ISSN 2616-5767 Vol.5, Issue 2: 30-35, DOI: 10.25236/AJETS.2022.050206

Wind Turbine Blade Icing Fault Prediction Based on SCADA Data by XGBoost

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

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

Abstract: *To deal with the icing problem of wind turbine blades, the traditional classification methods are introduced firstly in this paper, and the XGBoost model based on monitoring and data acquisition (SCADA) system is introduced to estimate the icing conditions of blades. Meanwhile, the generation process of the XGBoost model is introduced in detail. Finally, the superiority of the XGBoost model is verified by experiments. The results show that XGBoost has higher precision and efficiency than other methods.*

Keywords: *XGBoost; Wind Turbine Blade Icing Detection; Supervisory Control and Data Acquisition (SCADA); Data-driven*

1. Introduction

With the environmental problems becoming increasingly prominent, renewable energy is being extensively used. Meanwhile, wind power generation catches on among countries and regions of the world as a non-pollution and low-cost electricity industry [1]. To acquire greater development values, wind turbines usually built-in high-altitude areas, which quickly cause some blade icing problems, which leads to a series of consequences. Specifically, it has the following hazards: Firstly, after icing, the airfoil of the fan blade changes, leading to the decrease of wind energy capture capacity. In addition, the ice attached to the blade increases the energy required for blade rotation and finally leads to the power loss of the fan. Secondly, after the fan blade freezes, some structural parameters of the blade change directly, affecting its inherent modal parameters and inducing blade fracture. In the meantime, when the fan blade ice accumulates to a certain extent, the ice will break and fly out under the influence of dead weight, which will quickly hit the inspection personnel in the wind field and cause personal accidents. If it isn't disposed of in time, this will bring irreversible damages to the system. Even blades without anti-icing and de-icing protection will be at risk of downtime. Therefore, icing detection on wind turbines is particularly significant for enhancing efficiency in the electric power industry and prolonging the wind turbines' operational life [2].

In terms of traditional wind turbines (WTs) blade icing detection techniques, foreign countries are more mature than domestic ones [3]. The mechanism of the detection method measures the corresponding changes due to ice accretion to detect whether blades are frozen. For instance, icing detection of WT blade based on the ultrasonic guided-wave way measures changes of quality, refection characteristic, conductivity, heat conductivity, and permittivity [4]. We can also take advantage of the mechanical properties. When there is ice on the sensor, ice will augment the rigidity and resonant frequency of the sensor to calculate the thickness of the ice [5]. However, traditional WTs blade icing detection methods can result in high costs and aggrandize the mechanical complexity of WTs. From the above discussion, it is very significant to find a safe and reliable detection method. Many researchers have started using data-driven modeling based on SCADA systems, reducing maintenance costs [6].

The method of blade icing detection based on data-driven is equal to classification problems in mathematics. Also, there are many ways to figure out the classification problems. The most common way is logistic regression which is often used to solve dichotomous problems. K-Nearest Neighbor (KNN) is also one of the most fundamental algorithms in machine learning. The Fisher discrimination criterion is a classic supervised data dimension reduction method. But the fact is that the above three methods only use the distance between data to make a judgment, and they didn’t excavate the nonlinear characteristic between variables.

ISSN 2616-5767 Vol.5, Issue 2: 30-35, DOI: 10.25236/AJETS.2022.050206

To excavate the correlations between variables deeply, many algorithms on machine learning solve the problem. Firstly, decision tree is a sorting technique based on the tree structure. Secondly, support vector machine (SVM) is a classifier with the most significant spacing in the feature space. The hidden Markov model (HMM) is a classification way for processing time-series data. Although the above three approaches dig the relevancy of information correctly, they are single-mode models. Industrial data generally has multi-modal characteristics. Under the influence of working conditions, the relationship between variables is varied. To solve this difficulty, ensemble learning inspires us.

Currently, there are two ways commonly used in ensemble learning. One is random trees based on bagging. The other is XGBoost based on boosting. The ideology of bagging is not complicated. The weak learners need to be independent of each other. However, most methods cannot guarantee the independence of each learner. Also, there is no way to differentiate the weak learner impression. And the output prediction speed is slow. Compared with bagging, the ideology of boosting is more straightforward and practical. In recent years, XGBoost based on boosting caught on. Compared to the traditional boosting algorithm, its unique point is a lot of optimizations are made. For example, the loss function is optimized by using second-order Taylor expansion. They are using regularization to avoid overfitting. In engineering, Zhang et al. adopt the XGBoost algorithm to diagnose the fault in bearing [7]. Its specific process is divided into two parts. The first thing is collecting data using a vibrating sensor as input of the XGBoost model. Then, vibration data are extracted and influenced by equipment complexity and other factors.

