ISSN 2616-5775 Vol. 5, Issue 2: 81-86, DOI: 10.25236/AJCIS.2022.050213

Wind Turbine Blade Icing Detection Based on Random Forest

Yanlong Zhao[1,*], Liwen Wang[2]

[1]*School of Electrical Engineering, University of Jinan, Jinan, China*
[2]*Faculty of Electric Power Engineering, Kunming University of Science and Technology, Kunming, China*
**Corresponding author*

Abstract: *In this paper, a multimodal random forest method is proposed to solve the problem that the traditional linear method has low accuracy in predicting the icing of generator blade. This paper uses the random forest algorithm to model and analyze the icing event of wind turbine blades, and describes in detail the process of using the C4.5 algorithm to generate a decision tree, and then randomly selecting samples and sample features to train to generate a random forest. The effectiveness of the method in this paper is experimentally verified by SCADA data, and the results show that the multi-modal random forest algorithm reduces the error rate to 1.97% in wind turbine blade icing prediction, which is more accurate than the traditional method.*

Keywords: *Ensemble Learning; Random Forest; Icing Detection; SCADA Data*

1. Introduction

Wind power generation accounts for a large proportion of global new energy. It is still on the rise [1], and wind turbines are installed in large numbers in areas with abundant wind resources, and it is essential to obtain stable wind power generation. Many unfavorable factors can lead to unstable wind power generation and waste of resources. The more profound impact is the icing of wind turbine blades. When the blades of wind turbines freeze, the load on the blades becomes more extensive, which causes a large part of the wind energy to be wasted on the rotation of the dragging blades. In addition, the icing of the blades causes a burden for subsequent maintenance. The current method to solve this problem is to manually observe whether the blade is icing or an external icing condition monitoring-based system [2]. These methods are usually that the blade icing has occurred and cannot effectively solve the waste of resources and staffing in a timely and effective manner. Therefore, finding a timely and accurate prediction method for wind turbine blade icing is significant.

In recent years, the detection of icing on wind turbine blades based on mechanism models has emerged. For example, a technology based on passive thermal infrared is used to detect ice on wind turbine blades [3]. The temperature of the blades is determined by collecting thermal infrared images of the wind turbine blades. And the vibration frequency to predict whether the fan blade is icing, and the piezoelectric ceramic patch is installed on the blade of the fan to record different wavelet packet energy (WPE). Then the icing of the blade is analyzed, and the temperature of the blade is recorded by a fiber-optic sensor Change to identify and freeze time [5]. However, most of the analysis mentioned above of the event mechanism of blade icing has some inevitable shortcomings: high cost, complicated interpretation of test results, reference standards are required, test operators need to be trained, etc. Therefore, it is essential to establish a concise and reliable non-mechanical model to explain blade icing.

With the development of artificial intelligence, data-driven technology is widely used [4], [8], [16]. There are many algorithms to solve the classification problem. Tian et al. applied K-Nearest Neighbor (KNN) algorithm to detect motor-bearing faults by extracting fault features [6]. Bodla et al. used the logistic regression method for early fault detection on the condition of wind turbines [7]. The classifiers established by these algorithms are all linear classifiers with simple structure and low complexity, but they cannot mine the nonlinear relationship between variables.

In order to dig deeper into the nonlinear relationship between variables, we should build a binary classifier that can learn the nonlinear relationship. There are many algorithms to choose from. For example, support vector machines (SVM) can dig the nonlinearity between variables. Laouti et al. use

ISSN 2616-5775 Vol. 5, Issue 2: 81-86, DOI: 10.25236/AJCIS.2022.050213

periodic sampling to detect wind turbine failures based on SVM [9]. Another classifier decision tree (DT) based on a tree structure can also mine the nonlinear relationship between variables. Abdallah et al. implemented DT based on Apache Hadoop and Spark to detect damage and failure of wind turbines [10]. Although these algorithms can mine complex correlations between variables, a single classifier often makes unavoidable errors when solving multimodal relationships. Therefore, it is necessary to find a way to comprehensively solve this multimodal classification problem.

