*Research Article*

# Assessment of Equipment Operation State with Improved Random Forest

**Na Yang** [ID], **She Liu** [ID], **Jie Liu** [ID], **and Changjie Li** [ID]

*School of Mechanical Engineering, Shenyang University of Technology, Shenyang 110870, China*

Correspondence should be addressed to She Liu; liushe@sut.edu.cn

To accurately assess the state of a generator in wind turbines and find abnormalities in time, the method based on improved random forest (IRF) is proposed. The balancing strategy that is a combination of oversampling technique (SMOTE) and undersampling is applied for imbalanced data. Bootstrap is applied to resample original data sets of generator side from the supervisory control and data acquisition (SCADA) system, and decision trees are generated. After the decision trees with different classification capabilities are weighted, an IRF model is established. The accuracy and performance of the model are based on 10-fold cross-validation and confusion matrix. The 60 testing sets are assessed, and the accuracy is 95.67%. It is more than 1.67% higher than traditional classifiers. The probabilities of 60 data sets at each class are calculated, and the corresponding state class is determined. The results show that the proposed IRF has higher accuracy, and the state can be assessed effectively. The method has a good application prospect in the state assessment of wind power equipment.

## 1. Introduction

In recent years, the number of wind turbines is increasing with the extensive application of wind energy. The generator is the key component of wind turbines. So, its operation state has a direct impact on the power generation of wind turbines. Generally, wind farms are built in complex and harsh places such as the Gobi. The wind turbine generator has failed frequently under the influence of extremely harsh working environment, complex working conditions, and extreme weather [1, 2]. A major problem is maintenance which is difficult and expensive. So, it is crucial to assess the operation state of a wind turbine generator.

There is a growing body of scholars that recognized the importance of state assessment in the wind turbine generator. Some methods such as fuzzy comprehensive evaluation, support vector machine (SVM), and neural networks (NN) are main methods to assess. At the same time, the SCADA system has been installed in many wind farms. It is a powerful way to obtain operating data of wind turbines. Extensive research has shown that state parameter data from the SCADA system is critical. It is

related to the operation state of wind turbines, and the state assessment is realized. Wang et al. established an evaluative model based on fuzzy mathematics to assess the design performance of wind turbines comprehensively, but its operating state was not assessed effectively [3]. Zheng et al. considered randomness into the fuzzy method. Combination weight was applied to determine the index weight for higher accuracy, and the health state was assessed effectively [4]. An et al. comprehensively considered multisource information such as wind speed and rotational speed. The experimental data of different faults was verified based on SVM. Finally, fault diagnosis of the wind turbine was realized [5]. Liang and Fang considered the coupling relationship among components of the wind turbine and established a regression prediction model with SVM [6]. Lin et al. proposed an adaptive immune fruit fly optimization algorithm (AIFOA) to optimize the parameters of SVM. The feature index was predicted more accurately. And performance assessment was realized comparing deviation with normal value [7]. With the usage of the traditional SCADA alarm system, the fault diagnosis was more convenient. Li et al. established a normal behavior model based on NN to assess the wind

turbine operation state. The health class was proposed to measure differences between operating state and normal state in the paper. Finally, the wind power generation system was effectively assessed [8]. Wang proposed a two-level NN recognition method used for fault classification and fault diagnosis, respectively [9]. Zhao et al. proposed a deep learning method (DLM) with a deep autoencoder (DAE) network and established the DAE model. SCADA data was input, and early warning of fault components was realized [10]. Yang et al. established an assessment index system for wind turbines firstly. Quantitative assessment of main components such as blades and generator bearings in wind turbines was realized based on SCADA data [11]. Tautz and Watson realized state and fault monitoring of wind turbines based on SCADA data and five aspects which concluded clustering methods, normal behavior modeling, damage modeling, and expert system review estimate [12]. Hu et al. proposed an evaluation method based on temperature characteristic parameters and deterioration degree function. The early deterioration of the wind turbine generator system was detected successfully [13]. Qian et al. proposed an online sequential extreme learning machine (OS-ELM) algorithm for wind turbine condition monitoring. The long-term deterioration characteristics and the short-term faults of the gearbox were detected efficiently based on SCADA data and the proposed method [14]. Hsu et al. regarded control charts based on an exponentially weighted moving average (EWMA) model as a main assessment method and set upper and lower limits to monitor state variables. The operation state of wind turbines could be reacted at all times, but the data was limited in this process [15].

These research findings provide the theoretical and experimental foundations for assessment of wind turbines or their core components. However, some problems exist. Fuzzy comprehensive assessment is defective in determination of index weight; it is highly subjective. SVM is suitable for classification based on the small size of samples, but it is difficult for the large size of samples. NN is simple to learn and implement compared with the first two methods, but results must be obtained accurately based on a large number of data sets. Importantly, the obtained results are highly dependent on the parameters, and a lot of work and experience are required in the process of fine tuning these parameters require [16].

However, ensemble learning is a learning paradigm where many classifiers are combined to solve a problem. The generalization ability of a single classifier can be significantly improved based on a classifier ensemble. For example, random forest (RF) is widely applied because of strong classification ability, strong learning ability, and no requirements for samples. Importantly, the method is more suitable for classification or regression problems with less noise. It is insensitive to the adjustment of parameters. Classification attributes are not divided too much, and data dimensions should be under tens of dimensions. Wang et al. proposed a panoramic crack detection method based on structured RF to realize condition monitoring and fault diagnosis. Finally, the surface cracks of a panoramic steel beam were found efficiently [17]. But the same weight is given to decision trees with different classification capabilities in the final voting stage of RF, which weakened classification performance.

