**, *179*, 109491. [CrossRef]
7. Encalada-Dávila, Á.; Puruncajas, B.; Tutivén, C.; Vidal, Y. Wind Turbine Main Bearing Fault Prognosis Based Solely on SCADA Data. *Sensors* **2021**, *21*, 2228. [CrossRef] [PubMed]
8. Chen, P.; Li, Y.; Wang, K.; Zuo, M.J.; Heyns, P.S.; Baggeröhr, S. A threshold self-setting condition monitoring scheme for wind turbine generator bearings based on deep convolutional generative adversarial networks. *Measurement* **2021**, *167*, 108234. [CrossRef]
9. Saari, J.; Strombergsson, D.; Lundberg, J.; Thomson, A. Detection and identification of windmill bearing faults using a one-class support vector machine (SVM). *Measurement* **2019**, *137*, 287–301. [CrossRef]
10. Yuan, T.K.; Sun, Z.F.; Ma, S.H. Gearbox Fault Prediction of Wind Turbines Based on a Stacking Model and Change-Point Detection. *Energies* **2019**, *12*, 20. [CrossRef]
11. Aziz, U.; Charbonnier, S.; Bérenguer, C.; Lebranchu, A.; Prevost, F. Critical comparison of power-based wind turbine fault-detection methods using a realistic framework for SCADA data simulation. *Renew. Sustain. Energy Rev.* **2021**, *144*, 110961. [CrossRef]
12. Natili, F.; Daga, A.; Castellani, F.; Garibaldi, L. Multi-Scale Wind Turbine Bearings Supervision Techniques Using Industrial SCADA and Vibration Data. *Appl. Sci.* **2021**, *11*, 6785. [CrossRef]
13. Azzam, B.; Harzendorf, F.; Schelenz, R.; Holweger, W.; Jacobs, G. Pattern Discovery in White Etching Crack Experimental Data Using Machine Learning Techniques. *Appl. Sci.* **2019**, *9*, 5502. [CrossRef]

14. Castellani, F.; Garibaldi, L.; Daga, A.P.; Astolfi, D.; Natili, F. Diagnosis of Faulty Wind Turbine Bearings Using Tower Vibration Measurements. *Energies* **2020**, *13*, 1474. [CrossRef]
15. Xiang, L.; Wang, P.; Yang, X.; Hu, A.; Su, H. Fault detection of wind turbine based on SCADA data analysis using CNN and LSTM with attention mechanism. *Measurement* **2021**, *175*, 109094. [CrossRef]
16. Kordestani, M.; Rezamand, M.; Orchard, M.; Carriveau, R.; Ting, D.S.K.; Saif, M. Planetary Gear Faults Detection in Wind Turbine Gearbox Based on a Ten Years Historical Data From Three Wind Farms. *IFAC-Pap.* **2020**, *53*, 10318–10323.
17. Trizoglou, P.; Liu, X.; Lin, Z. Fault detection by an ensemble framework of Extreme Gradient Boosting (XGBoost) in the operation of offshore wind turbines. *Renew. Energy* **2021**, *179*, 945–962. [CrossRef]
18. Liu, H.; Yu, C.; Yu, C. A new hybrid model based on secondary decomposition, reinforcement learning and SRU network for wind turbine gearbox oil temperature forecasting. *Measurement* **2021**, *178*, 109347. [CrossRef]
19. Pan, Y.; Hong, R.; Chen, J.; Wu, W. A hybrid DBN-SOM-PF-based prognostic approach of remaining useful life for wind turbine gearbox. *Renew. Energy* **2020**, *152*, 138–154. [CrossRef]
20. Ziegler, A.; Koenig, I.R. Mining data with random forests: Current options for real-world applications. *Wiley Interdiscip. Rev. -Data Min. Knowl. Discov.* **2014**, *4*, 55–63. [CrossRef]
21. Asadi, S.; Roshan, S.; Kattan, M.W. Random forest swarm optimization-based for heart diseases diagnosis. *J Biomed. Inf.* **2021**, *115*, 103690. [CrossRef] [PubMed]
22. Kumar, P.; Nair, G.G. An efficient classification framework for breast cancer using hyper parameter tuned Random Decision Forest Classifier and Bayesian Optimization. *Biomed. Signal Process. Control* **2021**, *68*, 102682.
