Research on Fault Diagnosis of Wind Turbine Based on SCADA Data

YIRONG LIU[1], ZIDONG WU[2], AND XIAOLI WANG[1]

[1]School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai 264209, China
[2]Nari Technology Company Ltd., Nanjing 211106, China

Corresponding author: Xiaoli Wang (wxl@sdu.edu.cn)

Yirong Liu and Zidong Wu are co-first authors.

ABSTRACT Effective early warning of wind turbine failures is of great significance to reduce the operation and maintenance costs of wind farms and improve power generation efficiency. At present, most wind farms are installed with supervisory control and data acquisition (SCADA) system, and SCADA data contains a lot of hidden information, which can be used for fault early warning. This paper uses the generator temperature and gearbox oil temperature in the SCADA data as the entry point for fault warning. Firstly, the eXtreme gradient boosting (XGBoost) algorithm is used to establish the normal temperature regression prediction model of wind turbine components. Then, the residual between the predicted value and the actual value is calculated, and the change trend of the residual is monitored by the principle of exponentially weighted moving-average (EWMA) control chart. Finally, by setting an appropriate threshold, the variation trend of the residual is judged to determine the occurrence and development of the fault. This paper uses two fault detection methods: fixed threshold and dynamic threshold based on adaptive algorithm, and compares the advantages and disadvantages of the two methods. Based on the SCADA data of a wind farm in Inner Mongolia (China), this paper designs the fault early warning test of the wind turbine generator and gearbox. The experimental results show that for the generator, the fixed fault threshold method can give the fault alarm 3 hours in advance, while the dynamic fault threshold determination method can give fault alarm 4.25 hours in advance. For gearbox, the fixed fault threshold method can give the fault alarm 2 hours in advance, while the dynamic threshold fault diagnosis method can send out the fault alarm 2.75 hours in advance.

INDEX TERMS Wind turbine, fault warning, eXtreme gradient boosting (XGBoost), exponentially weighted moving-average (EWMA), supervisory control and data acquisition (SCADA).

I. INTRODUCTION

In recent years, with the continuous deterioration of global ecological environment and the gradual depletion of fossil fuels, countries all over the world have increased the research on renewable energy [1], [2]. As a clean and pollution-free renewable energy, wind energy has the advantages of wide distribution and huge reserves [3]. Therefore, the use of wind power generation has gradually become a new way to replace the traditional power generation [4], [5]. The global wind power industry has entered a period of rapid development, and the cumulative installed capacity of global wind power is increasing year by year [6]. The traditional wind power plant maintenance strategy relies heavily on regular maintenance and after-maintenance, and the deployment of spare parts has a long cycle, which leads to the high cost of failure

maintenance and has a huge impact on the operation and maintenance economy of wind power plants [7]. Therefore, how to make early warning before the occurrence of wind turbine failure is of great significance to reduce the operation and maintenance cost of wind farms and the long-term development of wind energy industry [8], [9].

At present, many achievements have been made in the research on fault prediction and diagnosis of wind turbine [10], [11]. The research on fault diagnosis of wind turbine based on condition monitoring system (CMS) is relatively mature [12]. [13] analyzes the vibration signals in time domain and detects the fault of components. However, the CMS system relies on high-performance sensors, acquisition cards and other hardware equipment, so it needs to add corresponding interfaces on the wind turbine, which increases the cost and brings inconvenience to the actual operation. The fault diagnosis method based on wavelet theory and neural network is mature and widely used in wind turbine

The associate editor coordinating the review of this manuscript and approving it for publication was Fanbiao Li[ID].

gearbox and blade fault diagnosis [14], [15]. However, this method needs additional vibration sensors and is only suitable for fault diagnosis of limited parts. The fault diagnosis method based on analytical model needs to fully understand the operation mechanism of the system, and can achieve the ideal diagnosis effect under the construction of appropriate mathematical model [16]. Wind turbine is a typical nonlinear dynamic system with complex structure. It is difficult to establish an accurate system mechanism model, and the model can't be applied to other industrial systems.

How to use more convenient data to study the fault of wind turbine has become a new research hotspot. In recent years, more and more wind farms use supervisory control and data acquisition (SCADA) system to record the operation status of wind turbines. SCADA data is widely used in wind turbine condition monitoring [17]. Hong Wang *et al.* proposed a deep confidence network based on SCADA data for feature learning and classification to realize sensor fault detection [18]. In addition to learning spatial correlation information between several different variables, this method can also capture the temporal characteristics of each variable. However, this method is only verified on the general benchmark model, which requires a large amount of real SCADA running data for further verification. [19] fits a support vector machines (SVM) regression to model gearbox oil temperature using selected variables in SCADA data as predictors, and uses the residual between the predicted value and the real value to predict the gearbox failure in advance. However, this method has two shortcomings. Firstly, SVM is a non-linear method in essence. When processing a large number of sample data, SVM will become obviously weak and sensitive to missing values in the data. However, the wind farm SCADA data accumulates quickly. In the long run, the regression prediction model of SVM algorithm is not suitable for processing a large number of SCADA data in the future. On the other hand, the conclusion given in this paper is that the wind turbine fault can be warned ten days in advance, which is obviously lack of practicability in practical application. If the wind turbine is stopped for maintenance ten days in advance due to the possibility of early failure, the wind turbine will stop for a long time and a large amount of power generation time will be lost. Therefore, when using SCADA data for fault warning, attention should be paid to the large amount of SCADA data and the setting of fault warning thresholds.

The development of machine learning and multi-sensor data fusion technology provides new ideas for wind turbine status monitoring and fault warning [20], [21]. Aiming at the problem of early warning of key parts of wind turbine, this paper proposes an early warning method for key parts of wind turbine based on SCADA data. This method uses XGBoost to establish the normal temperature regression prediction model of wind turbine components, and uses the residual change trend between the predicted value and the actual value as an early warning indicator. When setting the alarm threshold, this paper adopts the EWMA principle to control and monitor the change trend of the residual error,

and divides the normal, early warning, and alarm intervals by setting the control line to realize the gradual and multi-level early warning of the key parts of the wind turbine. However, since the complex operation conditions of wind turbine, the normal value deviation of SCADA parameters of wind turbine may occur in acceptable range under the influence of some uncontrollable objective factors. If the fault alarm system adopts fixed threshold setting, the residual extreme value may be higher than the threshold, which may lead to false alarm. In order to solve the problem of false alarm caused by fixed fault threshold, this paper designs a dynamic threshold setting method based on adaptive principle. In order to verify the effectiveness of the algorithm, this paper uses the SCADA of a wind farm in Inner Mongolia (China) in 2019 to conduct early warning experiments on wind turbine generator and gearbox failures. The results show that the fault early warning method based on fixed threshold proposed in this paper can detect the early fault characteristics of wind turbine generator 7.25 hours in advance, send out the generator fault alarm 3 hours in advance, find the early fault characteristics of gearbox 3.75 hours in advance, and send out the fault alarm of gearbox 1.75 hours in advance. While use the dynamic fault threshold method determined the fault, not only can adapt to the sudden residual extreme value and avoid false alarm, but also, it can send out generator fault alarm 4.25 hours in advance and gearbox fault alarm 2.75 hours in advance, which is 1.25 hours and 0.75 hours earlier than the fixed fault threshold.