However, a multi-modal learning classification model based on ensemble learning is proposed based on the DT to solve the classification problem in one mode. There are two major categories of boosting and bagging. Boosting is an algorithm that promotes a weak learner to a strong learner. When each sub-learner is generated, it focuses on learning the previous error samples, thereby reducing the skewness of the model. Bagging uses self-service sampling to produce multiple sub-learners in parallel, and the results are decided by voting. It mainly reduces the variance of the model. Que et al. used extreme gradient boosting (XGBoost) to realize fault detection and remaining life prediction of steam turbines [11]. However, the boosting category model is relatively high in complexity and sensitive to abnormal data. The bagging model has the same level of complexity as a single classifier and can automatically generate training and test sets. The representative random forest (RF) of bagging generates multiple training samples by randomly sampling feature variables and samples to learn nonlinear relationship mining between variables in multiple modes. Through the above discussion, in order to solve the shortcomings of traditional methods in blade icing detection, feature variables related to wind turbine blade icing will be extracted, and sample data will be collected for self-sampling to generate multiple sub-classifiers in parallel to construct RF. The difference between each sub-learner is used to improve the generalization ability of the model to realize the detection of blade icing.

2. Concept and Approach

Since random forests are constructed using decision trees that have not been pruned as sub-learners, an indispensable part of studying random forests is to find a suitable method to generate decision trees. In this paper, multiple training samples are generated by the self-sampling method and random selection of attributes. The C4.5 algorithm based on the information gain rate to select the root node and the intermediate node is introduced to generate a decision tree a random forest with classification function is integrated.

As a classic machine learning algorithm, the decision tree model can be regarded as a recursive process when constructed [12]. First, an attribute is selected in the sample set D to be split. At this time, the node generated by the attribute is called the root node. After such an attribute test, the set D is divided into several subsets, and then each subset can be split by an attribute test. The node generated at this time is called an internal node. If an attribute test is performed, the samples of the node are all the same category, or all samples have the same category on all attributes. Then mark the node as a leaf node, where the category with the most contained samples is taken as the node's category. If there is no sample set on the node, mark the node as a leaf node, its category is that the category of the node is the same as the category of its parent node. Through such multiple splits and the labeling of leaf nodes, a classifier with a tree structure is produced, and the generalization ability of the decision tree largely depends on the optimal choice of attributes at the time of splitting and a division principle based on information gain rate is introduced below.

Information entropy is used as a measure of sample purity [13]. Suppose there are k classes in sample D, and the proportion of each class in the total sample is p_i(i=1,2, 3....k). Then the definition of information entropy is expressed as:

$$H(D) = -\sum_{i=1}^{k} p_k \log_2 p_k \quad (1)$$

If the value of H(D) is larger, the purity of the sample is greater. When we select attribute A in the sample as the root node to split, D can be divided into several subsets$\{D_1, D_2, D_3, \ldots D_n\}$. Split the information entropy of the Sample D as:

$$H(D)_A = -\sum_{j=1}^{n} \frac{|D_j|}{|D|} H(D_j) \quad (2)$$

Therefore, the information gain of taking attribute A as the root node is defined as:

$$G(D, A) = H(D) - H(D)_A \quad (3)$$

In the early decision tree generation, the classic ID3 algorithm uses attributes with large information

ISSN 2616-5775 Vol. 5, Issue 2: 81-86, DOI: 10.25236/AJCIS.2022.050213

gain as the root node, and this often results in a tilt towards attributes with many values, so the C4.5 algorithm that uses information gain as an index can avoid this preference [14], the information gain rate is defined as:

$$GR(D,A) = \frac{G(D,A)}{P(A)} \quad (4)$$

where $P(A)$ can be expressed as:

$$P(A) = -\sum_j^n \frac{|D_j|}{|D|} \log_2 \frac{|D_j|}{|D|} \quad (5)$$

Gradually select the attribute with a large information gain rate as the node split and finally reach the leaf node through the recursive method.

In the random forest, two random ideas, including bagging and random selection of feature variables, are used to generate multiple decision trees. The final result is output by voting when solving the classification problem. Suppose that multiple decision trees $\{T_1, T_2, T_3, \dots, T_N\}$ are generated by the above method, and each sub-learner outputs a mark in the category set $\{C_1, C_2, C_3, \dots, C_n\}$ We input sample X into each sub-learner T_i and the output obtained is $\{T(X)_i^1, T(X)_i^2, T(X)_i^3, \dots, T(X)_i^n\}$, where T_i^j represents the output of T_i on category C_j, and the integrated learner uses the voting method to output the final classification result. The principles of voting law are:

$$R(X) = \arg\max_j \sum_{i=1}^N T_i^j \quad (6)$$

We describe the construction process of the random forest in Fig. 1.

