energies

MDPI

Article

Ice Detection Model of Wind Turbine Blades Based on Random Forest Classifier

Lijun Zhang [1,*], Kai Liu [1], Yufeng Wang [1] and Zachary Bosire Omariba [1,2]

[1] National Center for Materials Service Safety, University of Science and Technology Beijing, Beijing 100083, China; s20161185@xs.ustb.edu.cn (K.L.); s20171176@xs.ustb.edu.cn (Y.W.); zomariba@egerton.ac.ke (Z.B.O.)

[2] Computer Science Department, Egerton University, Egerton 20115, Kenya

* Correspondence: ljzhang@ustb.edu.cn; Tel.: +86-10-6232-1017

Received: 14 September 2018; Accepted: 21 September 2018; Published: 25 September 2018

Abstract: When wind turbine blades are icing, the output power of a wind turbine tends to reduce, thus informing the selection of two basic variables of wind speed and power. Then other features, such as the degree of power deviation from the power curve fitted by normal sample data, are extracted to build the model based on the random forest classifier with the confusion matrix for result assessment. The model indicates that it has high accuracy and good generalization ability verified with the data from the China Industrial Big Data Innovation Competition. This study looks at ice detection on wind turbine blades using supervisory control and data acquisition (SCADA) data and thereafter a model based on the random forest classifier is proposed. Compared with other classification models, the model based on the random forest classifier is more accurate and more efficient in terms of computing capabilities, making it more suitable for the practical application on ice detection.

Keywords: ice detection; wind turbine blades; SCADA data; random forest classifier; power curve; confusion matrix

1. Introduction

With the gradual depletion of traditional fossil fuels such as coal, oil and natural gas, the development and use of new energy such as wind power has received increasing attention making wind power one of the fastest growing energy sources in the world [1]. In 2017, the newly installed capacity of wind power worldwide reached 52,492 MW, and the cumulative installed capacity reached 539,123 MW. Among them, the newly installed capacity of wind power in China accounted for 37%, and the cumulative installed capacity accounted for 35%. The newly installed capacity of wind power accounts for more than 15% of the total installed capacity in recent years, and the cumulative installed capacity accounts for a steady increase. The Global Wind Energy Council (GWEC) predicts that as costs drop and the market begins to recover at the end of this decade; global wind power installed capacity will increase by more than 50% over the next five years. According to GWEC, as countries around the world develop renewable energy sources to achieve emission reduction targets, wind energy costs continue to decline, and by the end of 2022 installed capacity global wind power is expected to increase to 840 GW [2].

Wind power however faces some challenges that restrict its development with cost making of the list of the important issue. According to the study of Department of Energy (DOE), United States, 20% revenue growth of wind farms by 2030 will come from improvement of wind turbine working status and reduction of maintenance costs. Using the appropriate maintenance and maintenance strategy to reduce the cost of operation and maintenance is an important way to increase wind farm income [3].

Land-based wind farms are established mostly based on high altitude mountainous areas. These regions experience low temperature and high humidity, which makes it possible for wind turbine blades to form varying degrees of icing easily, especially in winter. However, there are shutdown events caused by wind turbine blade icing, which seriously threatens the normal operation of wind power plants. Any wind turbine blade icing will cause power loss, mechanical failure, equipment failure, and safety issues [4]. The freezing of wind turbine blades changes the aerodynamic performance of the blades, which yield into power generation loss. Equally, the uneven distribution of ice from the blade changes the original mass distribution, making the wind turbine to run unstably, and causing severe damage to the blade in varying degrees, which not only lead to huge economic losses, but also have serious security risks [5]. Therefore, a reliable detection method for icing wind turbine blade is very important, especially in the early stage icing detection.

Now there are many standards and guidelines that have been developed by the IEC (International Electro technical Commission), and it has helped us a lot in analyzing turbine faults [6]. For the problem of ice detection in the blades, the existing methods mainly use the mechanism of icing to conduct theoretical analysis and research and establish the physical model of icing, then according to the monitoring data, make a judgment whether the wind turbine blades are frozen at the current moment. Davies et al. studied three methods of creating a power threshold curve to distinguish the ice growth period from the non-icing period to identify the power loss caused by icing [7]. Wang et al. proposed a numerical simulation method for three-dimensional wind turbine blade icing and compared it with experimental results to verify the effectiveness of the method [8]. Shu et al. studied the characteristics of leaf icing and the severity of icing on the power characteristics of wind turbines under natural icing conditions [9]. Blasco et al. performed a quantitative analysis of the power loss of a representative 1.5 MW wind turbine under various icing conditions, attempting to reduce the loss of wind farms in cold regions by formulating some control strategies [10]. Based on the analysis of supervisory control and data acquisition (SCADA) data, Li et al. proposed a method for detection of blade icing based on logistic regression [11]. Aral et al. proposes and demonstrates Phase-based Motion Estimation (PME) and a motion magnification algorithm to perform non-contact structural damage detection of a wind turbine blade [12]. Yu et al. developed a simple method to detect damage based on a discrete mathematical model for fan blades using changes in natural frequencies combined with a fluid-structure analysis [13]. The above research generally requires additional sensor placement for wind turbine blades. The disadvantages such as inconvenient practical application and increased wind farm operation and maintenance cost make them unable to be widely used in practice.

Vibration signal analysis [14] and SCADA system data analysis are two different aspects. The former pays more attention to the analysis of the equipment mechanism, while the latter tends to analyze the data. Both have their own advantages. Wind turbine blades work at high altitude, and they are inconvenient to measure the vibration acceleration signal offline. The amount of acceleration signal on the line is larger, which is inconvenient to transmit and store. Therefore, more and more people are committed to using SCADA data to predict and diagnose wind turbine faults. The SCADA system is the most widely used and technologically advanced data acquisition and monitoring system, in fault diagnosis of a large amount of wind power equipment [14–19]. This system collects environmental parameters and operating parameters of wind power equipment, which can fully characterize the operational status of the wind turbine. More and more people use it for data modeling and analysis to mine information of equipment fault and blade icing detection, etc. [20–24].