Therefore, a state assessment method based on IRF is proposed in the paper. Undersampling and SMOTE are introduced for imbalanced data sets. Right weight is introduced to the final voting stage of RF. To improve accuracy, different weights are set according to the different capabilities of decision trees. 10-fold cross-validation and improved assessment criteria based on a confusion matrix are applied for model assessment. Finally, the method is verified based on data sets from the SCADA system in wind turbines. The state of generators is assessed correctly. The efficiency of the method is verified compared with traditional classifiers.

## 2. Method: Improved Random Forest

*2.1. Processing for Imbalanced Data.* Unbalanced data mean that some categories have a large number of samples and others have a small number of samples, which forms an imbalance of each category in the data sets. Generally, a small number of samples is called a minority. It is easy to be misclassified with a small number of categories, and classification accuracy is bad because of imbalance. At present, the imbalance is solved by data processing and algorithms. The data processing is that increasing minority samples is based on undersampling or reducing majority of samples based on oversampling. However, oversampling is a copy of a minority of samples, which causes overlap and overfitting of data sets. An undersampling method is to randomly delete some data to balance samples, which causes some important sample information lost, and the result is affected. Someone has studied the combination of two methods, and good results are obtained, so the method combined between undersampling and Synthetic Minority Oversampling Technique (SMOTE) is introduced in the paper [18].

SMOTE is linear interpolation realized between a few neighboring samples to synthesize new minority samples. That is, $k$ (usually 5)-nearest neighbor samples are found for each sample in the minority data sets. According to the sampling magnification $N$, $N$ same kind of samples are randomly selected from $k$-nearest neighbor samples, it is $y_1, y_2, \cdots, y_N$. The linear interpolation is realized between $x$ and $y_i(1, 2, \cdots, N)$ of minority samples to synthesize new minority samples new_data. The formula is expressed as follows:

$$\text{new\_data} = x + \text{rand} * (y_i - x)(i = 1, 2, \cdots), \qquad (1)$$

where rand is a random number between $(0, 1)$. new_data is new samples.

### 2.2. Random Forest

*2.2.1. Establishment of RF Model.* RF is a classifier integration algorithm that combines "random subspace method" and "bootstrap aggregate" to establish decision trees (DT). RF is established as follows [19].

The bootstrap resampling method is applied to allocate training set and testing set. The original sample sets are replaced and randomly sampled $N$ times to form a new
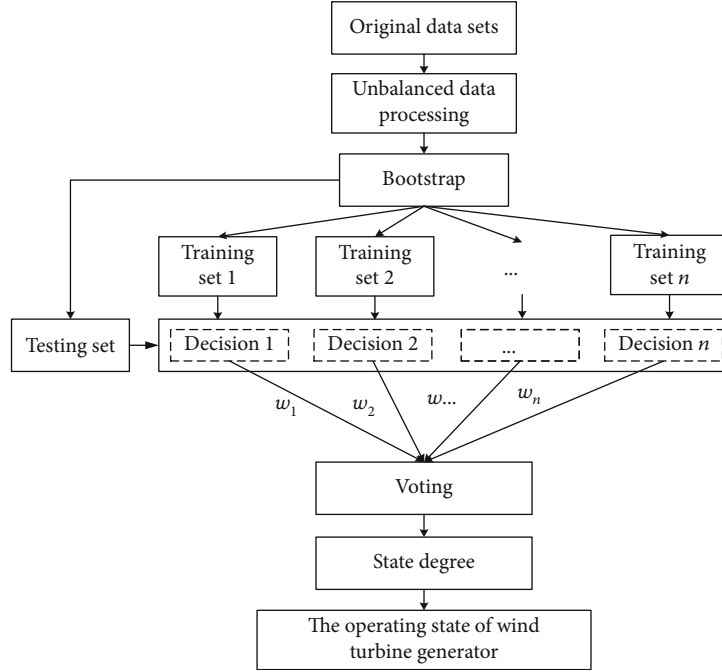
FIGURE 1: The assessment flowchart of IRF.

training set that is the same size as the original sample sets. According to the distribution of probability, 60+% of samples are repeatedly collected as the training set. About 36% of samples not collected are the testing set.

Each generated training set is applied to establish a corresponding decision tree $C_1, C_2, \cdots, C_n$. $m(m \leq M)$ attributes are extracted from each node of decision trees and as the split attribute of the current node for classification. During the growth of an entire forest, $m$ is determined by the Gini index of each node. The Gini index indicates impurity of each node. The purity is inversely proportional to the Gini index. The formula is expressed as follows:

$$\text{Gini}(X) = 1 - \sum_{j=1}^{m} \left( P_j \right)^2, \qquad (2)$$

where $P_j$ is probability that the sample $X$ contains the attribute $j$.

The input testing set is verified by each decision tree, and results are obtained based on the number of votes. The RF model is expressed as follows [20]:

$$H(x) = \arg \max_Y \sum_i^n I(h_i(x) = Y), \qquad (3)$$

where $h_i(x) = Y$ is output by the $i$th decision tree. $Y = 1, 2, \cdots, c$ is the corresponding category, $i = 1, 2, \cdots, n$, $n$ is the number of decision trees in the random forest. $I(\cdot)$ is the indicator function. $\cdot$ is classed correctly; its value is 1; otherwise, it is 0.

The number of decision trees in RF has crucial influences on generalization ability. The data sets that are not extracted are set to $Q$. $Q$ is input to the above RF model; the corresponding classified result is output. The number of incorrect classification is set to $R$, so the probability of incorrect classification about $Q$ is $E_{\text{OOB}} = R/Q$. Finally, the number of decision trees is determined effectively [21].