The rest of this paper is organized as follows. Section II introduces the overall framework and main algorithm of fault early warning method. Fault warning of generator and gearbox is described in Section III and section IV respectively. Section V discusses the advantages of XGBoost compared with SVM regression prediction model. Finally, section VI is the conclusion and prospect of the whole paper.

II. ANALYSIS OF FAULT EARLY WARNING ALGORITHM
A. ALGORITHM OVERVIEW
The generator and gearbox of the wind turbine are the most frequent fault locations [17], [22]. In order to ensure the normal operation of the generator, the wind turbine generator set will set the cooling system and lubrication system to prevent the generator from overheating. The cooling system mainly uses the high-power cooling fan to draw out the heat generated by the generator operation to the outside of the engine room to achieve the cooling effect. In the whole process of heat dissipation, the temperature of the inner winding of the generator will be high when occurs some failure in generator such as the power of the fan motor decreases, In theory, when the ambient temperature and wind speed and power are in the same condition, if the temperature of a certain wind turbine generator is higher or the temperature rises faster than other turbines, it is necessary to stop and check if there is a fault in the generator. On the other hand, through the lubrication system, the bearing wear of

the generator is minimized to ensure the stable operation of the generator. Generally, when the bearing is damaged, it is often accompanied by bearing high temperature. The hidden danger of gearbox can be detected by the operator on duty by observing the temperature and temperature rise of the gearbox or comparing the temperature of the same position with other wind turbines. Usually, the wind turbine is cooled by radiator and fan when the wind turbine runs at high speed with strong wind. The oil temperature of gearbox is usually controlled at 60 °C. However, if it is observed that the gearbox oil temperature reaches 75 °C or even close to 80 °C, It is necessary to check whether the gearbox has radiator blockage, temperature control valve failure, fan blade damage, bearing wear and other problems. The generator and gearbox have complex internal structures, and the failure principles are diverse and coupled. However, most of the failures will eventually lead to abnormal temperature rise of the generator and gearbox oil [16], [23].

Therefore, the generator temperature and gearbox oil temperature in SCADA data can be used as the entry point of fault warning. This paper designs a fault warning algorithm based on temperature prediction of key components of wind turbines. The specific fault warning algorithm diagram is shown in Fig.1.
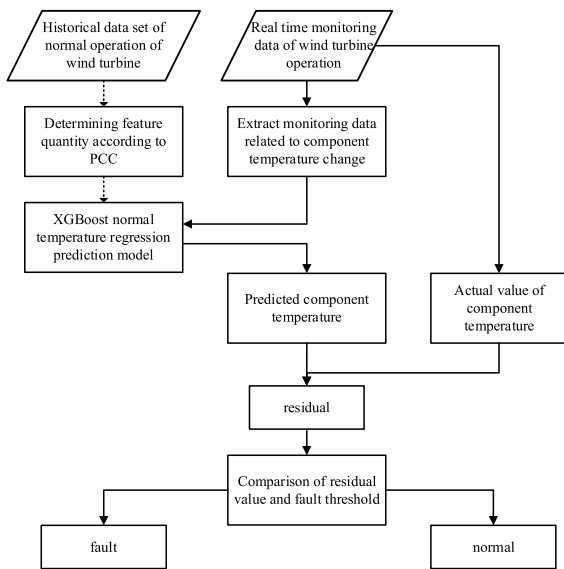

FIGURE 1. Block diagram of wind turbine fault warning algorithm.

The algorithm flow is as follows:
1) Get the modeling data set. The SCADA historical data of a normal operating wind turbine is selected to establish the normal temperature prediction model of the key parts of the wind turbine.
2) Select the characteristic quantity. Pearson correlation coefficient (PCC) is used to determine the monitoring items related to the temperature changes of wind turbine components.
3) Establish a normal temperature regression prediction model based on XGBoost algorithm. According to the

TABLE 1. Relationship between correlation coefficient value and correlation intensity.

Value range of correlation coefficient	Correlation intensity
0.8~1.0	Highly correlation
0.6~0.8	Strong correlation
0.4~0.6	General correlation
0.2~0.4	Weak correlation
0.0~0.2	Not correlation

temperature related characteristic parameters selected in step (2), a normal temperature regression prediction model of wind turbine components based on XGBoost algorithm is established.
4) The established model is used to predict the temperature of the key parts of the wind turbine running in real time.
5) Fault diagnosis. By setting the threshold to judge the change trend of the residual error, so as to judge the occurrence and development of the fault.

B. SELECT THE CHARACTERISTIC QUANTITY
The first work of constructing the temperature regression prediction model is to determine the input characteristic quantity [24]. Since the working state of the wind turbine is easily affected by weather conditions such as wind speed and wind direction, the temperature changes of the internal components of the wind turbine have the characteristics of randomness and volatility. In this paper, the PCC of internal component temperature data and other monitoring project data of wind turbine is calculated to determine the relevant monitoring data of the component temperature change.

This paper calculates the correlation coefficient by uses the SCADA data of wind turbine under normal operation condition as the experimental data set. The temperature data of a certain part of wind turbine in the experimental data set is taken as X, and the other observation characteristic data series are respectively taken as Y. The PCC R_i of X and Y_i can be obtained by substituting the temperature series X of a certain part of wind turbine and other observed characteristic series $Y_1, Y_2 \ldots \ldots Y_n$ into Eq.1.

$$R_i = r_{X,Y_i} = \frac{COV(X, Y_i)}{\sigma_X \sigma_{Y_i}}$$
$$= \frac{E(XY_i) - E(X)E(Y_i)}{\sqrt{E(X^2) - E^2(X)}\sqrt{E(Y_i^2) - E^2(Y_i)}} \quad (1)$$

After n times of calculation, n correlation coefficient values can be obtained. According to Eq.1, the value range of correlation coefficient is in the interval [0,1]. The greater the absolute value of correlation coefficient of any two vectors, the stronger the correlation between the two vectors. The relationship between the correlation value and correlation strength adopted in this paper is shown in Table 1.

In this paper, the monitoring items with correlation coefficient greater than 0.6 are selected as part of the input characteristics of the regression prediction model of wind turbine component temperature. In addition, due to the time continuity of the temperature signal, this paper uses the idea of time series prediction to add the component temperature value of the previous period to the input vector of XGBoost temperature regression prediction model. Finally, the input vector of the temperature regression prediction model consists of two parts: the observed characteristic parameters of SCADA system related to the temperature of the component and the average temperature value of the component within 15 minutes.

C. REGRESSION PREDICTION MODEL BASED ON XGBoost

XGBoost is a special gradient boosting decision tree (GBDT) algorithm, which is an improvement of the basic GBDT algorithm. GBDT is a machine learning algorithm composed of multiple classification and regression tree (CART) iterations in accordance with the gradient lift method and integrated learning ideas [25].