When the wind turbine blade is early frozen and detected by the model in this paper, this warning information will be fed back to the wind farm owners (or managers, or controllers). At this time, the wind farm has not experienced a serious accident. This early warning information of early icing gave them time to deal with the icing of the wind turbine blades. During this time, they could use other methods to reasonably arrange when to take measures such as deicing the blades, to prevent loss and damage due to severe icing of wind turbine blades.

This paper analyzes the SCADA data of a wind farm, combines the mechanism analysis and data analysis of the wind turbine icing to extract features that are sensitive to wind turbine icing, and then uses a random forest-based classification algorithm to achieve the detection of wind turbine blade icing. The first section introduces the related theories of icing of wind turbine blades and the research ideas of this paper; the second section introduces the related theories of the model based on the random forest classifier and the model assessment method selected for this paper; the third section is the data preprocessing which extracts the sensitive characteristics of early icing of wind turbine blades by analyzing the SCADA data; the fourth section is optimization and comparison of the model based on the random forest classifier with the results of other classifiers; and the last section is the conclusion.

2. Materials and Methods

2.1. Ice Detection on Wind Turbine Blades

2.1.1. Theory and Process of Icing

Icing is a physical phenomenon with a complete and specialized theory and research system. Blade icing of wind turbines is a type of atmospheric icing. The international standard ISO12494: 2017 [25] describes in detail the definition, scope, classification, principle, characteristics, and effects of such an icing. For wind turbines, atmospheric icing refers to the process of icing in the air frozen or adhered to objects exposed in the atmosphere under certain atmospheric conditions, including water droplets, rain, drizzle, snow, and other forms.

There are three forms of blade icing: cloud ice, sedimentation ice and accumulation of frost. Cloud ice refers to icing condensed from sub-cooled water droplets floating in clouds; sedimentation ice refers to icing caused by freezing rain or wet snow under low temperature conditions; frost accumulation refers to the direct phase change of water vapor. The icing process usually occurs at low temperatures. Among them, cloud ice and sedimentation ice are more common in wind turbine icing, and once it occurs, it will have a serious impact on the wind turbine and cause more damage.

2.1.2. Ice Detection Method

Ice detection analysis on blades is generally composed of several parts such as physical principle analysis, icing process analysis, feature extraction, detection model establishment and result presentation. This paper adopts blade detection model construction process based on a random forest classification, as shown in Figure 1.

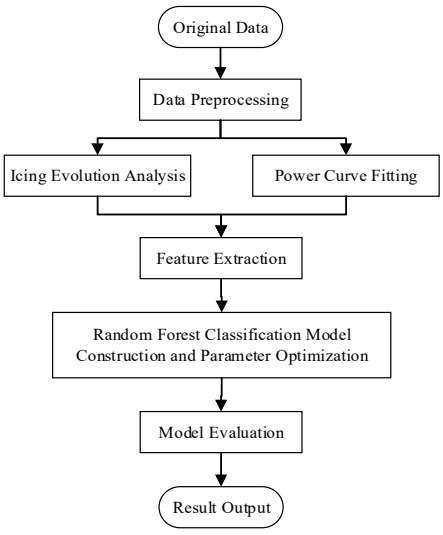

Figure 1. Flow chart of model construction.

Severe icing detection during the actual operation of the wind turbine is easily established, but automatically deicing by the wind turbine deicing system is also a challenge. However, the icing of the wind turbine blade is a slow process. In the early days of icing, the impact on the wind turbine is generally small and difficult to find. Besides, early icing will cause certain changes to the shape of the blades, which will cause water droplets in the atmosphere to stick to and freeze at the surface of the blades. Eventually, the probability of serious icing is greatly increased. The treatment of early icing is easier and has less impact on the wind turbine. It has a certain early warning effect of the occurrence of severe icing. Therefore, the detection of early icing is very important.

2.2. Model Based on Random Forest Classifier

2.2.1. Random Forest Classifier

The random forest [26] is a machine learning algorithm first published in 2001 by Breiman, L. which combines bagging ensemble learning theory [27] proposed in 1996, with the stochastic subspace method proposed by Ho, T. in 1998 [28]. This model adopts bootstrapping re-sampling technology to randomly select n samples from the original training sample set N and put it back randomly to generate a new training sample set to train a decision tree. Then, the above steps generate m decision trees to form a random forest. The classification results of new data-based upon the score formed by how many classification trees vote. Its essence is an improvement in the decision tree algorithm, with multiple decision trees merged together. The establishment of each tree depends on the independently extracted samples. Figure 2 shows the basic structure of the random forest classifier.

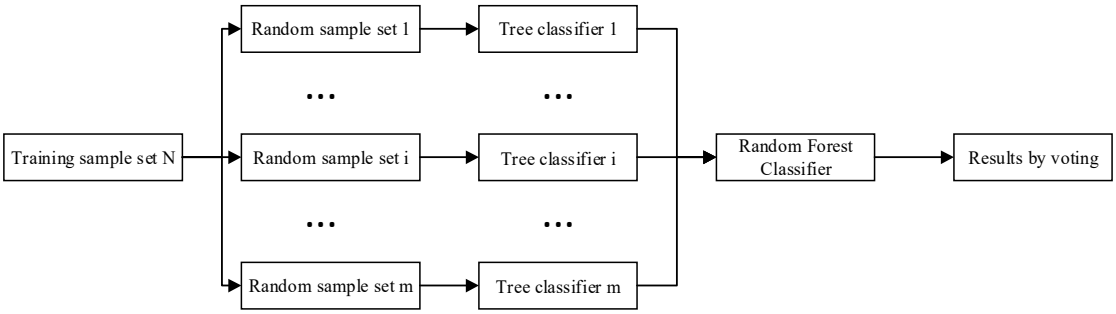

Figure 2. Basic structure of random forest classifier.