*2.2.2. Improvement Process of RF.* To avoid the effect that traditional RF makes on assessment accuracy owing to the same weight for each decision tree. Weight is introduced to the voting process. Different weights are endowed to different decision trees, and generalization ability is improved. The formula of weight is expressed as follows:

$$w_i = \frac{X'_{\text{correct},i}}{X'}, \quad i = 1, 2, 3, \cdots, n, \qquad (4)$$

where $X'$ is the pretested sample (it is part of the training set) and $X'_{\text{correct},i}$ is the number of samples classed correctly.

The improved RF model is expressed as follows:

$$H(x) = \arg \max_Y \left\{ \sum_i^n I(h_i(x) = Y) w_i \right\}. \qquad (5)$$

The assessment process of the wind turbine generator with IRF is shown in Figure 1. The setting of weight is shown in Figure 2.

*2.3. Online Assessment Strategy.* To assess errors caused by noise, online assessment is introduced. The class of operation state is set to $c$; data at time $t$ is input to IRF model $H(x)$. The

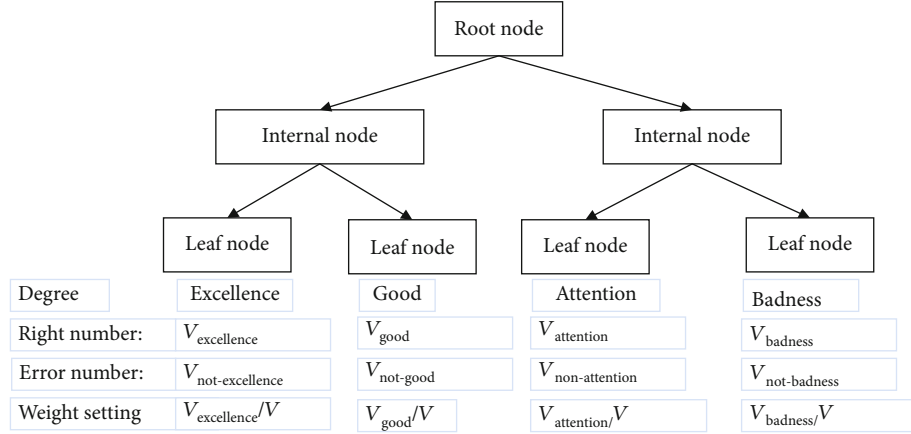| Degree | Excellence | Good | Attention | Badness |
|---|---|---|---|---|
| Right number: | $V_{\text{excellence}}$ | $V_{\text{good}}$ | $V_{\text{attention}}$ | $V_{\text{badness}}$ |
| Error number: | $V_{\text{not-excellence}}$ | $V_{\text{not-good}}$ | $V_{\text{non-attention}}$ | $V_{\text{not-badness}}$ |
| Weight setting | $V_{\text{excellence}}/V$ | $V_{\text{good}}/V$ | $V_{\text{attention}}/V$ | $V_{\text{badness}}/V$ |

FIGURE 2: The setting of voting weight for improved random forest.

voting results $Y_i$ of each decision tree are output, so degree probability of the $c$th operation state is expressed as follows:

$$P(c|x_t) = \frac{\sum_n I(Y_i = c)w_i}{\sum_c \sum_n I(Y_i = c)w_i}. \tag{6}$$

The state degrees of the generator are converted between adjacent degrees. Finally, the corresponding state degree of the data $x_t$ is expressed as follows:

$$c_t{}^* = \arg\left\{ \max_{Y=1,2,\cdots,c} P(c|x_t) \right\}. \tag{7}$$

### 2.4. Assessment of Model Performance

#### 2.4.1. K-Fold Cross-Validation.
To assure statistical significance in the results, 10-fold cross-validation is applied to verify the model performance. The samples are randomly divided into 10 servings to ensure the uniform distribution of the data sets as much as possible. Nine of them are taken as the training set, and the remaining 1 is taken as the test set. 10 times are repeated, and the results are the average of 100 values.

#### 2.4.2. Confusion Matrix.
Generally, classification accuracy (ACC) is used as an assessment standard of the model. But the model performance is often ignored for the minority samples from unbalanced data sets. So, a confusion matrix is introduced in the paper [22]. The relationship between the true category of the samples and the classification result is described by the confusion matrix to present the assessment standard of model performance. The confusion matrix is shown in Table 1.

$N$ is the majority class; $P$ is the minority class; $TP, TN$ is the number of majority classes and minority classes classified correctly, respectively; and $FN, FP$ is the number of majority classes and minority classes misclassified.

To assess the classification model more accurately, harmonic average of minority class accuracy and $F-$measure, geometric average correct rate ($G-$mean), and Matthew correlation coefficient (MCC) are determined as the assess-

ment standard of the model based on the confusion matrix. The specific formula is expressed as follows:

$$\text{recall} = \frac{TP}{TP + FN},$$

$$\text{specificity} = \frac{TN}{FP + TN},$$

$$\text{precision} = \frac{TP}{FP + TP},$$

$$G-\text{mean} = \sqrt{\text{recall} * \text{specificity}},$$

$$F-\text{measure} = \frac{2 * \text{recall} * \text{precision}}{\text{recall} + \text{precision}},$$

$$\text{MCC} = \frac{TP * TN - FN * FP}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}},$$

$$\text{ACC} = \frac{TP + TN + FN}{TP + FP + TN + FN}. \tag{8}$$

However, $F-$measure, $G-$mean, and MCC are only applicable for binary problems. A "one vs. one" strategy is introduced to solve multiclass problems in the paper. That is, multiclass is paired in pairs; the multiclass problem is converted into binary problems. The average is finally taken as a result. So, the improved assessment standard is applied in the paper. The formula is expressed as follows:

$$\overline{F-\text{measure}} = \frac{\sum_{i=1}^{N} F-\text{measure}_i}{N},$$

$$\overline{G-\text{mean}} = \left[ \prod_{i=1}^{N} (\text{recall}_i * \text{specificity}_i) \right]^{1/N},$$

$$\overline{\text{MCC}} = \frac{\sum_{i=1}^{N} N_i \times \text{MCC}_i}{N},$$

$$\overline{\text{ACC}} = \frac{\sum_{i=1}^{N} N_i \times \text{MCC}_i}{N}. \tag{9}$$