Like GBDT, XGBoost is composed of several CART trees. The negative gradient (the first derivative) of the loss function of the first mock exam model is needed in the GBDT training process to fit the negative gradient value when constructing the next model. In XGBoost, Taylor expansion is used to expand the loss function into binomial function (second derivative) to fit the tree model faster and better. The tree model of XGBoost can be represented by Eq.2.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \quad f_k \in F \tag{2}$$

where $\hat{y}_i$ is the predicted value; x_i is the i-th sample input; K is the total number of trees; F represents the function space of the decision tree (all CART trees); f_k is a function in the function space F. In order to better learn the above model, it is necessary to minimize the objective function. The objective function of XGBoost is shown in Eq.3:

$$Obj(\Phi) = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{3}$$

The objective function of XGBoost consists of two parts. The first part is the loss function l, which is used to measure the difference between the predicted value and the real value. The second part is the penalty term Ω of model complexity, which used to prevent over fitting of model output. The expansion of Ω is shown in Eq.4:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda||\omega||^2 \tag{4}$$

where γ is the regularization parameter of the number of leaf nodes, which is mainly used to inhibit the further splitting of nodes. λ represents the regularization parameter of leaf node weight to prevent leaf node weight from being too large. T is the number of leaf nodes; ω is the score of leaf node.

In the process of constructing CART decision tree, XGBoost algorithm solves the problem of bifurcation feature selection through greedy thought, and solves the problem of how to get the predicted score by finding the maximum value of objective function. When selecting bifurcation features, XGBoost uses greedy strategy to enumerate the objective function values and selects the feature with the minimum objective function value at the current time as the bifurcation feature. When calculating the prediction score of each leaf node, XGBoost calculates the minimum value of the objective function, and the maximum value point is the predicted score of the leaf node.

Since the internal use of XGBoost algorithm is gradient promotion strategy, in the construction of classification regression tree, not all the trees are obtained at once, but a new tree is added each time, and the previous test results are constantly patched while adding new trees. Assuming that the predicted value of the model after generating t-th trees is $\hat{y}_i^{(t)}$, the derivation process of the construction process of the XGBoost model is shown in Eq.5.

$$\begin{cases} \hat{y}_i^{(0)} = 0 \\ \hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \\ \hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \\ \vdots \\ \hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \end{cases} \tag{5}$$

The objective function of each layer is shown in Eq.6:

$$\begin{aligned} Obj^{(t)} &= \sum_{i=1}^{n} l(\hat{y}_i^{(t)}, y_i) + \sum_{i=1}^{t} \Omega(f_i) \\ &= \sum_{i=1}^{n} l(\hat{y}_i^{(t-1)} + f_t(x_i), y_i) + \Omega(f_t) \end{aligned} \tag{6}$$

The purpose of each layer of model construction is to find a f_t to minimize the objective function. The Taylor expansion of the objective function at $f_t = 0$ is approximately shown in Eq.7:

$$Obj^{(t)} \simeq \sum_{i=1}^{n} \left[l(\hat{y}_i^{(t-1)}, y_i) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{7}$$

where $g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$ is the first derivative; $h_i = \partial^2_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$ is the second derivative. By deleting the constant term in the formula, the objective function formula of step t is as shown in Eq.8:

$$Obj^{(t)} \simeq \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{8}$$

By introducing $\Omega(f_t)$ into Eq.8, it is concluded that:

$$Obj^{(t)} \simeq \sum_{i=1}^{n}\left[g_i w_q(x_i) + \frac{1}{2}h_i w_q^2(x_i)\right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

$$= \sum_{j=1}^{T}\left[G_j w_j + \frac{1}{2}(H_j + \lambda)w_j^2\right] + \gamma T \qquad (9)$$

where $G_j = \sum_{i \in I_j}g_i$, $H_j = \sum_{i \in I_j}h_i$ and $I_j = \{i|q(x_i) = j\}$ are samples in the j-th leaf node sample set. By setting the derivative of the objective function to zero, the point with the minimum derivative value of the function is the predicted score of the leaf node:

$$w_j^* = -\frac{G_j}{H_j + \lambda} \qquad (10)$$

The minimum value of the objective function is obtained as shown in Eq.11.

$$Obj^* = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j + \lambda} + \lambda T \qquad (11)$$

The core of wind turbine fault warning algorithm is to establish the normal temperature regression prediction model of wind turbine components by using the historical SCADA data of normal components. The construction process of regression prediction model of wind turbine component temperature is shown in Fig.2.
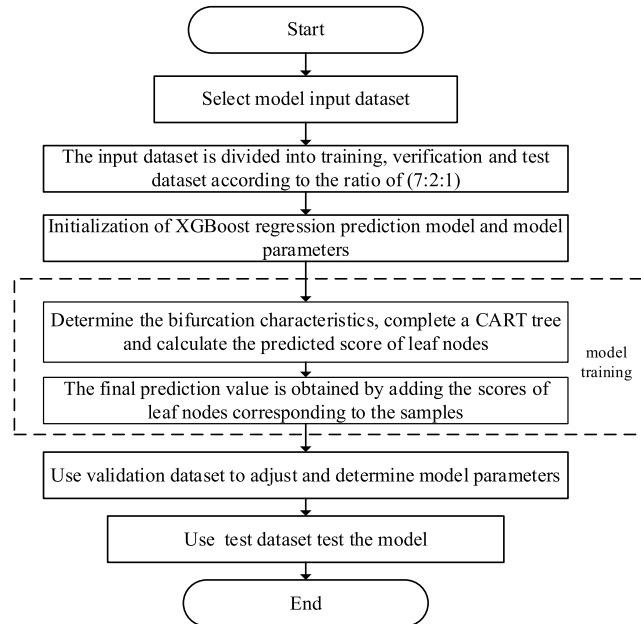

FIGURE 2. Temperature regression prediction process.

The specific construction process is as follows:
1) The model input parameters are extracted from the original SCADA monitoring data to construct the input data set. Selecting the relevant parameters and the temperature value of the previous period of time according to the PCC, and then use the historical data of these parameters to build the data set.

2) Divide the data set. According to the ratio of 7:2:1, the sorted data sets are divided into training, verification and test data sets.
3) Initialize XGBoost model. Set the model parameters, including the maximum depth of constructing decision tree, the learning rate of the model, the total number of training times, the number of threads used, and the method of specifying learning objectives and learning tasks.
4) Training model with training data set. In training the model, a CART decision tree is constructed firstly, and then the bifurcation characteristics are determined with the help of greedy algorithm to minimize the loss function, and the predicted score of leaf nodes is calculated to complete the construction of the second tree. By cycling the above steps, K trees with K classified feature nodes are finally constructed.
5) Adjust the parameter model by the training set. Constantly adjust the parameters for multiple validation set prediction experiments, and select a set of parameters with the highest accuracy of model prediction results as the final parameters of the model.
6) Verifying the accuracy of the prediction model by the test data set.