The classification ability of a single tree may be small, but after randomly generating many decision trees, a test sample through the statistics of the classification of each tree is selected to obtain the most likely classification.

1. The steps for the basic construction process of a random forest are: Use the bootstrapping method from the original training set to select n samples randomly in m times to generate m training sets.
2. For the newly generated m training sets, train m decision tree classification models.
3. For a single tree, every time a new node based upon the information gain or information gain ratio or the Gini is split to select the best split method.
4. Split each tree according to step 3 until the training sample is correctly classified at a certain node or reaches the maximum depth of the tree.
5. Organize the resulting multiple decision trees into the random forest classifier and the final classification results are determined by voting.

The randomness of each tree corresponding to the sampling of the training set and the way in which part of the features are selected when splitting to form a new node. The random forest does not need to be pruned and almost no over fitting occurs, and have good tolerance for noise and outliers, high stability, and strong generalization ability. In addition, the random forest is suitable for parallel

computing, and even for large samples and high latitude data, they have the higher training speed and the achieve efficient calculation.

This paper used a model based on the random forest classifier to identify early icing data from normal data to achieve the goal of predicting early icing failure, and then to determine if there would be icing failure in the next period.

2.2.2. Model Assessment Method

The confusion matrix [29] is a classical method for evaluating the results of classification models. Table 1 shows the confusion matrix representation.

Table 1. Confusion matrix representation.

Confusion Matrix	Prediction of Icing	Prediction of Non-Icing
Actual icing	*TP*	*FN*
Actual non-icing	*FP*	*TN*

where: *TP* indicates the proportion of all actual icing samples predicted to be icing samples; *TN* indicates the proportion of all actual non-icing samples predicted to be non-icing samples; *FP* indicates the proportion of all actual non-icing samples predicted to be icing samples; *FN* indicates the proportion of all actual icing samples predicted to be non-icing samples.

In addition, based upon the confusion matrix the precision of the test results and the recall rate assessed further to evaluate the model classification results [26].

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

3. Data Preprocessing

3.1. Data Sources and Introduction

In this paper, the test data are driven from the first China Industrial Big Data Innovation Competition [30], which contains two wind turbines SCADA data in a wind farm provided by Goldwind for predicting icing failures on blades. The SCADA data of each wind turbine contains 28 variables such as the time stamp, operating condition parameters, environmental parameters, and status parameters. The acquisition time was two months and with the sample size of about 580,000. Table 2 shows the statistical information of SCADA data. In addition, more detailed information of SCADA data can be seen in the Appendix A, Table A1.

Table 2. Statistical information of supervisory control and data acquisition (SCADA) data.

Data Set	Sample Size	Time Range
Wind turbine 15#	393,886	November & December 2015
Wind turbine 21#	190,494	November 2015

In addition, the organizers of the event conducted preliminary processing on the data, which removed severely frozen data and made the data not continuous; the data was also standardized, thus lost the physical meaning of the original data. Standardization means that making the mean of every variable in data is 0 and the variance is 1. The contest organizer has already set the labels for the data-icing and non-icing (due to the authority of the data owner and the supervisor, the accuracy of the data label is guaranteed, so the basis for judging whether the data is frozen or not is also credible); we only need to process the data that has been tagged.

3.2. Features Extraction

Some indicators in the raw data given by the contest organizers are sensitive to icing, and some indicators are almost not related to icing. So, the first step in this paper on the data is to pick out the icing-sensitive indicators from the raw data indicators. However, relying solely on these indicators does not well identify early icing data from non-icing data, this paper further processed the data and obtained some better indicators of icing and non-icing. In general, it is through the screening and supplementation of indicators to achieve better characterization of early icing with fewer features, which not only reduces the running time of the model but also gives better results.

This section will introduce the process of data preprocessing, including the screening of basic features and the construction of other features, with giving some figures to make features more intuitively judgment—whether it is easier to distinguish between icing and non-icing.

Extraction features, especially quantitative features [31] are very essential for the fault diagnosis of equipment. On the one hand, because the inertia of the wind turbine blade will reduce the correlation between the instantaneous power and the instantaneous wind speed, taking the average value from the data over a certain time span can reduce the inertial effect to some extent. On the other hand, in the original data, about 8 samples are collected every minute, but because the data provider has deleted some data, the sample interval time in the data is not fixed. So, the data are resampled in one-minute intervals, the specific process is as follows. According to the timestamp, the SCADA data grouped every minute for the time span, and then the mean of each group sample is taken as the new sample characteristics.

$$V = \frac{1}{n}\sum_{i=1}^{n} wind\_speed_i \tag{3}$$

where V is the average wind speed—the new sample characteristics; n is the number of wind_speed in one minute. The solutions of average power P and other new variables are the same as Equation (3).

Then the data is filtered.

1. Filter unspecified data. In the given raw data set, the data covers normal sample data, icing sample data, and other unspecified data, according to the status tag. Unspecified data will affect the classification of normal sample data/icing sample data, due to the uncertainty of its information; it will be classified as invalid data.
2. Filter samples below 80% of full power. Taking wind turbine 21# as an example, it can be found that when the wind turbine is at more than 80% full power icing status data does not exist by comparison of the original instantaneous power-wind speed scatter plot (shown in Figure 3), and the processed average power-wind speed scatter plot (shown in Figure 4). In other words, the wind turbine power cannot reach 80% of full power after the blade's freeze. Filter the samples below 80% of full power can make it easier to identify early icing data.