TABLE 1: Confusion matrix.

| Real value | Classification value | |
|---|---|---|
| | Positive instances | Negative instances |
| Positive instances | True positive ($TP$) | False positive ($FN$) |
| Negative instances | False positive ($FP$) | True positive ($TN$) |

TABLE 2: Details of original data sets.

| Samples | Features | Class | Class distribution | Imbalance rate |
|---|---|---|---|---|
| 1266 | 9 | 4 | 354/457/403/52 | 8.79 |
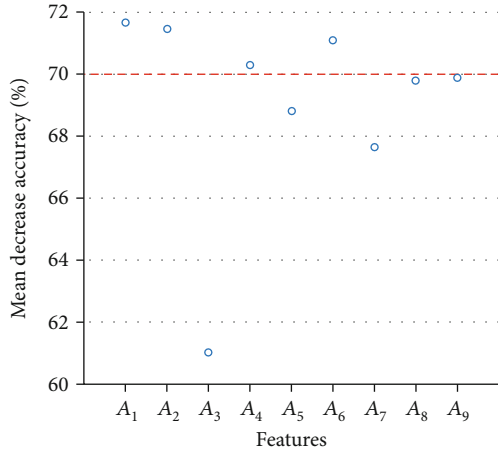


FIGURE 3: Importance of each feature based on Gini index.



FIGURE 4: The fitness of RF.

TABLE 3: Generator feature data sets from SCADA system.

| Features | Time | | | | | |
|---|---|---|---|---|---|---|
| | 4:30 | 6:30 | 8:30 | 10:30 | 12:30 | 14:00 |
| $A_1$ | 44.0 | 46.6 | 50.7 | 48.7 | 55.1 | 58.5 |
| $A_2$ | 44.4 | 46.9 | 49.7 | 47.8 | 49.8 | 52.3 |
| $A_3$ | 54.2 | 39.5 | 44.0 | 44.7 | 45.5 | 48.2 |
| $A_4$ | 78.6 | 77.8 | 84.0 | 81.5 | 81.6 | 85.3 |
| $A_5$ | 75.5 | 75.3 | 81.0 | 76.5 | 78.5 | 83.2 |
| $A_6$ | 78.1 | 77.2 | 82.9 | 79.0 | 80.9 | 85.9 |
| $A_7$ | 77.3 | 76.7 | 82.5 | 78.6 | 79.1 | 85.8 |
| $A_8$ | 78.6 | 79.0 | 84.2 | 80.2 | 80.7 | 87.3 |
| $A_9$ | 77.4 | 76.3 | 82.6 | 78.8 | 80.3 | 84.7 |

## 3. Simulation Results and Discussion

### 3.1. Preparation for Simulation

*3.1.1. Data Collection.* The data sets in the paper come from the SCADA system of wind turbines in a wind farm. The SCADA is a distributed control system (DCS) and power automation monitoring system based on a computer. It achieves data collection, equipment control, measurement, and parameter adjustment of core components in wind turbines such as generators, gearboxes, and blades by monitoring and controlling on-site equipment, namely, wind turbines. The purpose is to correctly grasp the state of system and each component, to make a decision quickly, to help diagnose the fault state, and so on. The F7 wind turbine failed in the wind farm at 14:01 on July 1, 2017. The generator data sets are collected before failure from the SCADA system in the paper. The data details are shown in Table 2, which includes the number of samples, features, classes, class distribution, and imbalance rate. The imbalance rate is obtained by the largest sample and the smallest sample.

*3.1.2. The Setting of Assessment Features.* According to features related to the generator in the SCADA system, nine feature data sets of generator that have a greater impact are determined, respectively, $A_1$: front shaft temperature;
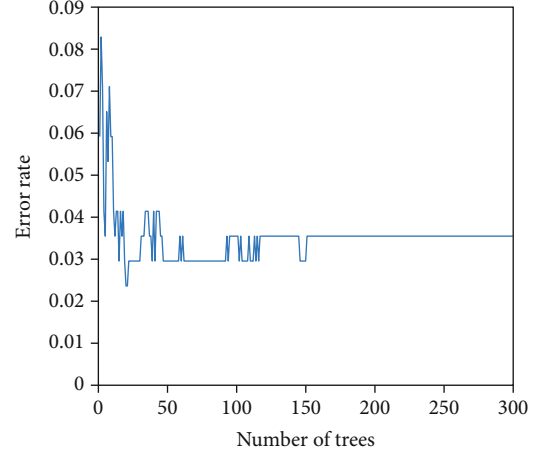
$A_2$: rear bearing temperature; $A_3$: cooling water inlet temperature; $A_4$: $u_1$ winding temperature; $A_5$: $u_2$ winding temperature; $A_6$: $v_1$ winding temperature; $A_7$: $v_2$ winding temperature; $A_8$: $w_1$ winding temperature; and $A_9$: $w_2$ winding temperature.

*3.1.3. The Setting of State Degree.* Generally, it is appropriate to divide state degrees into 3-5. In this paper, the state degrees of the wind turbine generator are finally determined as 4, namely, "excellent," "good," "attention," and "badness."