D. FAULT DIAGNOSIS
By inputting the SCAD data of the running wind turbine into the above prediction model, the predicted temperature values of the wind turbine components under normal operation can be obtained. The residual difference between the predicted temperature value and the measured temperature value in SCADA data represents the degree to which the current temperature state of the component deviates from the normal state. Therefore, by setting appropriate threshold, the occurrence and development of fault can be judged according to the trend of residual error. Based on Exponentially weighted moving-average (EWMA) control chart principle and 3sigma theory, this paper designe two fault threshold setting methods.

1) RESIDUAL TREND CHART BASED ON EWMA PRINCIPLE
EWMA is often used in statistical data processing, which fully considers the information of all previous observations in the form of setting weighting coefficient to reflect the recent change trend of target quantity [26], [27]. In this paper, the control chart based on EWMA principle is used to monitor the change trend of residual value, and the normal, early warning and alarm intervals are divided by setting partition. The expression of EWMA control point value is shown in Eq.12:

$$v_t = \beta Re_t + (1 - \beta)v_{t-1} \qquad (12)$$

where Re_t is the residual at time t. The coefficient β represents the weight coefficient of EWMA control chart to historical data, $\beta \in (0, 1]$, set $\beta = 0.9$. v_0 is the mean value of the first four sampling point. In addition, since model prediction is always unavoidable with errors, processing residual by

EWMA not only reduces the fluctuation range of residual values, but also effectively eliminates the number of false alarm points, making the alarm algorithm more stable and accurate.

2) THE SETTING OF FIXED THRESHOLD BASED ON RESIDUAL TREND GRAPH

Firstly, the residual data set is obtained by calculating the residual between the predicted value and the measured value.

$$\text{Re}_t = Y_t - X_t \tag{13}$$

where Y_t is the predicted value, X_t is the measured value, and Re_t is the residual value. Then, calculate the expected value E and deviation σ of the residual data set. The fixed threshold function is designed as shown in Eq.14.

$$T = E(\text{Re}) + z\sigma \tag{14}$$

According to the 3-sigma criterion, if the residuals obey normal distribution, then 99.73% of the residual values are concentrated in the range of $(E - 3\sigma, E + 3\sigma)$, and almost all the values are in the range of $(E - 6\sigma, E + 6\sigma)$. Considering the inevitable error of the prediction model, a certain margin should be reserved when setting the threshold value. In this paper, the threshold value of 4 is set as the threshold of early fault warning, and the threshold value of 8 is set as the alert line of abnormal component temperature.

The temperature prediction residuals are calculated after the actual running wind turbine data is input into the model, and the calculated temperature residuals are compared with the set threshold. If the residual exceeds the early warning threshold, it indicates that the wind turbine component is in the initial stage of failure. If the residual continues to increase to the warning threshold, it indicates that the wind turbine component is about to fail and it is necessary to alert for the component failure.

3) THE DYNAMIC THRESHOLD SETTING METHOD BASED ON ADAPTIVE PRINCIPLE

The operation condition of wind turbine is complex. Under the influence of some uncontrollable objective factors, the normal value deviation of wind turbine SCADA parameters may occur within the acceptable range, which is specifically manifested as the extreme point on the residual value. If the fault alarm system adopts fixed threshold setting, it may lead to false alarm when the residual extreme value is higher than the constant threshold. Based on the fixed threshold setting method, this paper designed a dynamic threshold setting method according to adaptive principle. The specific steps are as follows:

Step 1: Set data window size. According to the principle of K-S test, if the k-value of K-S test of two data sets is more than 0.05, it can be considered that the two data sets have the same distribution law [28], [29]. In this paper, according to the K-S test principle, the length of the smallest data subset which can reflect the characteristics of the original data set is

selected as the size of the sliding window. As shown in Fig.3, firstly, select a certain range of data from the beginning of the data as the sub data set, and take K-S test with the original parent data set to test the similarity between the two data set. Then expand the data subset range to the right in turn until the k-value between the subset and the parent set is greater than 0.05, and record the length of the subset at this time, which is the window size N.
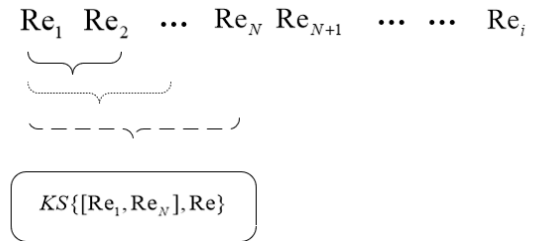

FIGURE 3. The schematic diagram of sub dataset selection.
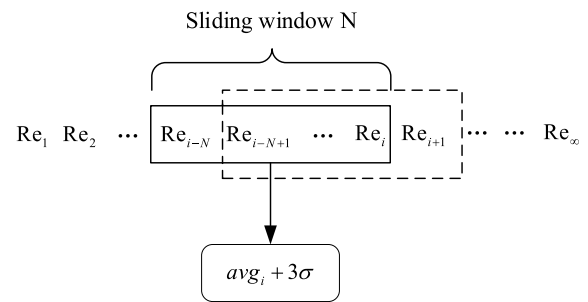

FIGURE 4. The schematic diagram of dynamic threshold setting.

Step 2: Calculate the threshold. As shown in Fig.4, according to the sliding window size N determined in the previous step, select the data within $\{\text{Re}_{i-N}, \text{Re}_{i-N+1}, \cdots, \text{Re}_i\}$ range to calculate the threshold value. Since the temperature of wind turbine changes slowly in normal state, the change amplitude of residual value is small. Therefore, when setting the dynamic threshold, this paper fully considers the residual change trend of the previous period in the window. This paper uses the method shown in Eq.15 to calculate the residual value in the window.

$$\frac{\sum\limits_{t=i-N}^{i} \text{Re}_t}{N} + 3\sigma \tag{15}$$

where Re_t is the residual value at time t, N is the size of the sliding window, σ is the standard deviation of residual under normal condition.

Step 3: Move the data window frame by frame and set a new threshold according to step 2.

Step 4: Repeat step 3 to get the threshold values at all times, and then connect them to form an adaptive threshold graph fitting the trend of residual Re_t.

III. GENERATOR FAULT WARNING

In order to verify the reliability of the fault warning algorithm, this paper uses the SCADA data record of a wind farm in

Inner Mongolia (China) as the experimental data. The rated power of the wind turbine used in the wind farm is 2000 *kw*, the cut-in wind speed is 3 *m/s*, the cut-out wind speed is 20 *m/s*, the impeller diameter is 110 *m*, and the data acquisition period of SCADA system is 30 *s*. In this paper, the normal operation of wind turbines and wind turbines with generator failures are selected as the control for the generator failure warning experiment.

A. GENERATOR TEMPERATURE RELATED PARAMETER SELECTION

This paper analyzes the data records of the SCADA system to find the wind turbine that has experienced a generator failure. The time of the wind turbine failure is 3:42:20 on October 20, 2019. In order to verify the validity of the temperature regression prediction algorithm, another normal operation data model is selected in this paper, and the data recording time interval is from August 1, 2019 to October 31, 2019. After deleting null values, removing outliers and normalizing the original data, the generator temperature in the model data is selected as X and the other monitoring items are Y according to Eq.1. The PCC between each monitoring item and the generator temperature is calculated as shown in Table 2.