In following figures, the green points are in the wind turbine normal state and the red points are in the icing state. The blue lines in Figures 3 and 4 represent the dividing lines representing 80% of full power.

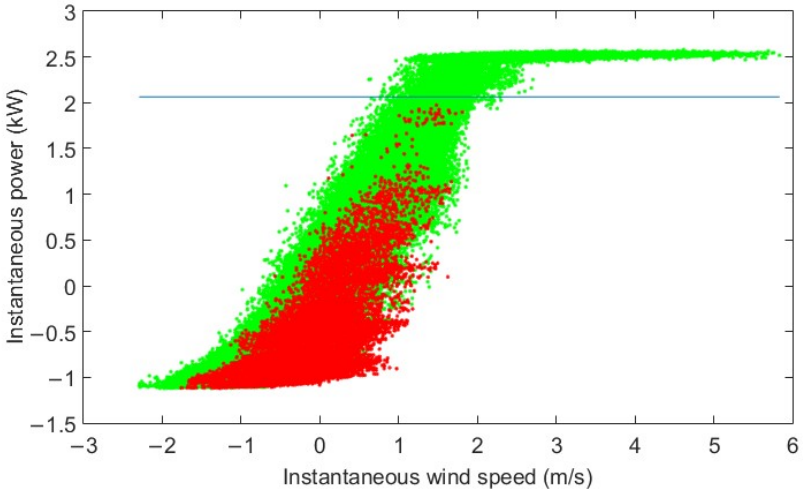

Figure 3. Original instantaneous power-wind speed scatter plot.

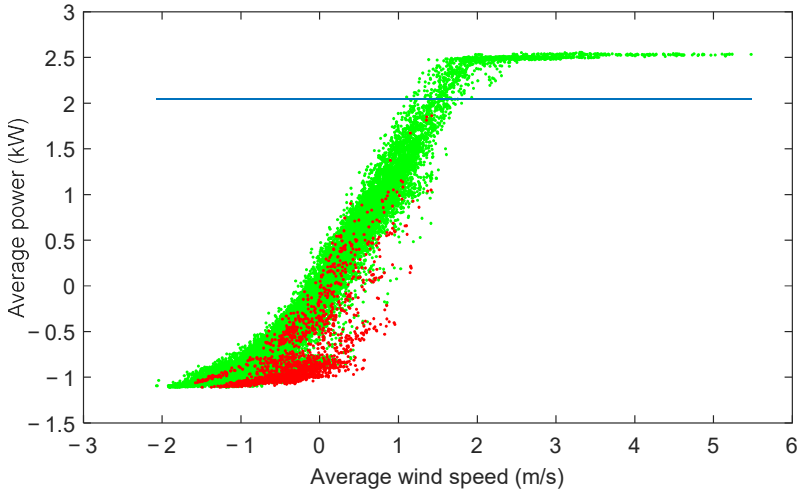

Figure 4. Average power-wind speed scatter plot.

Filter unspecified data and samples below 80% of full power and normalize the remaining data (making the scale from 0 to 1), then plot the average power and average wind speed as a scatter plot (shown in Figure 5).

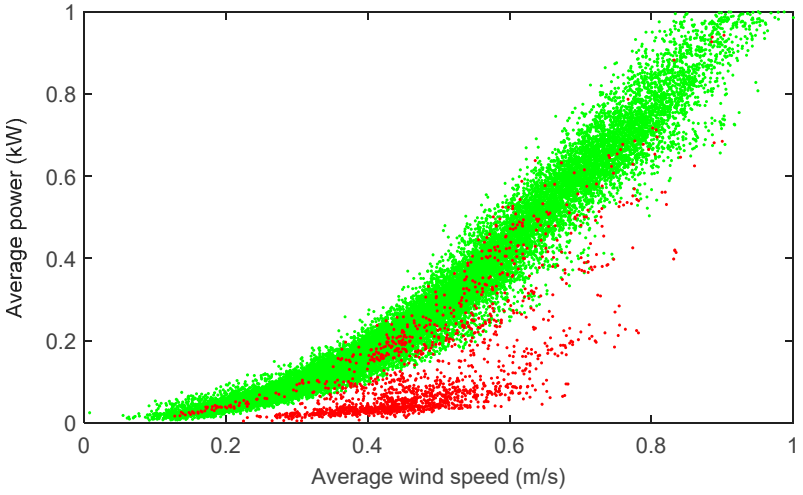

Figure 5. Average power-wind speed with removing more than 80% full power data scatter plot.

Wind turbines are devices that convert wind energy into mechanical energy and then into electrical energy, where wind speed and power are regarded as the two basic features of icing prediction. When the blades freeze, the shape and aerodynamic characteristics of the blades will change, reducing the power output. Therefore, when the wind turbine blades freeze, the relationship of the output power and the wind speed will be changed.

In the non-icing condition, the wind machine operates according to the wind turbine power characteristic curve in the normal mode (the green part of Figure 5). After the icing formation, the actual operation state of the wind turbine will deviate, and the power cannot reach the rated power. When the normal state sample data is used, the abnormal point eliminated, the power characteristic curve of the wind turbine is fitted to obtain a baseline model of the power characteristic curve [32], and then this model is used to predict the output power at the corresponding wind speed. The baseline model obtained by curve fitting is shown in Figure 6.