*3.1.4. The Setting of Optimal Characteristic.* The mean decrease accuracy is calculated based on the Gini index (as shown in Figure 3). The number of optimal characteristics is the same as above 70%. Namely, the number of optimal characteristics is set to 4 in the paper, and decision trees are branched.

As shown in Figure 4, the out-of-band (OOB) error rate decreases as the number of decision trees increases. After $n > 150$, the OOB error rate remains stable and below 4%. To assess more accurately, it is set to $n = 200$.
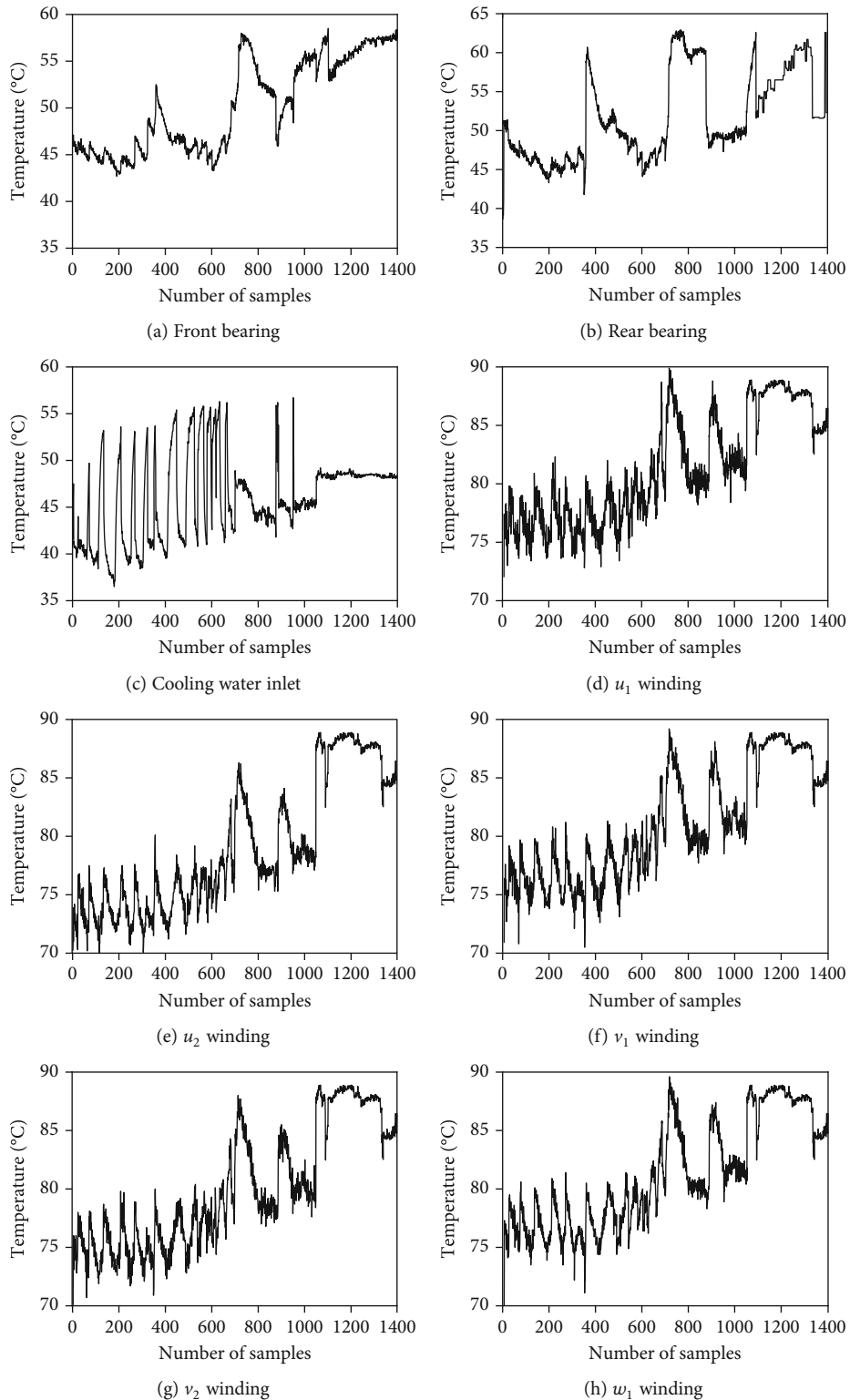
(a) Front bearing



(b) Rear bearing



(c) Cooling water inlet



(d) $u_1$ winding



(e) $u_2$ winding



(f) $v_1$ winding



(g) $v_2$ winding



(h) $w_1$ winding

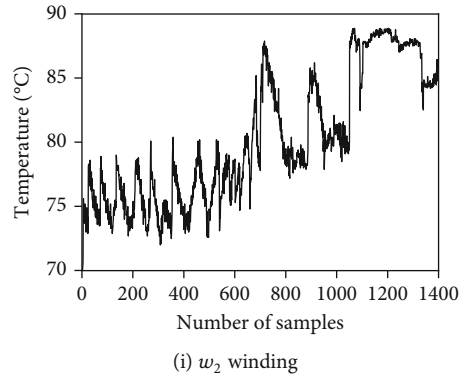Figure 5: Continued.

(i) $w_2$ winding

FIGURE 5: The temperature of features in wind turbine generator.

TABLE 4: The assessment results of IRF based on different sizes of data sets.