The above literature review shows that the XGBoost model based on SCADA system performs quite well and convincingly for condition monitoring and fault diagnosis. The novel and advantages of the proposed method are summarized as follows: (a) deep learning is used to adaptively extract multilevel nonlinear features from SCADA data, which improves the feature extraction process and acquired feature performance, and lays a foundation for improving the diagnostic accuracy of the model. (b) Automatic use of CPU multi-threading parallel computing, while the algorithm's accuracy is also improved. Therefore, compared with traditional machine learning models, the XGBoost icing detection model based on monitoring and SCADA system has better detection accuracy and generalization ability.

2. Concept and Approach

Before introducing XGBoost, we should understand the concept of the decision tree. The decision tree is a standard machine learning method whose purpose is to classify new examples using models learned from a given training data set. As the name suggests, it makes classification decisions based on the tree structure. The final conclusion of the decision process corresponds to the decision result we predict. Each question posed in the decision process is a test of some properties.

Generally, a decision tree consists of a root node, several internal nodes, and several leaf nodes. The leaf node corresponds to the decision result. Each of the other nodes corresponds to a property test, and each node contains a sample set divided into child nodes according to the test attributes. The root node has the complete set of samples. The path from the root to each leaf corresponds to a sequence of decision trees. Classification and Regression Tree (CART) [8] is a kind of decision tree. The tree is a binary tree, meaning that each split produces two leaves.

XGBoost is an improved algorithm to integrate boosting based on Gradient Boosting Decision (GBDT) [9]. Its core idea is to use CART as the weak learner, each iteration based on the existing tree, adding a tree to fit the residuals.

First, determining the predicted value of the initial tree:

$$\hat{y}_i^{(0)} = f_0(x_i) = 0 \quad (1)$$

Then the prediction function of the t-th tree is:

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) \quad (2)$$

Where x_i denotes the i_{th} sample, t denotes the number of CART in the model, $f_k(x_i)$ denotes the predicted value of the i_{th} sample in the t_{th} tree.

ISSN 2616-5767 Vol.5, Issue 2: 30-35, DOI: 10.25236/AJETS.2022.050206

The objective function of the XGBoost:

$$obj^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \Omega(f_i)$$
$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|\omega\|^2 \quad (3)$$

We can conclude that XGBoost adds a regular item based on the error function. Its purpose is to prevent overfitting, reduce the complexity of the trees and improve generalization ability, where T represents the number of leaves in a tree, and ω denotes the score of the leaf nodes. γ and λ are the penalty coefficients. y_i is the label value of the i_{th} sample. $\hat{y}_i^{(t)}$ represents the predicted value went through t iterations.

The objective function can be simplified by the second-order Taylor expansion:

$$obj^{(t)} = \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i \right) \omega_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \beta \right) \omega_j^2 \right] + \gamma T \quad (4)$$

We can define:

$$G_j = \sum_{i \in I_j} g_i, H_j = \sum_{i \in I_j} h_i \quad (5)$$

To minimize this objective function, we can set its derivative to zero and the optimal fraction of each leaf node is obtained:

$$\omega_j^* = -\frac{G_j}{H_j + \beta} \quad (6)$$

Plugging in the objective function, the minimum loss is expressed as:

$$obj^{(t)} = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j^2 + \beta} + \gamma T \quad (7)$$

Fig. 1 can clearly show the running process of the XGBoost.