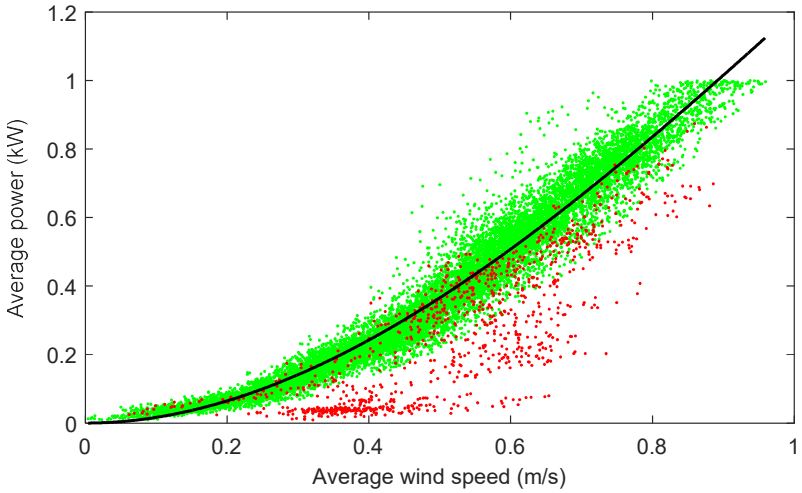

Figure 6. Fitted power curve.

From Figure 6 the icing sample is more deviating from the baseline model than the normal sample, thus constructing another feature of icing prediction, which can distinguish then better: the degree of deviation from the output power.

$$C = \frac{P_{pre} - P_{real}}{P_{\text{pre}}} = 1 - \frac{P_{real}}{P_{pre}} \tag{4}$$

where P_{real} is the actual measured output power and P_{pre} is the output power estimated by the actual wind speed and power curve.

After calculating the power degree by Equation (4), to facilitate visual observation of whether the variable is helpful for model classification, we draw a figure about relationship between the power degree and the average wind speed, as shown in Figure 7. As can be seen from Figure 7, there are more red dots (icing samples) that are distinguished from green dots (non-icing samples).

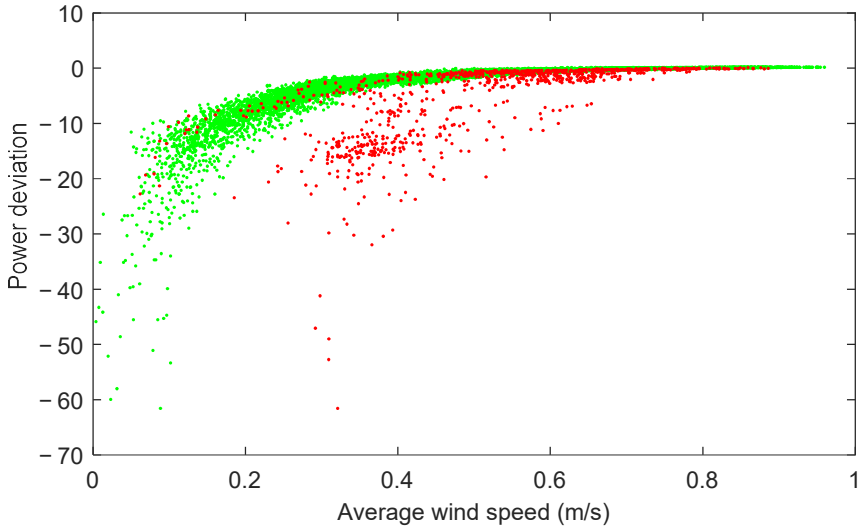

Figure 7. Relationship between degree of deviation and average wind speed.

In the early stage of icing, the operation state of the wind turbine is similar to the normal state, and it is difficult to separate the icing state from the normal state. However, the detection of early icing conditions is a very important process, and the healthy operation of the wind turbine unit is of utmost importance. It can minimize the loss to the unit due to icing on the blades. Because icing is a cumulative process, instantaneous characteristics such as the wind speed, the power, and the degree of deviation make it difficult to characterize fully icing conditions, especially in the early icing part. Therefore, it is necessary to analyze the evolution of the icing process and extract features that can characterize icing changes to better distinguish early icing conditions and achieve early icing prediction.

This paper mainly extracts features of early icing based upon the characteristics of degree of deviation. The icing process of the wind turbine contains certain periodicity, thus calling for serialization of the original data. Calculate the average rate of change (ΔC) of the degree of deviation at the corresponding time in each time segment.

$$\Delta C = \frac{C_t - C_{t_1}}{t - t_1} \tag{5}$$

where ΔC represents the average rate of change of the current degree of deviation; C_t represents the current degree of deviation; C_{t_1} represents the degree of deviation from the previous moment; $t - t_1$ represents the time span (t is the current time after digitization and t_1 is the previous time after digitization).

Then, according to the sliding window method, take ten minutes as the window length and one minute as the moving step length to obtain the maximum value max$C$ to the degree of deviation and the cumulative value sumΔC of the ΔC within 10 min before the current time.

First, this paper selects two basic features from the 28-dimensional features in the given data, and then adds four additional features based upon the mechanism of the wind turbine operation and icing. Finally, Table 3 represents the six groups of icing prediction features obtained.

Table 3. Features of ice detection on wind turbine blades.

Features	Description
V	Average wind speed
P	Average power
C	Degree of deviation of P
ΔC	Average rate of change of C
max$C$	Maximum value of C
sumΔC	Cumulative value of ΔC

4. Results

4.1. Classification Model Optimization

This paper mainly adjusts two important parameters of the model based on the random forest classifier to optimize the model: the number of trees and the maximum depth of the tree. Divide 70% of sample data of wind turbine 15# into the training set, 30% of sample data into the test set, adjust the number of decision tree and maximum depth of a tree in the random forest classifier, and then calculate the output from the model. The confusion matrix is chosen as the evaluation index, and the calculation results are shown in Table 4.

Table 4. Results of the model based on random forest classifier optimization.