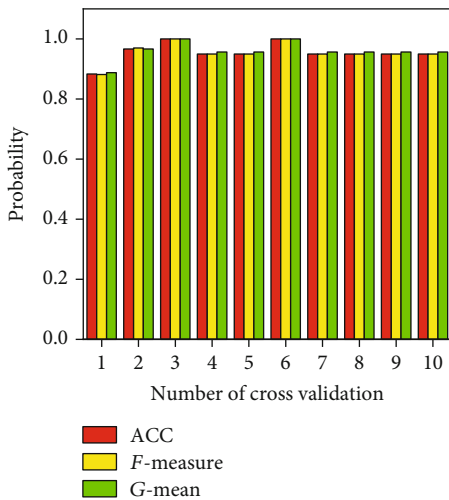| The size of data sets | Classes | Size | Average number of corrected assessment | Assessment standard | | |
|---|---|---|---|---|---|---|
| | | | | $\overline{ACC}$ | $\overline{F-measure}$ | $\overline{G-mean}$ |
| 1400 | Excellence | 35 | 32 | | | |
| | Good | 35 | 31 | | | |
| | Attention | 35 | 35 | 0.9529 | 0.9536 | 0.9546 |
| | Badness | 35 | 35 | | | |
| 1000 | Excellence | 25 | 23 | | | |
| | Good | 25 | 23 | | | |
| | Attention | 25 | 25 | 0.9560 | 0.9548 | 0.9615 |
| | Badness | 25 | 24 | | | |
| 600 | Excellence | 15 | 13 | | | |
| | Good | 15 | 14 | | | |
| | Attention | 15 | 15 | 0.9567 | 0.9545 | 0.9600 |
| | Badness | 15 | 15 | | | |



FIGURE 6: The result based on 10-fold cross-validation.

3.2. State Assessment and Analysis of Simulation Results. The proposed model in the paper is verified based on the data sets in Table 2. The data sets are imbalanced for the classification problem. Part of data sets is shown in Table 3.

TABLE 5: Comparison of different classifier results.

| Classifiers | Assessment standard | | | |
|---|---|---|---|---|
| | $\overline{ACC}$ | $\overline{F-measure}$ | $\overline{G-mean}$ | $\overline{MCC}$ |
| IRF | 0.9567 | 0.9545 | 0.9600 | 0.9159 |
| RF | 0.9400 | 0.9379 | 0.9428 | 0.8824 |
| LVQ | 0.9400 | 0.9386 | 0.9502 | 0.8926 |
| SVM | 0.9067 | 0.9046 | 0.9188 | 0.8282 |
| DT | 0.9083 | 0.9013 | 0.9283 | 0.8380 |
| PNN | 0.9083 | 0.9045 | 0.9245 | 0.8358 |

The feature distribution is shown in Figure 5 after unbalanced data set processing based on a combination of under-sampling and SMOTE.

In the paper, a simulation testing is completed on MATLAB R2016a. Firstly, the impact of the sample size is analyzed on the accuracy of the model. The total number of samples is determined as 1400, 1000, and 600. The accuracy of the improved random forest model is verified by 10-fold cross-validation and improved assessment criteria. The results are shown in Table 4.

TABLE 6: Definition of all the symbols used in the paper.

| | |
|---|---|
| $A_1$: front bearing temperature | $I(\cdot)$: indicator function |
| $A_2$: rear bearing temperature | $k$: 5 number of nearest neighbor samples |
| $A_3$: cooling water inlet temperature | $M$: the number of attributes |
| $A_4$: $u_1$ winding temperature | $m$: the number of optimal attributes |
| $A_5$: $u_2$ winding temperature | MCC: Matthew correlation coefficient |
| $A_6$: $v_1$ winding temperature | $\overline{MCC}$: average of MCC |
| $A_7$: $v_2$ winding temperature | $N$: the number of data sets |
| $A_8$: $w_1$ winding temperature | $n$: the number of decision trees |
| $A_9$: $w_2$ winding temperature | new_data: new samples |
| ACC: classification accuracy | $P_j$: $j = 1$-$m$, probability that the sample $X$ contains the attribute $j$ |
| $\overline{ACC}$: average of ACC | $P(c\|x_t)$: probability of $x_t$ being in $c$ degree |
| $C_n$: the $n$th decision tree | $Q$: the number of data sets not extracted |
| $c$: state degree | $R$: the number of incorrect classification from data sets not extracted |
| $c_t{}^*$: state degree at time $t$ | $t$: time |
| $E_{OOB}$: error rate of classification | $w_i$: $i = 1$-$n$, weight for the $i$th decision tree |
| FN: false positive | $x_i$: data value of generator at time $t$ |
| FP: false positive | $X'_{correct,i}$: $i = 1$-$n$, the number of samples classified correctly for the $i$th tree |
| $F$ – measure: harmonic average of minority class accuracy and recall | $X'$: the number of pretested sample |
| $\overline{F - measure}$: average of $F$ – measure | $x$: some sample |
| Gini$(X)$: Gini index | $y_i$: samples are randomly selected from nearest neighbor samples |
| $G$ – mean: geometric average correct rate | rand: random number between (0,1) |
| $\overline{G - mean}$: average of $G$ – mean | TP: true positive |
| $H(x)$: comprehensive assessment result | TN: true positive |
| $h_i(x)$: $i = 1$-$n$, assessment result of $i$th decision tree | $Y$: the corresponding category |

According to the assessment results in Table 4, it is not significant for the number of the training set to affect accuracy of the assessment model, and final assessment accuracy will fluctuate around 95%-96%. Meantime, the assessment results for 10-fold cross-validation based on 600 data sets are shown in Figure 6.

To further verify that IRF has higher generalization ability and classification ability on state assessment of a generator, in the same condition, DT, RF, Probabilistic Neural Network (PNN), Learning Vector Quantization (LVQ), and SVM are adopted separately for training and testing. The data set size of 600 is taken as an example, and comparison of the assessment accuracy of different classifiers based on 10-fold cross-validation and confusion matrix is shown in Table 5.