TABLE 2. Generator related parameters.

| Monitoring items	PCC
Ambient wind velocity	0.891220162
Active power	0.826173287
Generator speed	0.785334182
Bearing temperature on drive side of generator	0.709255765
Ambient temperature	0.627165099
Generator slip ring temperature	0.617951245

B. XGBOOST REGRESSION PREDICTION OF GENERATOR TEMPERATURE

Since the temperature values of the generator change slowly and the time span of the original data is large, data prediction based on one sample point every 30 seconds will result in an excessive amount of data, which will slow down the prediction speed. Therefore, this paper predicts the average temperature of the generator every 15 minutes, that is, the monitoring item data of the input prediction model is the average value every 15 minutes. According to the construction process shown in Fig.2, the generator temperature regression prediction model was constructed. The minimum root mean square error (RMSE) of the final model was 0.484, and the minimum mean absolute error (MAE) was 0.335. Using the operating data of normal wind turbine from October 25, 2019 to October 31, 2019 to predict the generator temperature during this period, the results are shown in Fig.5.

The red curve in Fig.5 represents the predicted curve of generator temperature, and the green curve represents the
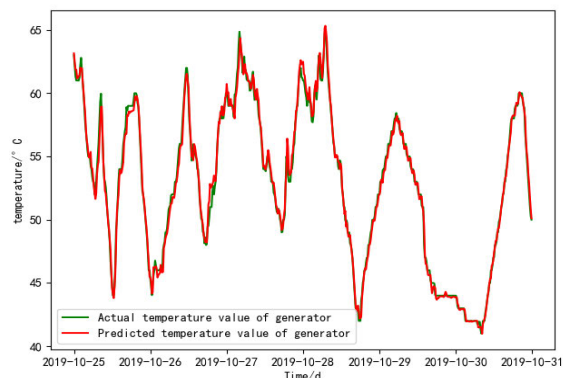
FIGURE 5. The temperature prediction result of the generator in normal operation.
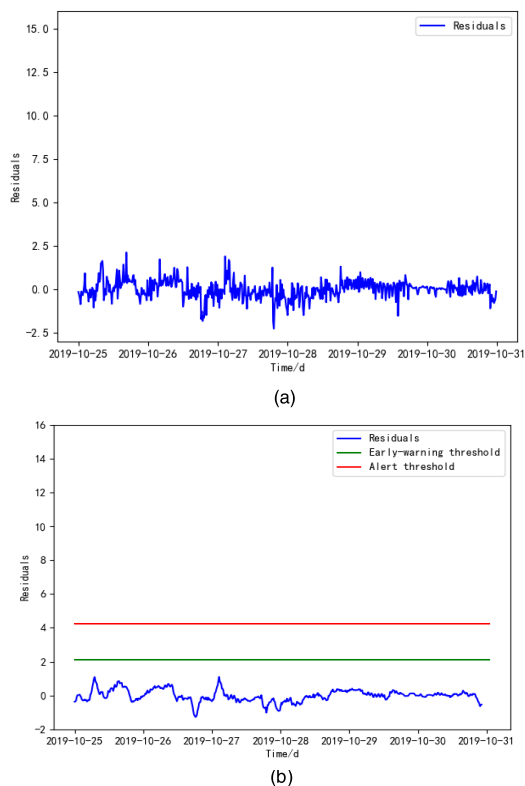
FIGURE 6. Residual distribution of normal generator temperature prediction model: (a). Before EWMA processing, (b). After EWMA processing.

actual temperature curve of generator. When the wind turbine generator is in normal operation state, the predicted temperature value of the model can better fit the actual temperature value of the generator. By calculating the residual between the predicted value and the actual value of the generator temperature, the distribution of the predicted residual of the generator temperature during normal operation is shown in Fig.6(a). It can be seen from the figure that during normal operation of the wind turbine, the predicted residual value of the model will show a certain symmetry near the zero point. Although there will be a few points with larger absolute residual value, the overall distribution is relatively uniform. The mean value

of the residual data set is -0.01 and the standard deviation is 0.503. By use the *kstest()* test function in Python 3.8, this paper tests the k-value between residual data and standard normal distribution is 0.125, which is greater than 0.05, so the residual data set conforms to normal distribution.

Fig.6 (b) shows the residual distribution map and alarm threshold after the residual is processed by EWMA principle. In the figure, the blue curve represents the residual distribution curve after calculation, the green line represents the early warning threshold, and the red line represents the alarm threshold. When the residual is below the green early warning threshold, it indicates that the generator temperature is normal. When the residual is between the green early warning threshold and the red alarm threshold, it indicates that the generator temperature has a high trend, which should be paid attention to. When the residual is above the red alarm threshold, it indicates that the generator is about to break down, and the corresponding fault treatment preparation should be made.

C. GENERATOR FAULT WARNING

In this paper, the wind turbine with generator failure is selected for early warning experiment. The failure time of the wind turbine is 3:42:20 on October 20, 2019, and the time interval of temperature prediction data set is from October 14, 2019 to 7:00, October 20, 2019. Input the preprocessed data set into the temperature regression prediction model, and the temperature prediction results are shown in Fig.7.
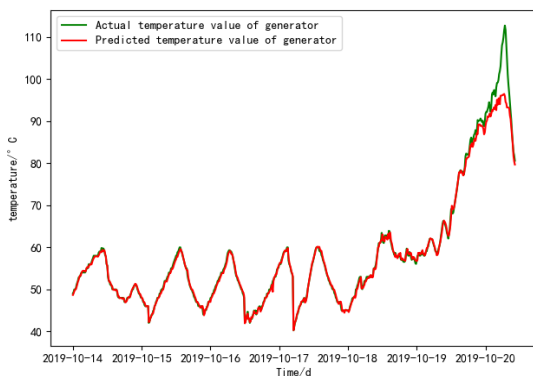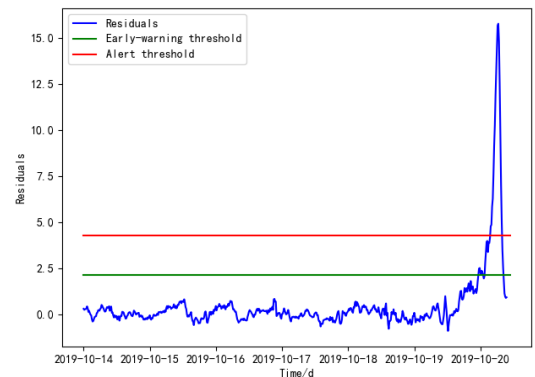
FIGURE 7. Temperature prediction results of fault generator.