Confusion Matrix/%		Number of Trees			
		10	20	30	40
Maximum depth of the tree	5	$\begin{pmatrix} 64.8 & 35.2 \\ 17.4 & 82.8 \end{pmatrix}$	$\begin{pmatrix} 66.7 & 33.3 \\ 15.6 & 84.4 \end{pmatrix}$	$\begin{pmatrix} 67.8 & 32.2 \\ 14.3 & 85.7 \end{pmatrix}$	$\begin{pmatrix} 68.5 & 31.5 \\ 13.1 & 86.9 \end{pmatrix}$
	10	$\begin{pmatrix} 77.6 & 22.4 \\ 12.3 & 87.7 \end{pmatrix}$	$\begin{pmatrix} 80.1 & 19.9 \\ 8.8 & 91.2 \end{pmatrix}$	$\begin{pmatrix} 84.2 & 15.8 \\ 7.9 & 92.1 \end{pmatrix}$	$\begin{pmatrix} 88.5 & 11.5 \\ 5.6 & 94.4 \end{pmatrix}$
	15	$\begin{pmatrix} 86.7 & 13.3 \\ 9.8 & 90.2 \end{pmatrix}$	$\begin{pmatrix} 89.8 & 10.2 \\ 4.4 & 95.6 \end{pmatrix}$	$\begin{pmatrix} 90.6 & 9.4 \\ 3.2 & 96.8 \end{pmatrix}$	$\begin{pmatrix} 91.0 & 9.0 \\ 3.5 & 96.5 \end{pmatrix}$
	20	$\begin{pmatrix} 90.6 & 9.4 \\ 5.6 & 94.4 \end{pmatrix}$	$\begin{pmatrix} 91.3 & 8.7 \\ 2.4 & 97.6 \end{pmatrix}$	$\begin{pmatrix} 91.8 & 8.2 \\ 2.5 & 97.5 \end{pmatrix}$	$\begin{pmatrix} 91.9 & 8.1 \\ 2.6 & 97.4 \end{pmatrix}$
	25	$\begin{pmatrix} 91.2 & 8.8 \\ 2.8 & 97.2 \end{pmatrix}$	$\begin{pmatrix} 92.2 & 7.8 \\ 2.5 & 97.5 \end{pmatrix}$	$\begin{pmatrix} 92.1 & 7.9 \\ 2.6 & 97.4 \end{pmatrix}$	$\begin{pmatrix} 92.0 & 8.0 \\ 2.5 & 97.5 \end{pmatrix}$
	30	$\begin{pmatrix} 91.1 & 8.9 \\ 2.8 & 97.2 \end{pmatrix}$	$\begin{pmatrix} 92.1 & 7.9 \\ 2.6 & 97.4 \end{pmatrix}$	$\begin{pmatrix} 92.1 & 7.9 \\ 2.6 & 97.4 \end{pmatrix}$	$\begin{pmatrix} 92.1 & 7.9 \\ 2.7 & 97.3 \end{pmatrix}$

From Table 4, random forest model parameters selections are: the number of trees 20, and the maximum depth 25.

4.2. Test Results

After the optimization of the model based on the random forest classifier in the previous section, the parameters of the model based on the random forest classifier are determined.

The next four groups of tests are about the different classification results of the model based on the random forest classifier between data of wind turbines 15# and 21#. The train and test set details and the classification results from each test are as follows. In addition, in the following tables, the indicator, the running time, refers to the total time the model takes to training and predicts the data on the experimental computer.

In the Test No. 1, 70% of the sample data of the wind turbine 15# was divided into training sets, and 30% of the sample data was divided into a test set. The results are shown in Table 5.

Table 5. Result of Test No. 1 (Running time: 27.0 s).

Confusion Matrix	Prediction of Icing	Prediction of Non-Icing
Actual icing	92.2%	7.8%
Actual non-icing	2.5%	97.5%

In the Test No. 2, 70% of the sample data of the wind turbine 21# divided into training sets, and 30% of the sample data is the test set, as shown in Table 6.

Table 6. Result of Test No. 2 (Running time: 14.2 s).

Confusion Matrix	Prediction of Icing	Prediction of Non-Icing
Actual icing	85.6%	14.4%
Actual non-icing	5.0%	95.0%

Test No. 3 takes all the sample data of wind turbine 21# into the training set and the sample data of 15# wind turbine into test set, as shown in Table 7.

Table 7. Result of Test No. 3 (Running time: 38.1 s).

Confusion Matrix	Prediction of Icing	Prediction of Non-Icing
Actual icing	82.2%	17.8%
Actual non-icing	18.2%	81.8%

Test No. 4 takes all the sample data of wind turbine 15# into the training set and the sample data of wind turbine 21# into the test set, as shown in Table 8.

Table 8. Result of Test No. 4 (Running time: 26.8 s).

Confusion Matrix	Prediction of Icing	Prediction of Non-Icing
Actual icing	89.8%	10.2%
Actual non-icing	8.8%	91.2%

Consequently, the classification results between different classification models are also compared. This paper selects the logistic regression classifier, the GBDT (Gradient Boosting Decision Tree) classifier and the random forest classifier for comparison. In these tests, 70% of the sample data of the wind turbine 15# is the training set, and 30% of the sample data is set as the test set.

Test No.5 used a logistic regression classification model. After the optimization, the classification threshold was set to 0.86, as results shown in Table 9 demonstrate.

Table 9. Result of Test No. 5 (Running time: 86.7 s).

Confusion Matrix	Prediction of Icing	Prediction of Non-Icing
Actual icing	88.8%	11.2%
Actual non-icing	8.8%	91.2%

The GBDT classification model was used in the Test No. 6. In addition, after the model optimization, the classification result is presented in Table 10.

Table 10. Result of Test No. 6 (Running time: 56.8 s).

Confusion Matrix	Prediction of Icing	Prediction of Non-Icing
Actual icing	76.6%	23.4%
Actual non-icing	5.7%	94.3%

The precision and recall of the six tests computed separately, and the obtained results are shown in Table 11.