As you can see from Table 5, IRF is significantly improved based on the RF. The assessment standard is sorted by model as $\overline{ACC}$: IRF > RF = LVQ > DT = PNN > SVM, $\overline{F - measure}$: IRF > LVQ > RF > SVM > PNN > DT, $\overline{G - mean}$: IRF > LVQ > RF > DT > PNN > SVM, and $\overline{MCC}$: IRF > LVQ > RF > DT > PNN > SVM. The results are analyzed comprehensively. Average accuracy is higher by 1.67% than RF. The IRF model has the best performance, which reflects that it has good prediction accuracy, extrapolation ability,

and robust performance. Definition of all the symbols used in the paper is shown in Table 6.

Meanwhile, 60 testing sets are used to calculate the probability of state classes. The results of probability are shown in Figure 7.

In Figure 7, the probability of "excellent" is the largest for the first 13 samples. From the 14th sample, the probability of "good" gradually increases. From the 14th sample to the 23rd sample, the probability of "good" is the largest. From the 24th sample, the probability of "good" gradually decreases, and the probability of "attention" gradually increases. From the 31th sample to the 37th sample, the probability of "attention" is the largest. From the 38th sample, the probability of "attention" gradually decreases, and the probability of "badness" gradually increases. From the 46th sample to the 60th sample, the probability of "badness" is the largest.

The state assessment class and original class are shown in Figure 8. The assessment results of 60 data sets are determined: 1st-13th: excellent; 14th-30th: good; 31st-45th: attention; 46th-60th: badness. That is, the original testing set at "excellent" class is mistakenly classified as good twice, and original testing set at "good" class is mistakenly classified as "excellent" class once. The accuracy of results reaches
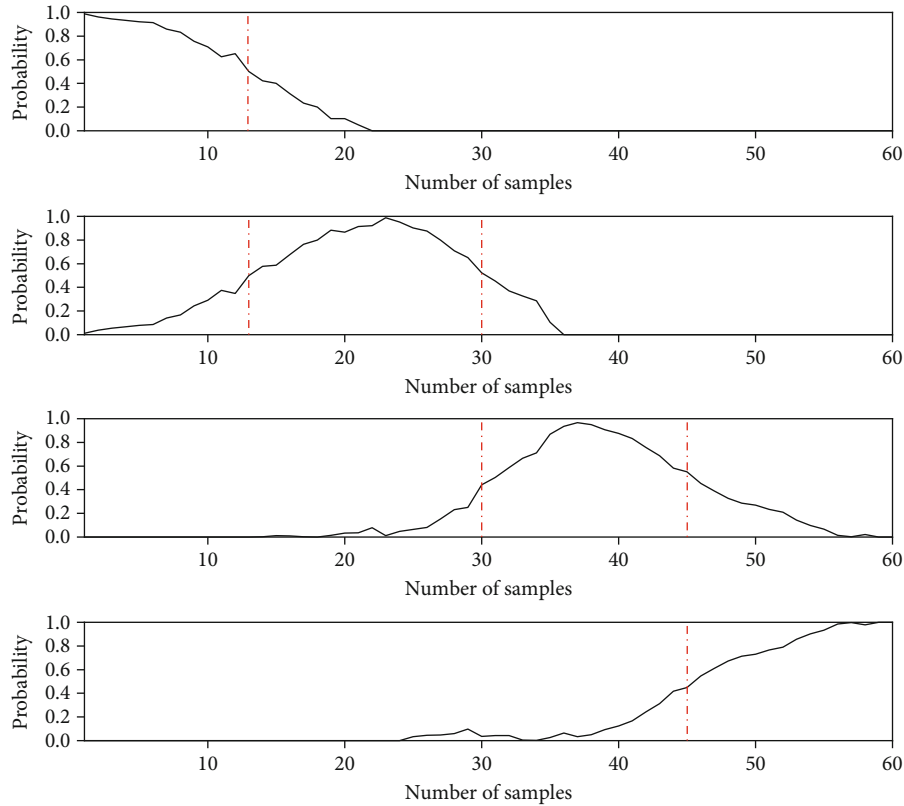
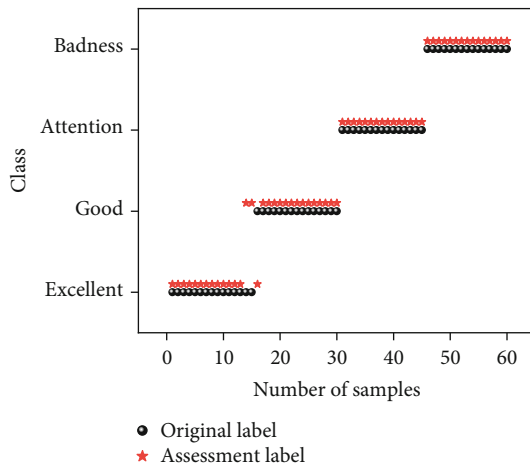FIGURE 7: The probability of each state degree.



FIGURE 8: Assessment degree of operating state.

95.67% compared with the state degree of real operation. So, the effectiveness of the proposed method is verified.

## 4. Conclusion

The state assessment method for a wind turbine generator based on improved random forest (IRF) is proposed. Firstly, data sets containing nine features of generator are determined from the supervisory control and data acquisition (SCADA) system. Undersampling and Synthetic Minority Oversampling Technique (SMOTE) is introduced to solve the imbalanced data problems in the paper. Bootstrap is applied to resample original data sets, and then, decision trees are generated. The weight is determined according to classification performance of different decision trees. The IRF model and corresponding online evaluation strategy are established. Finally, 60 data sets selected are input to verify the established model based on 10-fold cross-validation and confusion matrix. The state of wind turbine generator is assessed correctly, and then, the same data sets are applied to realize online assessment. The accuracy reaches 95.67%. The proposed method can not only ensure accuracy and effectiveness of assessment but also improve efficiency. The accuracy of the proposed method is better than traditional classifiers. It provides a certain reference for state assessment of wind power equipment.