23. Guo, Z.; Yu, B.; Hao, M.; Wang, W.; Jiang, Y.; Zong, F. A novel hybrid method for flight departure delay prediction using Random Forest Regression and Maximal Information Coefficient. *Aerosp. Sci. Technol.* **2021**, *116*, 106822. [CrossRef]
24. Makungwe, M.; Chabala, L.M.; Chishala, B.H.; Lark, R.M. Performance of linear mixed models and random forests for spatial prediction of soil pH. *Geoderma* **2021**, *397*, 115079. [CrossRef]
25. Jain, D.; Singh, V. Feature selection and classification systems for chronic disease prediction: A review. *Egypt. Inform. J.* **2018**, *19*, 179–189. [CrossRef]
26. Wang, J.; Xu, J.; Zhao, C.; Peng, Y.; Wang, H. An ensemble feature selection method for high-dimensional data based on sort aggregation. *Syst. Sci. Control Eng.* **2019**, *7*, 32–39. [CrossRef]
27. Long, W.; Jiao, J.; Liang, X.; Tang, M. Inspired grey wolf optimizer for solving large-scale function optimization problems. *Appl. Math. Model.* **2018**, *60*, 112–126. [CrossRef]
28. Long, W.; Jiao, J.; Liang, X.; Tang, M. An exploration-enhanced grey wolf optimizer to solve high-dimensional numerical optimization. *Eng. Appl. Artif. Intell.* **2018**, *68*, 63–80. [CrossRef]
29. Long, W.; Wu, T.; Xu, M.; Tang, M.; Cai, S. Parameters identification of photovoltaic models by using an enhanced adaptive butterfly optimization algorithm. *Energy* **2021**, *229*, 120750. [CrossRef]
30. Tang, M.; Ding, S.X.; Yang, C.; Cheng, F.; Shardt, Y.A.; Long, W.; Liu, D. Cost-sensitive large margin distribution machine for fault detection of wind turbines. *Clust. Comput.* **2019**, *22*, 7525–7537. [CrossRef]
31. Tang, M.; Chen, Y.; Wu, H.; Zhao, Q.; Long, W.; Sheng, V.S.; Yi, J. Cost-Sensitive Extremely Randomized Trees Algorithm for Online Fault Detection of Wind Turbine Generators. *Front. Energy Res.* **2021**, *9*, 234.
32. McKinnon, C.; Carroll, J.; McDonald, A.; Koukoura, S.; Infield, D.; Soraghan, C. Comparison of New Anomaly Detection Technique for Wind Turbine Condition Monitoring Using Gearbox SCADA Data. *Energies* **2020**, *13*, 5152. [CrossRef]
33. Yang, L.; Zhang, Z. Wind Turbine Gearbox Failure Detection Based on SCADA Data: A Deep Learning-Based Approach. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–11.
34. Corley, B.; Koukoura, S.; Carroll, J.; McDonald, A. Combination of Thermal Modelling and Machine Learning Approaches for Fault Detection in Wind Turbine Gearboxes. *Energies* **2021**, *14*, 1375. [CrossRef]
35. Tang, M.; Zhao, Q.; Ding, S.X.; Wu, H.; Li, L.; Long, W.; Huang, B. An Improved LightGBM Algorithm for Online Fault Detection of Wind Turbine Gearboxes. *Energies* **2020**, *13*, 807. [CrossRef]
36. Mohammed, H.M.; Umar, S.U.; Rashid, T.A. A Systematic and Meta-Analysis Survey of Whale Optimization Algorithm. *Comput. Intell. Neurosci.* **2019**, *2019*, 8718571. [CrossRef]
37. Roy, S.S.; Dey, S.; Chatterjee, S. Autocorrelation Aided Random Forest Classifier-Based Bearing Fault Detection Framework. *IEEE Sens. J.* **2020**, *20*, 10792–10800. [CrossRef]
38. Sheikhi, S. An effective fake news detection method using WOA-xgbTree algorithm and content-based features. *Appl. Soft Comput.* **2021**, *109*, 107559. [CrossRef]
39. Long, W.; Jiao, J.; Liang, X.; Wu, T.; Xu, M.; Cai, S. Pinhole-imaging-based learning butterfly optimization algorithm for global optimization and feature selection. *Appl. Soft Comput.* **2021**, *103*, 107146. [CrossRef]
40. Cho, M.-Y.; Hoang, T.T. Feature Selection and Parameters Optimization of SVM Using Particle Swarm Optimization for Fault Classification in Power Distribution Systems. *Comput. Intell. Neurosci.* **2017**, *2017*, 4135465. [CrossRef]