Table 11. Precision and recall tests.

Test Number	Train Data	Test Data	Model	Precision	Recall
No. 1	70% of 15#	30% of 15#	RF	97.4%	92.2%
No. 2	70% of 21#	30% of 21#	RF	94.5%	85.6%
No. 3	all of 21#	all of 15#	RF	81.9%	82.2%
No. 4	all of 15#	all of 21#	RF	91.1%	89.8%
No. 5	70% of 15#	30% of 15#	LR	97.4%	88.8%
No. 6	70% of 15#	30% of 15#	GBDT	93.1%	76.6%

Where RF means random forest classifier, and LR means logistic regression classifier.

As we all known, for a classification model, when both precision and recall have higher values at the same time without considering other factors, the model is thought to have a better performance. Through the comparison of test results we draw, the following summary is concluded:

1. Considering precision and recall at the same time, the results of the model based on the random forest classifier are better than other models (as the result of Table 11 shown), so that the model based on the random forest classifier has high accuracy in ice detection of wind turbine blades and can identify the early failure.
2. The trained model based on the random forest classifier still performs well on a new test set (as the results of Tests No. 3 and No. 4 shown), so that the model has good generalization ability, thus it has a strong adaptability to the new sample data.
3. The data of wind turbine 15# is more than the data of 21#. From the results of these tests (as the results of Tests No. 1, No. 2, No. 3, and No. 4), when only focusing on precision and recall without considering the running time, increasing the sample size of the training set can improve the classification performance of the model.
4. Considering the running time of these models, the model based on the random forest classifier shorter running time than the logistic regression and the GBDT classification models (as the results of Tests No. 5 and No. 6), so that the model based on the random forest classifier has more efficient calculation ability.

5. Conclusions

To detect wind turbine blade icing, a model based on the random forest classifier was proposed. The model with high accuracy and good generalization ability was verified by the data of the China Industrial Big Data Innovation Competition.

1. The features extracted in this paper well reflect the characteristics of icing failure. Two basic variables of wind speed and output power from SCADA data and more instantaneous and statistical features extracted from a power curve can be used to characterize the early icing on wind turbine blades. The various models selected in this paper have good results on this data set. So, it has high practical application value for ice detection in wind farms, which, as reflected in the effective prevention of severe ice formation in the blades, increase of wind farm's profit, and reduction of safety accident.
2. The model based on the random forest classifier has very high accuracy and good generalization ability for ice detection on a wind turbine's blades, which is used to identify the early icing. Compared with other classification models, the model based on the random forest classifier has higher accuracy and more efficiency in terms of computing capabilities, making it more suitable for the practical application of ice detection.

3. In the identification of new data, the accuracy of the model has reached more than 80%, which shows that there is room for further improvement. In the future, it can expand better features or use other models to ensure the best results in both accuracy and generalization.

Author Contributions: Conceptualization, L.Z.; Data curation, K.L.; Methodology, K.L. and Y.W.; Supervision, L.Z.; Writing–original draft, K.L.; Writing–review and editing, L.Z. and Z.B.O.

Funding: This research was funded by the National Key Research and Development Program of China (No. 2016YFF0203800), the Fundamental Research Funds for Central Universities of China (No. FRF-BD-18-001A) and the National Natural Science Foundation of China (No. 51775037).

Acknowledgments: The authors would like to thank the anonymous reviewers for their valuable comments and suggestions that helped improve the quality of this manuscript.

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

Appendix A

Table A1. Variable name and description.

Number	Name	Description
1	time	timestamp
2	wind_speed	wind speed
3	generator_speed	generator speed
4	power	grid side active power
5	wind_direction	wind direction
6	wind_direction_mean	mean of wind direction
7	yaw_position	yaw position
8	yaw_speed	yaw speed
9	pitch1_angle	pitch 1 angle
10	pitch2_angle	pitch 2 angle
11	pitch3_angle	pitch 3 angle
12	pitch1_speed	pitch 1 speed
13	pitch2_speed	pitch 2 speed
14	pitch3_speed	pitch 3 speed
15	pitch1_moto_tmp	pitch motor 1 temperature
16	Pitch2_moto_tmp	pitch motor 2 temperature
17	Pitch3_moto_tmp	pitch motor 3 temperature
18	acc_x	X-direction acceleration
19	acc_y	Y-direction acceleration
20	environment_tmp	environment temperature
21	int_tmp	cabin temperature
22	pitch1_ng5_tmp	ng5 1 temperature
23	pitch2_ng5_tmp	ng5 2 temperature
24	pitch3_ng5_tmp	ng5 3 temperature
25	pitch1_ng5_DC	ng5 1 charger DC current
26	pitch2_ng5_DC	ng5 2 charger DC current
27	pitch3_ng5_DC	ng5 3 charger DC current
28	group	data group identification

References

1. Leite, G.D.N.P.; Araújo, A.M.; Rosas, P.A.C. Prognostic techniques applied to maintenance of wind turbines: a concise and specific review. *Renew. Sustain. Energy Rev.* **2018**, *81*, 1917–1925. [CrossRef]
2. Global Wind Energy Council (GWEC). Available online: http://gwec.net/global-figures/graphs/ (accessed on 20 June 2018).
3. Oh, K.Y.; Nam, W.; Ryu, M.S.; Kim, J.Y.; Epureanu, B.I. A review of foundations of offshore wind energy convertors: Current status and future perspectives. *Renew. Sustain. Energy Rev.* **2018**, *88*, 16–36. [CrossRef]
4. Gantasala, S.; Luneno, J.C.; Aidanpää, J.O. Influence of icing on the modal behavior of wind turbine blades. *Energies* **2016**, *9*, 862. [CrossRef]