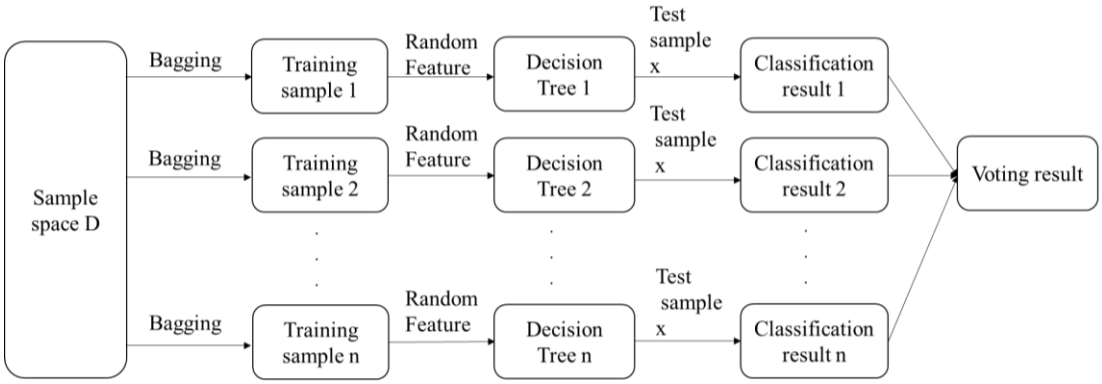

Figure 1: Random Forest construction process

Since about 37% of the samples are not selected during bagging [15], these unselected samples can be used as the test set to evaluate the performance of the model after the construction of the random forest.

3. Experiment and Discussion

Supervisory Control and Data Acquisition (SCADA) system can detect and collect industrial field data in real-time to realize automatic remote control of the local industrial field and comprehensively monitor the execution of the production process to provide necessary data support for production and management. In this paper, the operating status data of each wind turbine in the wind farm monitored by the SCADA system is used as support to verify the advantages of the random forest model for wind turbine blade icing detection compared to other single-mode linear models.

Based on the data collected by SCADA, each piece of data is time-stamped data. That is, each piece of data contains the detection of multiple continuous variables of the wind turbine at the moment, including 26 variable parameters, as shown in Table 1. We used 300 pieces of data to train the model, in which the sample data was marked as normal operation or icing, and multiple blade icing failures appeared in the data set, and the normal operation data and the data when the blades were icing were randomly mixed together. After the model training is completed, the past 1263 pieces of training data are used to evaluate and analyze the model training results.

ISSN 2616-5775 Vol. 5, Issue 2: 81-86, DOI: 10.25236/AJCIS.2022.050213

Table 1: Process variables

Variable name	Variable name	Variable name
Wind_speed	Pitch3_angle	Environment_tmp
Generator_speed	Pitch1_speed	Int_tmp
Power	Pitch2_speed	Pitch1_ng5_tmp
Wind_direction	Pitch3_speed	Pitch2_ng5_tmp
Wind_direction_mean	Pitch1_moto_tmp	Pitch3_ng5_tmp
Yaw_position	Pitch2_moto_tmp	Pitch1_ng5_DC
Yaw_speed	Pitch3_moto_tmp	Pitch2_ng5_DC
Pitch1_angle	Acc_x	Pitch3_ng5_DC
Pitch2_angle	Acc_y	

Among them, Logistic Regression (LR) model, Naive Bayes (NB), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Tree (DT) are selected to compare with the Random Forest (RF), and make the error rate, precision rate, recall rate, F1 score of each model in the test set to evaluate the effect of each model in detecting blade icing, the test results of each model in each index are given in Table 2. The error rate of each model is visualized as shown in Figure 1, where the lower the error rate, the better the model classification effect. The precision rate indicates how much data of the units that the model considers to be in normal operation is really normal data, while the recall rate indicates how much data of the units that the model classifies as normal data. Finally, the F1 score combines the accuracy rate and the recall rate. To evaluate the model, the larger the three values, the better the classification effect of the model. The visualization results are shown in Figure 2. At the same time, the receiver operating characteristic (ROC) curve and area under curve (AUC) value of each model are plotted to observe the advantages of the multi-modal integration algorithm, as shown in Figure 3.

Table 2: Results of different models

Model	Error rate	Precision rate	Recall rate	F1-score
RF	1.97%	96.24%	100%	98.08%
SVM	4.82%	91.41%	99.84%	95.44%
LR	4.35%	92.45%	99.53%	95.86%
NB	12.41%	80.30%	100%	89.07%
KNN	4.35%	92.57%	99.37%	95.85%
DT	2.61%	96.33%	98.59%	97.45%

It can be seen from Table 2 and Figure 1 that the error rates of linear classifiers such as KNN, LR, and Naive Bayes are as high as 4.35%, 4.35%, and 12.41% incorrectly distinguishing the working state of wind turbines, while the non-linear classifiers, the error rate of DT and SVM is 2.61% and 4.82%. Compared with linear classifiers, the error rate has decreased, but due to the limitation of single mode, the effect is still not ideal. However, the error rate of the random forest model of the integrated learning algorithm reached 1.97%, its classification effect shows superiority compared with other models.