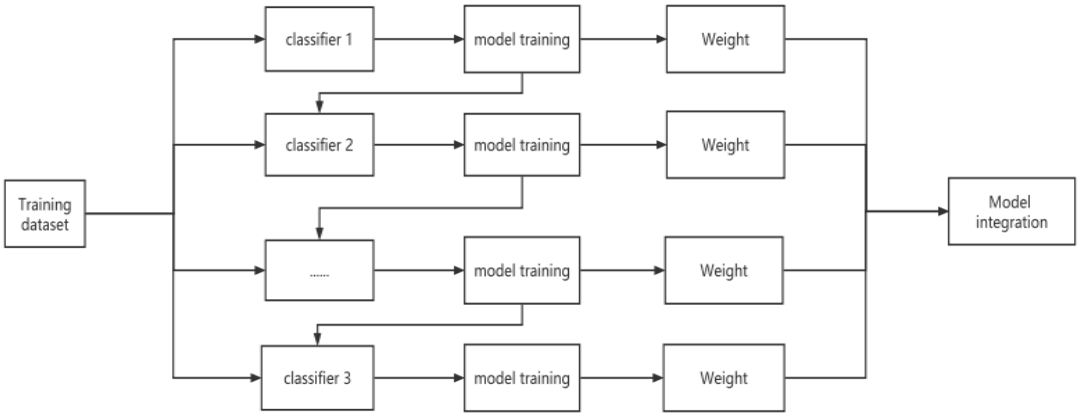

Figure 1: XGBoost running process

The greedy algorithm of enumerating all tree structures can be used to solve the problem of how to split a leaf node. In the meanwhile, set the tree depth and the tree stops growing when the gain is below a set threshold [10].

ISSN 2616-5767 Vol.5, Issue 2: 30-35, DOI: 10.25236/AJETS.2022.050206

3. Experiment and Discussion

SCADA systems are currently used in large wind turbines. SCADA system is controlled by the older generation algorithm, which is the integrated use of the signal acquisition, on-line monitoring and signals analysis function of the system, which can make the data acquisition, parameter regulation of wind turbines, equipment control, and fault alarm, etc., is generally used in wind turbine condition monitoring and fault diagnosis, can provide the reliable operation of the wind farm with powerful technology platform support. This paper mainly uses the SCADA system collecting and recording the data information to detect whether the fan blades freeze.

The positive and negative samples were mixed and randomly shuffled. Among them, 200 pieces of data are used as training sets. There are 26 variables. The variables are detailed in Table 1.

Table 1: Variables name

Variables			
Wind_speed	generator_speed	power	wind_direction
wind_direction_mean	yaw_position	yaw_speed	pitch1_angle
pitch2_angle	pitch3_angle	pitch1_speed	pitch2_speed
pitch3_speed	pitch1_moto_tmp	pitch2_moto_tmp	pitch3_moto_tmp
acc_x	acc_y	environment_tmp	int_tmp
pitch1_ng5_tmp	pitch2_ng5_tmp	pitch3_ng5_tmp	pitch1_ng5_DC
pitch2_ng5_DC	pitch3_ng5_DC		

The model heap established above was used to analyze the 1879 test data to determine whether ice was formed. We use XGBoost(XGB), K-Nearest Neighbors(KNN), decision tree classifier(DTC), Logistic Regression(LR), Naive Bayes(NB), Support Vector Classification(SVC) for comparison. We use error rate, accuracy rate, recall rate, and F1-score to represent the effect of the model to evaluate the quality of the model. The effects of different methods are detailed in Table2. The error rate represents the proportion of the samples with the wrong classification to the total samples, which is an indicator for evaluating classification models. To evaluate the model from the whole perspective, the evaluation criteria are based on the whole sample set. As the name implies, the lower the error rate, the better the model. The accuracy rate shows how many of the positive samples are genuinely positive. That is, the evaluation criteria are based on the predicted results. And the model with higher accuracy is more suitable for predicting blade icing. Recall rate represents how many favorable forces in the sample are correctly predicted. Recall rate refers to the original sample. Generally speaking, the higher the recall rate, the lower the accuracy. F1-score refers to the harmonic average of recall rate and precision rate. When F1-score is high, this model is ideal. Also, we draw the ROC curve of the predicted results and calculate the AUC value. The larger the integral of the ROC curve is. That means the more extensive the value of AUC is, the better the classifier can distinguish positive and negative samples.

Table 2: Results of different models

Model	**Error rate**	**Precision rate**	**Recall rate**	**F1-score**
XGB	**0.002**	**0.997**	**1.0**	**0.998**
SVC	0.014	0.977	0.996	0.986
LR	0.033	0.939	0.998	0.967
NB	0.138	0.782	1.0	0.878
KNN	0.035	0.936	0.997	0.965
DTC	0.012	0.976	1.0	0.988

From Table 2 and Figure 1, it can be concluded that the error rates of linear classifiers such as KNN, LR, and NB are as high as 0.035, 0.033, and 0.138, which incorrectly distinguish the operating state of wind turbines, while the error rates of nonlinear classifiers, DTC and SVC are 0.012 and 0.014. Compared to linear classifiers, the error rate is reduced, but the effect is still not ideal due to the limitations of single-mode processing. The error rate of XGBoost reaches 0.002, and its classification effect is better than other models.