5. Shu, L.; Liang, J.; Hu, Q.; Jiang, X.; Ren, X.; Qiu, G. Study on small wind turbine icing and its performance. *Cold Reg. Sci Technol.* **2017**, *134*, 11–19. [CrossRef]
6. British Standards Institution. Overhead Transmission Lines-Design Criteria. IEC 60826: 2017. Available online: https://webstore.ansi.org/RecordDetail.aspx?sku=BS+IEC+60826%3A2017 (accessed on 18 September 2018).
7. Davis, N.N.; Byrkjedal, Ø.; Hahmann, A.N.; Clausen, N.; Mark, Ž. Ice detection on wind turbines using the observed power curve. *Wind Energy* **2016**, *19*, 999–1010. [CrossRef]
8. Wang, Z.; Zhu, C. Numerical simulation for in-cloud icing of three-dimensional wind turbine blades. *Simulation* **2017**, *94*, 31–41. [CrossRef]
9. Shu, L.; Li, H.; Hu, Q.; Jiang, X.; Qiu, G.; McClure, G.; Yang, H. Study of ice accretion feature and power characteristics of wind turbines at natural icing environment. *Cold Reg. Sci. Technol.* **2018**, *147*, 45–54. [CrossRef]
10. Blasco, P.; Palacios, J.; Schmitz, S. Effect of icing roughness on wind turbine power production. *Wind Energy* **2017**, *20*, 601–617. [CrossRef]
11. Li, N.; Yan, T.; Li, N.; Kong, D.; Liu, Q.; Lei, Y. Ice detection method by using SCADA data on wind turbine blades. *Power Gener. Technol.* **2018**, *39*, 58–62.
12. Aral, S.; Zhu, M.; Christopher, N.; Peyman, P. Vibration-based damage detection in wind turbine blades using Phase-based Motion Estimation and motion magnification. *J. Sound Vib.* **2018**, *421*, 300–318.
13. Yu, M.Y.; Fu, S.; Gao, Y.B.; Zheng, H.; Xu, Y.G. Crack detection of fan blade based on natural frequencies. *Int. J. Rotating Mach.* **2018**, *2018*, 1–13. [CrossRef]
14. Qin, Y.; Zou, J.Q.; Cao, F.L. Adaptively detecting the transient feature of faulty wind turbine planetary gearboxes by the improved kurtosis and iterative thresholding algorithm. *IEEE Access.* **2018**, *6*, 14602–14612. [CrossRef]
15. Tautz-Weinert, J.; Watson, S.J. Using SCADA data for wind turbine condition monitoring—A review. *IET Renew. Power Gener.* **2017**, *11*, 382–394. [CrossRef]
16. Dai, J.; Yang, W.; Cao, J.; Liu, D.; Long, X. Ageing assessment of a wind turbine over time by interpreting wind farm SCADA data. *Renew. Energy* **2017**, *116*, 199–208. [CrossRef]
17. Chen, N.; Yu, R.; Chen, Y.; Xie, H. Hierarchical method for wind turbine prognosis using SCADA data. *IET Renew. Power Gener.* **2017**, *11*, 403–410. [CrossRef]
18. Dao, P.B.; Staszewski, W.J.; Barszcz, T.; Uhl, T. Condition monitoring and fault detection in wind turbines based on cointegration analysis of SCADA data. *Renew. Energy* **2017**, *116*, 107–122. [CrossRef]
19. Dai, J.; Yang, X.; Hu, W.; Wen, L.; Tan, Y. Effect investigation of yaw on wind turbine performance based on SCADA data. *Energy* **2018**, *149*, 684–696. [CrossRef]
20. Bangalore, P.; Patriksson, M. Analysis of SCADA data for early fault detection, with application to the maintenance management of wind turbines. *Renew. Energy* **2017**, *115*, 521–532. [CrossRef]
21. Alvarez, E.J.; Ribaric, A.P. An improved-accuracy method for fatigue load analysis of wind turbine gearbox based on SCADA. *Renew. Energy* **2018**, *115*, 391–399. [CrossRef]
22. El-Asha, S.; Zhan, L.; Lungo, G.V. Quantification of power losses due to wind turbine wake interactions through SCADA, meteorological and wind LiDAR data. *Wind Energy* **2017**, *20*, 1823–1839. [CrossRef]
23. Cao, M.; Qiu, Y.; Feng, Y.; Wang, H.; Li, D. Study of wind turbine fault diagnosis based on unscented Kalman filter and SCADA data. *Energies* **2016**, *9*, 847. [CrossRef]
24. Sun, P.; Li, J.; Wang, C.; Lei, X. A generalized model for wind turbine anomaly identification based on SCADA data. *Appl. Energy* **2016**, *168*, 550–567. [CrossRef]
25. Atmospheric Icing of Structures. Available online: https://www.iso.org/standard/72443.html (accessed on 6 June 2018).
26. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
27. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
28. Ho, T. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 832–844.
29. Deng, X.; Liu, Q.; Deng, Y.; Mahadevan, S. An improved method to construct basic probability assignment based on the confusion matrix for classification problem. *Inf. Sci.* **2016**, *340*, 250–261. [CrossRef]
30. Industrial Big Data Innovation Competition. Available online: http://www.industrial-bigdata.com (accessed on 6 June 2018).

31. Cui, L.; Wu, N.; Ma, C.; Wang, H. Quantitative fault analysis of roller bearings based on a novel matching pursuit method with a new step-impulse dictionary. *Mech. Syst. Signal Process.* **2016**, *68–69*, 34–43. [CrossRef]
32. Yang, X.; Xue, W.; Bao, L. Application of IEC standards in the measurement of practical power curve of wind turbine generator. *Power Syst. Clean Energy* **2012**, *28*, 87–91.