In Fig.7, the red curve represents the change curve of the predicted generator temperature, and the green curve represents the variation curve of the measured generator temperature. It can be seen from the figure that when the generator is in normal operation, the predicted value of the model is in good agreement with the actual value. When the generator is about to fail, the deviation between the predicted value and the actual value will gradually increase. At 3:42 on October 20, 2019, the SCADA system detected the generator failure and took braking measures to stop the generator. Therefore, the actual temperature of the generator in the figure dropped sharply after reaching the maximum value.
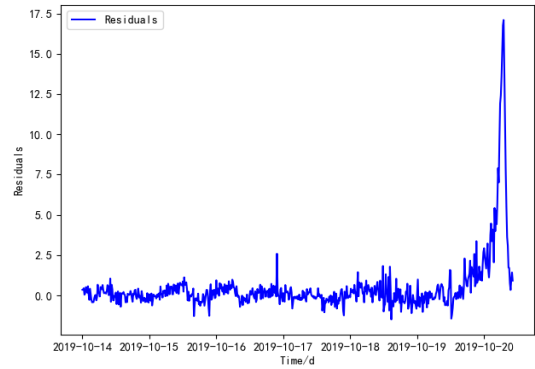
FIGURE 8. Residual distribution of fault generator temperature prediction model:(a). Before EWMA processing, (b). After EWMA processing.

1) FAULT DIAGNOSIS BASED ON FIXED THRESHOLD
The residual between the predicted value and the actual value of the model is shown in Fig.8(a). When the generator is in normal operation, the residual value is more evenly distributed. When the generator is about to fail, the predicted residual value of the model will gradually increase. Fig.8 (b) shows the residual value and fault threshold after EWMA processing.

It can be seen from Fig.8(b) that there is a big difference between the predicted residual value when the generator is about to fail and the residual value under normal operation. In order to show the effect of the model on fault prediction more clearly, this paper enlarges the curve of the rising part of the mean residual to get Fig.9.

In Fig.9, the blue curve represents the residual distribution curve, the green line represents the early warning threshold, and the red represents the alarm threshold.

It can be seen from Fig.9 that the residual value of the first crossing the early warning threshold point is the 44th sampling point, the first residual value crossing the alarm threshold point is the 61th sampling point, and the real fault point of the wind turbine is the 73th sampling point. Since there is a sampling point every 15 minutes, the early feature of generator fault can be detected 7.25 hours (29 sampling points) in advance by using the temperature regression early warning algorithm described in this paper, and it can give an
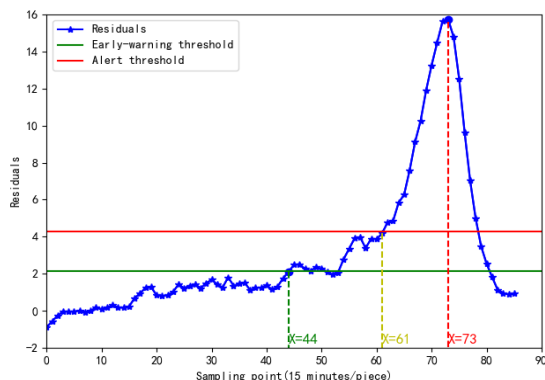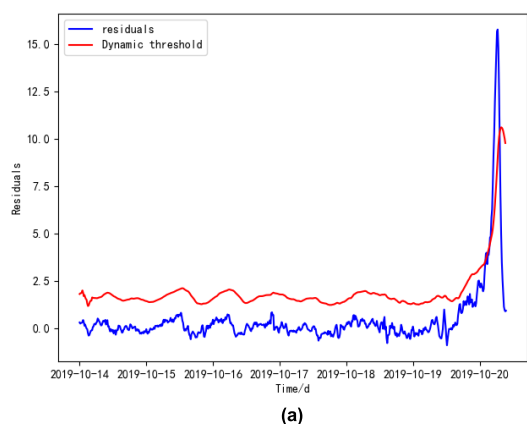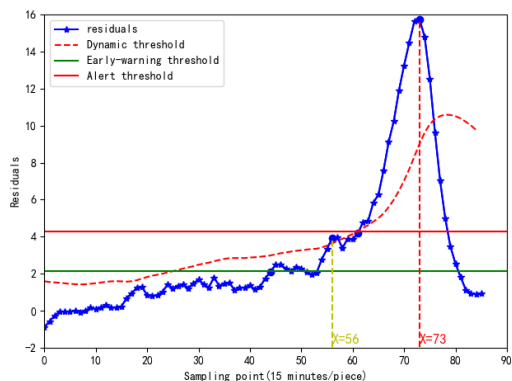
FIGURE 9. Local enlarged map of residual value of temperature prediction for fault generator.

alarm to the wind turbine fault 3 hours in advance (12 sampling points).

FIGURE 10. Dynamic fault threshold setting of fault generator: (a). Dynamic threshold graph, (b). Local enlarged drawing.

2) FAULT DIAGNOSIS BASED ON ADAPTIVE DYNAMIC THRESHOLD

According to the method in section II, the adaptive window size is determined as 25 sampling points, and the dynamic threshold of generator fault is set as shown in Fig.10 (a). It can be seen from the figure that the change trend of dynamic threshold basically matches the change trend of residual, and

the residual value exceeds the fault threshold value when the fault is about to occur. Similarly, in this paper, the curve of the rising part of the residual trend value is locally enlarged to obtain Fig.10 (b).

It can be seen from Fig.10 (b) that before the fault occurs, the dynamic threshold slowly increases with the change trend of the residual, and at the 56th sampling point, the residual value crosses the dynamic threshold for the first time. Therefore, using dynamic threshold to determine the fault can alarm the wind turbine 4.25 hours in advance (17 sampling points), which is 1.25 hours earlier than the fixed threshold (5 sampling points).

IV. GEARBOX FAULT WARNING

In order to further verify the universality of the fault warning method, this paper also carries out the early warning of gearbox fault. In this paper, a wind turbine with gearbox fault is selected from the actual wind farm as the experimental object, and a normal running wind turbine is selected as the modeling object. The fault time of the fault wind turbine is 18:32:33 on June 7, 2019, so the time for modeling the normal wind turbine is from May 1, 2019 to June 31, 2019. This paper selects the gearbox oil temperature in SCADA data to represent the gearbox temperature [16].

A. GEARBOX TEMPERATURE RELATED PARAMETER SELECTION

Before using the model to predict the gearbox temperature of wind turbine, it is necessary to determine the monitoring quantity related to the gearbox temperature change. Through the calculation of PCC, the final selection of gearbox temperature related monitoring items is shown in Table 3.

TABLE 3. Generator related parameters.

Monitoring items	PCC
Gearbox low speed bearing temperature	0.864347812
Active power	0.861943871
Gearbox high speed bearing temperature	0.837391140
Generator speed	0.728509119
Impeller speed	0.691370605
Ambient wind speed	0.682698046
Cabin temperature	0.648675280
Ambient temperature	0.603710142

B. REGRESSION PREDICTION MODEL OF GEARBOX TEMPERATURE

Like the generator fault warning, this paper also forecasts the average temperature of gearbox every 15 minutes, that is, the monitoring data input into the prediction model is the average value every 15 minutes. The generator temperature regression prediction model was built according to the construction process in Fig.2. The minimum RMSE and minimum MAE of the final model are 0.410 and 0.263 respectively.