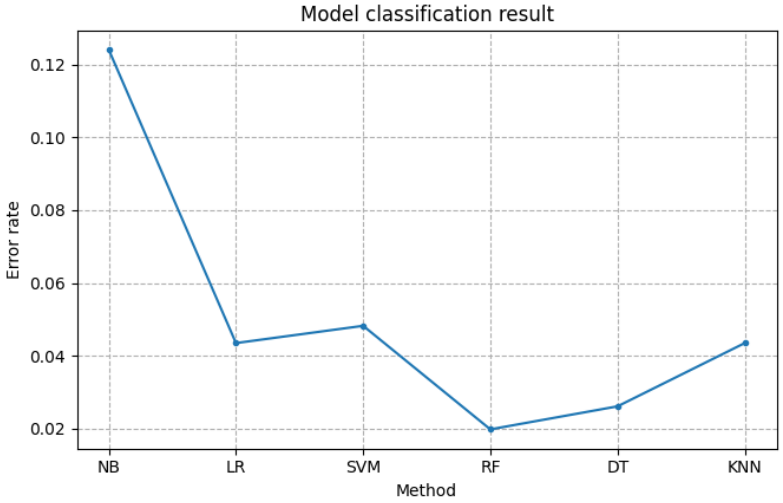

Figure 2: Different model error rate

ISSN 2616-5775 Vol. 5, Issue 2: 81-86, DOI: 10.25236/AJCIS.2022.050213

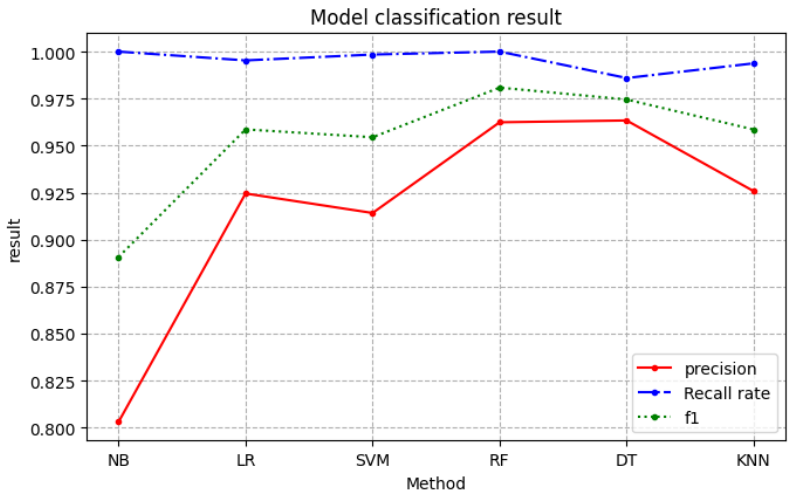

Figure 3: Different models score under each indicator

Combining Table 2 and Figure 3, it can be seen that the precision, recall, and F1 scores of the random forest model have reached 96.24%, 100%, and 98.08%, respectively. The scores of the three evaluation indicators are very advantageous compared to other classification models. On the whole, the random forest model has the best effect on the icing detection of wind turbine blades.

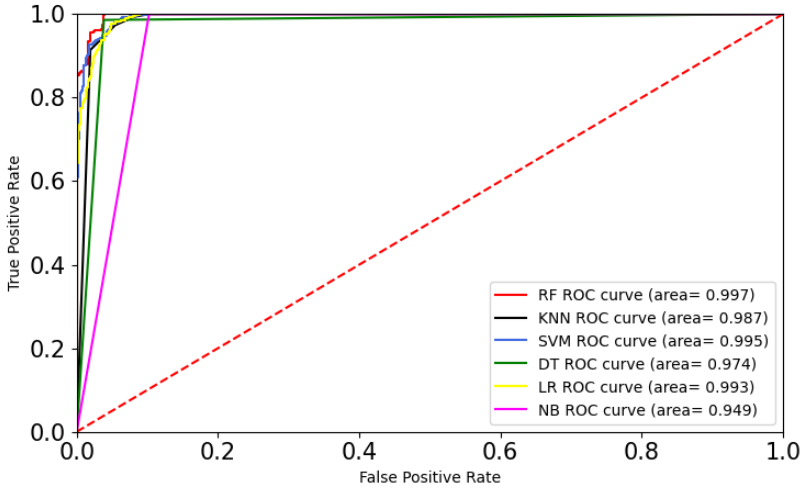

Figure 4: ROC curve and AUC value of different models

The closer each model in the ROC curve to the upper left corner of the coordinate plane and the more it deviates from the 45-degree diagonal, the better the classification effect of the model, and the AUC value represents the area under the ROC curve, and the larger the area, the better the effect of the model. Observing the ROC curves and AUC values of various models in Fig. 4, it can be clearly seen that the random forest model under ensemble learning has advantages in processing wind turbine blade icing detection compared with other models.