ISSN 2616-5767 Vol.5, Issue 2: 30-35, DOI: 10.25236/AJETS.2022.050206

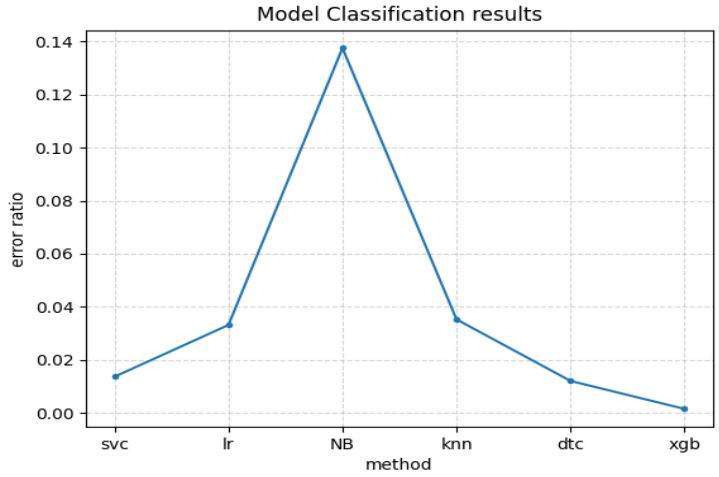

Figure 2: Error rate comparison

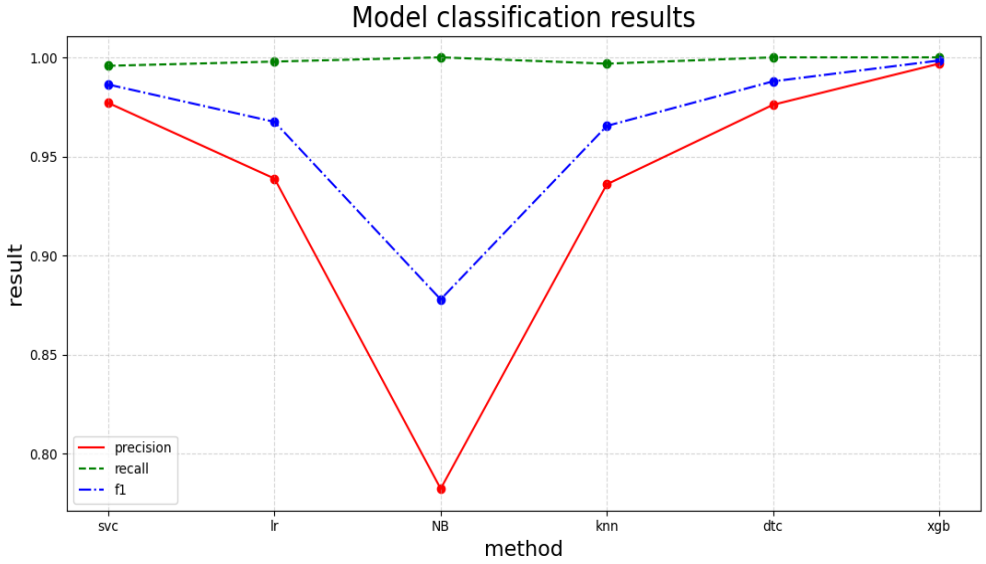

Figure 3: Comparison of other indicators

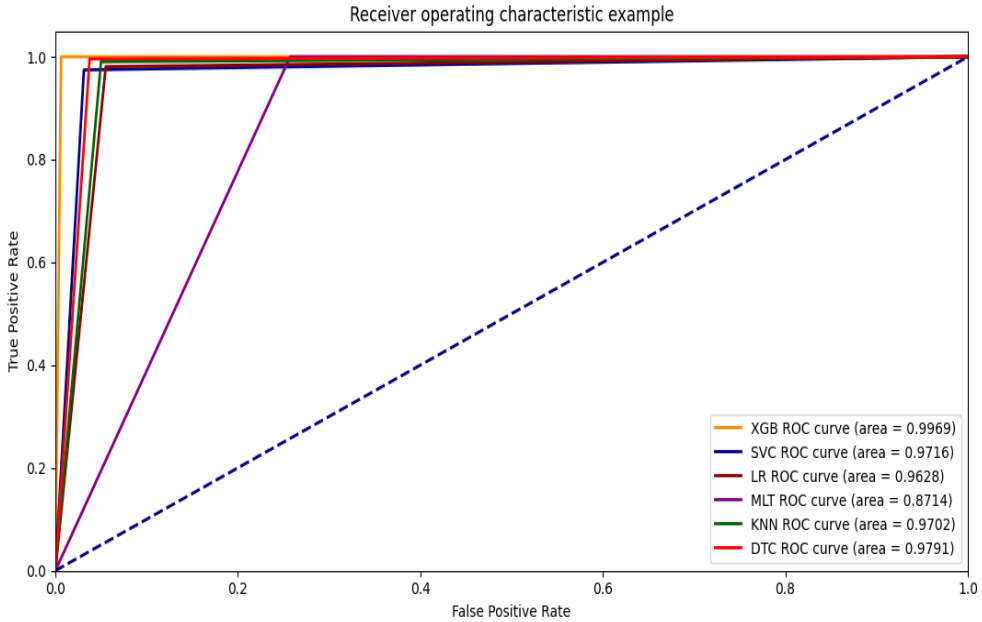

Figure 4: ROC curve and AUC value

ISSN 2616-5767 Vol.5, Issue 2: 30-35, DOI: 10.25236/AJETS.2022.050206

As shown from Figure 3, the accuracy rate is shown in the red line, and it is clear that XGBoost has the highest accuracy rate. Table 2 is 0.997, linear classifier KNN, LR, and NB are 0.936,0.939, and 0.782. Respectively, nonlinear classifier KNN and SVC are 0.936 and 0.977, which are not as low as XGBoost accuracy. The blue line represents the F1-score, and it can be clearly seen that XGBoost has the highest value (0.998), linear classifier KNN, LR and NB are 0.965, 0.967, and 0.878, respectively, and nonlinear classifier DTC and SVC are 0.965 and 0.986 respectively, all of which are lower than XGBoost F1-score.

The closer the ROC curve is to the upper left corner, the higher the totality of the model. The point on the ROC curve closest to the upper left corner is the best threshold for the fewest classification errors, with the fewest total number of false positives and false negatives. The AUC value represents the area under the ROC curve, and the more significant the area, the better the effect of the model. From Figure 4, we can see the ROC curve and AUC value of each model, of which the AUC value of XGBoost is 0.9969, and the AUC value of other models is below it XGBoost has more advantages in handling the icing detection of fan blades.

From the above results, it can be seen intuitively that XGBoost is significantly better than the other five methods. Three methods of Logistic Regression, Naive Bayes, KNN do not excavate the nonlinear characteristics between data, so the effect is relatively poor. SVC and DTC deal with the nonlinear model but not a multimodal model. So the effect is better than the above three methods. The XGBoost uses the idea of integrated learning. Both nonlinear models and multimodal models are processed. Therefore, it is better than the above five methods in error rate, precision rate, recall rate, and F1-score.