The historical data generated during the normal operation of the wind trubine is input into the model, and the gearbox temperature prediction results under normal state are finally obtained, as shown in Fig.11.
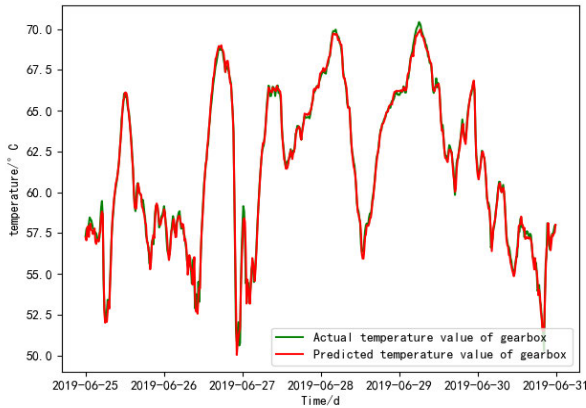

FIGURE 11. Temperature prediction results of normal operation gearbox.
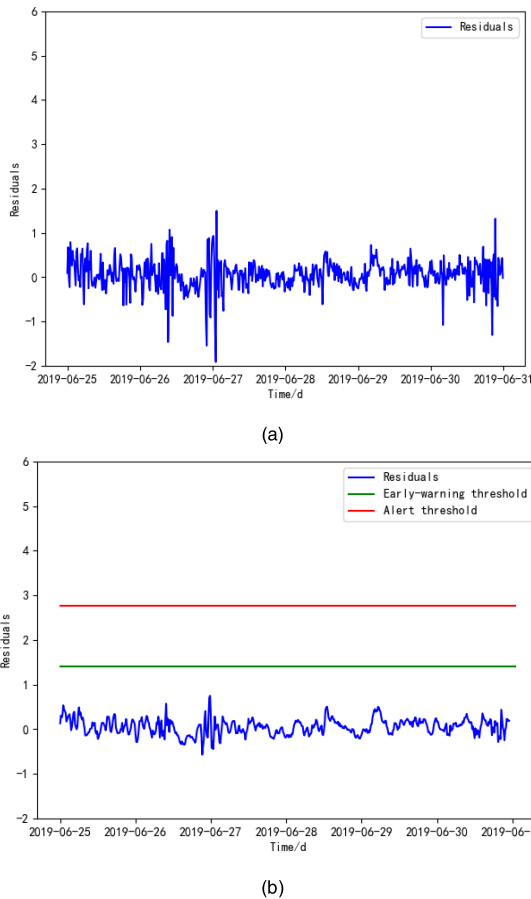

(a)

(b)

FIGURE 12. Residual distribution of normal gearbox temperature prediction model: (a). Before EWMA processing, (b). After EWMA processing.

The red curve in Fig.11 represents the predicted value of gearbox temperature, and the green curve represents the actual value of gearbox temperature. From the figure, when the gearbox is in normal operation, the predicted temperature

value of the model can better fit the actual value of gearbox temperature. The variation of predicted residual value of temperature during normal operation of gearbox is shown in Fig.12(a). The residual value distribution during normal operation of wind turbine will show a certain symmetry near the zero point. Although there are a few points with large residual error, the distribution is relatively uniform on the whole. The mean value of the residual data set is 0.052 and the standard deviation is 0.338. By use the *kstest*() test function in Python 3.8, this paper tests the k-value between residual data and standard normal distribution is 0.062, which is greater than 0.05, so the residual data set conforms to normal distribution.

The residual and threshold control areas processed by EWMA control principle are shown in Fig.12(b). The blue curve represents the residual mean change curve, and the red line represents the fault threshold.

C. GEARBOX FAULT WARNING

The wind turbine data record with gearbox fault is extracted from the SCADA system data record and used as the fault test data set. The time interval of the data set is from June 1, 2019 to 19:00, June 7, 2019. After preprocessing the data set, input the data into the established temperature regression prediction model, and the final temperature prediction results are shown in Fig.13.
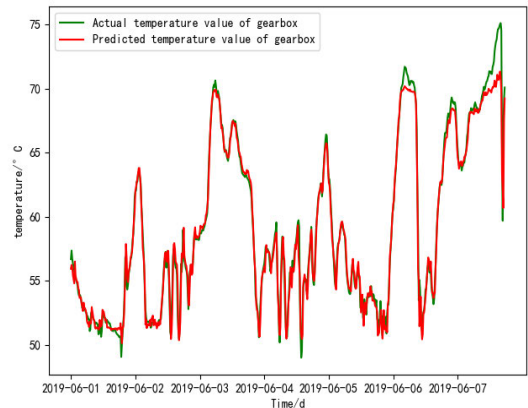

FIGURE 13. Prediction results of gearbox temperature under fault condition.

In Fig.13, the red curve represents the predicted gearbox temperature and the green curve represents the actual gearbox temperature change. When the gearbox is in normal operation, the predicted value and the actual value fit well. While when the gearbox is about to fail, the deviation between the predicted value and the actual value will gradually increase.

At 18:32:33 on June 7, 2019, SCADA system detected that the gearbox oil temperature was too high, and took braking measures to stop the wind turbine and the gearbox. Therefore, in the Fig.13, the actual temperature of the gearbox will drop sharply after reaching the maximum value.

1) FAULT DIAGNOSIS BASED ON FIXED THRESHOLD

The residual distribution between predicted and actual values is shown in Fig.14(a), when the wind turbine is in normal
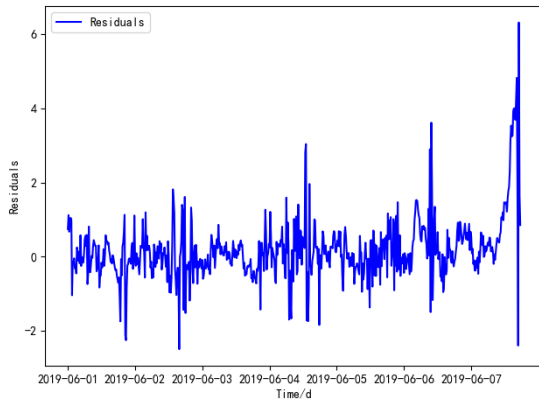
(a)
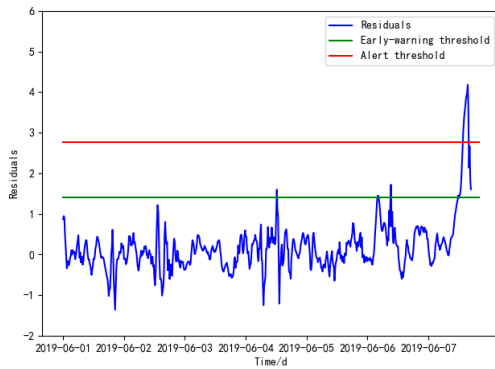
(b)

FIGURE 14. Residual distribution of fault gearbox temperature prediction model: (a). Before EWMA processing, (b). After EWMA processing.

operation, the residual value distribution is more uniform, and when the gearbox is about to fail, the residual value predicted by the model will show a gradual upward trend. According to EWMA control principle, residual threshold control area is divided as shown in Fig.14(b)

In Fig.14(b), the blue curve represents the residual distribution curve, the green straight line represents the early warning threshold, and the red represents the alarm threshold. Compared with the normal residual value distribution of gearbox, it can be found that the residual value fluctuates greatly when the gearbox is about to fail, and it will also show a certain upward trend. In order to show the effect of the model on fault prediction more clearly, the curve of the rising part of the residual is locally enlarged to get Fig. 15.