4. Conclusion

Aiming at the problem of blade icing faults of wind turbines in high altitude areas, a multimodal random forest algorithm is proposed to solve the detection and prediction of blade icing faults. The random forest model is obtained by using SCADA data to train and predict it on the test set and compared with support vector machine, logistic regression model, naive Bayes, KNN, decision tree under the same data set. The experimental results show that compared with the traditional methods, the random forest algorithm is more effective in detecting leaf icing. The error rate is reduced to 1.97%, which is significantly lower than other models (the error rates are 4.82%, 4.35%, 12.41%, 4.35%, and 2.61%, respectively), which produces better prediction accuracy and has practical application value.

ISSN 2616-5775 Vol. 5, Issue 2: 81-86, DOI: 10.25236/AJCIS.2022.050213

References

[1] Kreutz, Markus, et al. "Machine learning-based icing prediction on wind turbines." Procedia CIRP 81 (2019): 423-428.
[2] Liu, Yao, et al. "Intelligent wind turbine blade icing detection using supervisory control and data acquisition data and ensemble deep learning." Energy Science & Engineering 7.6 (2019): 2633-2645.
[3] Ghani, Rizwan, and S. Virk Muhammad. "Experimental study of atmospheric ice detection on wind turbine blade using thermal infrared technique." Wind Engineering 37.1 (2013): 71-77.
[4] Chen, Xu, and Chunhui Zhao. "Condition-Driven Soft Transition Modeling and Monitoring Strategy for Complex Nonstationary Process." IFAC-PapersOnLine 54.3 (2021): 445-450.
[5] Zhang, Zhao hui, Wen song Zhou, and Hui Li. "Icing estimation on wind turbine blade by the interface temperature using distributed fiber optic sensors." Structural Control and Health Monitoring 27.6 (2020): e2534.
[6] Tian, Jing, et al. "Motor bearing fault detection using spectral kurtosis-based feature extraction coupled with K-nearest neighbor distance analysis." IEEE Transactions on Industrial Electronics 63.3 (2015): 1793-1803.
[7] Bodla, Muhammad Kamran, et al. "Logistic regression and feature extraction based fault diagnosis of main bearing of wind turbines." 2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA). IEEE, 2016.
[8] Chen, Xu, and Chunhui Zhao. "Linear and nonlinear hierarchical multivariate time delay analytics for dynamic modeling and process monitoring." Journal of Process Control 107 (2021): 83-93.
[9] Laouti, Nassim, Nida Sheibat-Othman, and Sami Othman. "Support vector machines for fault detection in wind turbines." IFAC Proceedings Volumes 44.1 (2011): 7067-7072.
[10] Abdallah, Imad, et al. "Fault diagnosis of wind turbine structures using decision tree learning algorithms with big data." Safety and Reliability–Safe Societies in a Changing World. CRC Press, 2018. 3053-3061.
[11] Que, Zijun, and Zhengguo Xu. "A data-driven health prognostics approach for steam turbines based on xgboost and dtw." IEEE Access 7 (2019): 93131-93138.
[12] Myles, Anthony J., et al. "An introduction to decision tree modeling." Journal of Chemometrics: A Journal of the Chemometrics Society 18.6 (2004): 275-285.
[13] Fayyad, Usama M., and Keki B. Irani. "On the handling of continuous-valued attributes in decision tree generation." Machine learning 8.1 (1992): 87-102.
[14] Singh, Sonia, and Priyanka Gupta. "Comparative study ID3, cart and C4. 5 decision tree algorithm: a survey." International Journal of Advanced Information Science and Technology (IJAIST) 27.27 (2014): 97-103.
[15] Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." R news 2.3 (2002): 18-22.
[16] Chen, Xu, and Chunhui Zhao. "Multivariate Time Delay Estimation Based on Dynamic Characteristic Analytics." 2020 39th Chinese Control Conference (CCC). IEEE, 2020.