4. Conclusion

In order to predict blade icing more accurately, XGBoost method based on SCADA data is used to train and test. We compared five methods to validate the effectiveness and high efficiency of XGBoost. The research found that XGBoost has a 2% error rate, which was lower than other methods. XGBoost has the accuracy rate of 99.7%, which is higher than other methods. The results demonstrate the superiority and effectiveness of the proposed model.

References

[1] Tong W. Wind power generation and wind turbine design [M]. WIT press, 2010.
[2] Liu Y, Cheng H, Kong X, et al. Intelligent wind turbine blade icing detection using supervisory control and data acquisition data and ensemble deep learning [J]. Energy Science & Engineering, 2019, 7(6): 2633-2645.
[3] Homola M C, Nicklasson P J, Sundsbø P A. Ice sensors for wind turbines [J]. Cold regions science and technology, 2006, 46(2): 125-131.
[4] Dervilis N, Choi M, Taylor S G, et al. On damage diagnosis for a wind turbine blade using pattern recognition [J]. Journal of sound and vibration, 2014, 333(6): 1833-1850.
[5] Shi Q. Model-based detection for ice on wind turbine blades [D]. NTNU, 2017.
[6] Kreutz M, Ait-Alla A, Varasteh K, et al. Machine learning-based icing prediction on wind turbines [J]. Procedia CIRP, 2019, 81: 423-428.
[7] Zhang R, Li B, Jiao B. Application of XGboost algorithm in bearing fault diagnosis [C]//IOP Conference Series: Materials Science and Engineering. IOP Publishing, 2019, 490(7): 072062.
[8] Loh W Y. Classification and regression tree methods [J]. Encyclopedia of statistics in quality and reliability, 2008, 1: 315-323.
[9] Wang Y, Sun S, Chen X, et al. Short-term load forecasting of industrial customers based on SVMD and XGBoost [J]. International Journal of Electrical Power & Energy Systems, 2021, 129: 106830.
[10] Samat A, Li E, Wang W, et al. Meta-XGBoost for hyperspectral image classification using extended MSER-guided morphological profiles [J]. Remote Sensing, 2020, 12(12): 1973.