## Data Availability

All the data is in the manuscript. If the researchers are interested in obtaining the numerical solution files, please contact the email address: skyyangna@126.com.

## Conflicts of Interest

The authors declared no potential conflicts of interest to the research, authorship, and publication of this article.

## Acknowledgments

## References

[1] X. F. Chen, J. M. Li, H. Cheng, B. Li, and Z. He, "Research and application of condition monitoring and fault diagnosis technology in wind turbines," *Journal of Mechanical Engineering*, vol. 47, no. 9, pp. 45–52, 2011.

[2] H. Zhao, H. Liu, H. Liu, and Y. K. Lin, "Condition monitoring and fault diagnosis of wind turbine generator based on stacked autoencoder network," *Automation of Electric Power Systems*, vol. 42, no. 11, pp. 102–108, 2018.

[3] W. Zhiguo, M. Yitai, Y. Zhao, and L. Wei, "Fuzzy comprehensive evaluation method of wind power generation unit," *Acta Energiae Solaris Sinica*, vol. 25, no. 2, pp. 177–181, 2004.

[4] K. Zheng, L. N. Han, S. L. Guo, Z. Wang, X. Zhang, and X. Dong, "Fuzzy synthetic condition assessment of wind turbine based on combination weighting and cloud model," *Journal of Intelligent and Fuzzy Systems*, vol. 32, no. 6, pp. 4563–4572, 2017.

[5] X. L. An, M. H. Zhao, D. X. Jiang, and S. Li, "Direct-drive wind turbine fault diagnosis based on support vector machine and multi-source information," *Power System Technology*, vol. 35, no. 4, pp. 117–122, 2011.

[6] Y. Liang and R. M. Fang, "An online wind turbine condition assessment method based on SCADA and support vector regression," *Automation of Electric Power Systems*, vol. 37, no. 14, pp. 7–12+31, 2013.

[7] T. Lin, L. Zhang, R. Q. Cai, X. Yang, G. Liu, and W. Z. Liao, "Wind turbine performance evaluation based on support vector machine optimized by improved fruit fly optimization algorithm," *Renewable Energy Resources*, vol. 37, pp. 1132–1137, 2019.

[8] J. Li, X. Lei, H. Li, and L. Ran, "Normal behavior models for the condition assessment of wind turbine generator systems," *Electric Power Components and Systems*, vol. 42, no. 11, pp. 1201–1212, 2014.

[9] X. Y. Wang, "Fault diagnosis on transmission system of wind turbines based on wavelet packet transform and RBF neural networks," *ICMRA*, vol. 373-375, pp. 1102–1105, 2013.

[10] H. S. Zhao, H. H. Liu, W. J. Hu, and X. Yan, "Anomaly detection and fault analysis of wind turbine components based on deep learning network," *Renewable Energy*, vol. 127, pp. 825–834, 2018.

[11] W. X. Yang, R. Court, and J. S. Jiang, "Wind turbine condition monitoring by the approach of SCADA data analysis," *Renewable Energy*, vol. 53, pp. 365–376, 2013.

[12] W. J. Tautz and S. J. Watson, "Using SCADA data for wind turbine condition monitoring – a review," *IET Renewable Power Generation*, vol. 11, no. 4, pp. 382–394, 2017.

[13] Y. Hu, H. Li, X. Liao, E. Song, H. Liu, and Z. Chen, "A probability evaluation method of early deterioration condition for the critical components of wind turbine generator systems," *Mechanical Systems and Signal Processing*, vol. 76-77, pp. 729–741, 2016.

[14] P. Qian, X. Ma, and D. Zhang, "Estimating health condition of the wind turbine drivetrain system," *Energies*, vol. 10, no. 10, pp. 1583–1583, 2017.

[15] H. Y. Hsu, L. H. Mei, and M. L. Chun, "An approach combining data mining and control charts - based model for fault detection in wind turbines," *Renewable Energy*, vol. 115, pp. 808–816, 2018.

[16] A. Bustillo, J. F. Díez-Pastor, G. Quintana, and C. García-Osorio, "Avoiding neural network fine tuning by using ensemble learning: application to ball-end milling operations," *International Journal of Advanced Manufacturing Technology*, vol. 57, no. 5-8, pp. 521–532, 2011.

[17] S. Wang, X. Liu, T. Yang, and X. Wu, "Panoramic crack detection for steel beam based on structured random forests," *IEEE Access*, vol. 6, pp. 16432–16444, 2018.

[18] A. Bustillo and J. J. Rodriguez, "Online breakage detection of multitooth tools using classifier ensembles for imbalanced data," *International Journal of Systems Science*, vol. 45, pp. 2590–2602, 2014.

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

[20] S. Li, D. Chen, Q. Qiu et al., "Comprehensive assessment of power quality based on random forest," *Modern Electric Power*, vol. 36, pp. 81–87, 2019.

[21] X. Gao, J. Wen, and C. Zhang, "An improved random forest algorithm for predicting employee turnover," *Mathematical Problems in Engineering*, vol. 2019, Article ID 4140707, 12 pages, 2019.

[22] P. Santos, J. Maudes, and A. Bustillo, "Identifying maximum imbalance in datasets for fault diagnosis of gearboxes," *Journal of Intelligent Manufacturing*, vol. 29, pp. 331–351, 2018.