As can be seen from Fig.15, the first residual value crossing the early warning threshold point obtained by the prediction model appears at the 54th sampling point, the first residual value crossing the fault alarm threshold line appears at the 61th sampling point, and the gearbox fault point appears at the 69th sampling point. Through the calculation, it can be seen that the early fault characteristics of gearbox can be found 3.75 hours (15 sampling points) in advance by using the temperature regression early warning algorithm described in this paper, and send out the gearbox fault alarm 1.75 hours (8 sampling points) in advance.
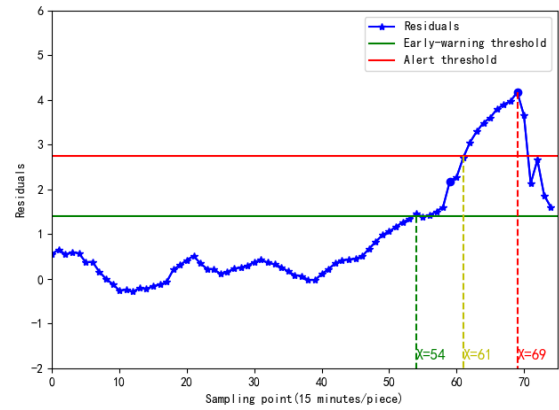
FIGURE 15. Local enlarged map of residual value of temperature prediction for fault gearbox.

2) FAULT DIAGNOSIS BASED ON ADAPTIVE DYNAMIC THRESHOLD

According to the method in section II, the adaptive window size is determined as 23 sampling points, and the dynamic threshold of gearbox fault is set as shown in Fig.16 (a). It can be seen from the figure that the change trend of dynamic threshold basically matches the change trend of residual, and the residual value exceeds the fault threshold value when the fault is about to occur. Compared with Fig.14 (b), it can be seen that using dynamic threshold can avoid the peak value of residual and avoid false alarm.

Similarly, in this paper, the curve of the rising part of the residual trend value is locally enlarged to obtain Fig.16 (b). It can be seen from Fig.16 (b) that before the fault occurs, the dynamic threshold slowly increases with the change trend of the residual, and the residual value crosses the dynamic threshold for the first time at the 58th sampling point. Therefore, using dynamic threshold to determine the fault can alarm the wind turbine 2.75 hours (11 sampling points) in advance, which is 0.75 hours (3 sampling points) earlier than the fixed threshold.

V. DISCUSSION

In order to effectively evaluate the performance of the temperature prediction model, this paper use the RMSE and MAE of the prediction model output to evaluation criteria to evaluate the model. The equation of RMSE is shown in Eq.15.

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2} \qquad (16)$$

where y_i is the true value and, $\hat{y}_i$ is the predicted value, and m is the number of data in the test set. The solution equation of MAE is shown in Eq.16.

$$MAE = \frac{1}{m} \sum_{i=1}^{m} \left| (y_i - \hat{y}_i) \right| \qquad (17)$$

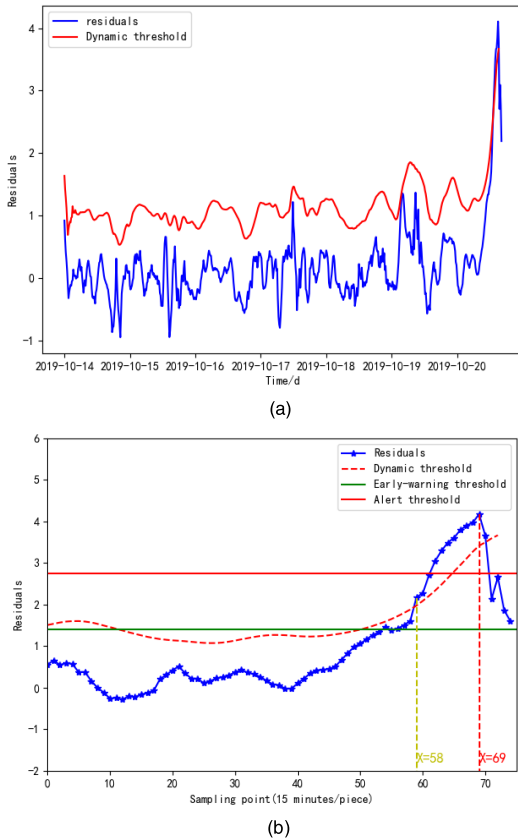where y_i is the true value and, $\hat{y}_i$ is the predicted value, and m is the number of data in the test set.

(a)

(b)

FIGURE 16. Dynamic fault threshold setting of fault gearbox:(a). Dynamic threshold graph, (b). Local enlarged drawing.

FIGURE 17. Evaluation results of regression prediction model.

In order to verify the prediction accuracy of the proposed algorithm, this paper uses experimental data to build a temperature prediction model based on SVR as a control experiment. In this paper, the wind turbine historical data in the generator early warning experiment is taken as the experimental data set. The data set is divided into training set, verification set and test set, which are used to train SVR temperature prediction model, adjust model parameters and verify the prediction effect. The final parameters of SVR temperature prediction model are: "C = 5", "kernel = RBF", "gamma = 0.01". The comparison between the final test results of the test set and the XGBoost temperature regression prediction model is shown in Fig.17.

It can be seen from Fig.17 that the RMSE and MAE obtained by using XGBoost to establish the regression prediction model are smaller than those obtained by using SVR modeling. The smaller the RMSE and MAE are, the more accurate the regression prediction results are. Therefore, the regression prediction model based on XGBoost described in this paper has the advantage of high accuracy.

VI. CONCLUSION AND PROSPECT
Early fault warning can effectively reduce the operation and maintenance costs of wind farms and improve the efficiency of power generation. In order to solve the problem of frequent faults in generator and gearbox of wind turbine, this paper
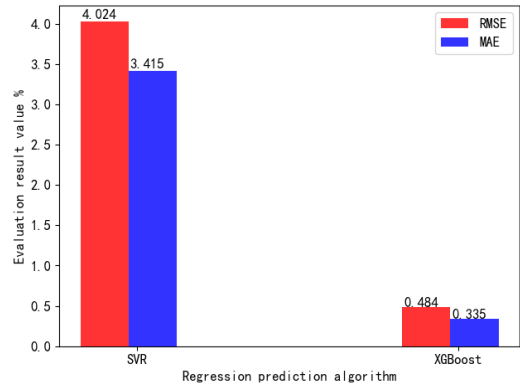
proposes a fault early warning method for key parts of wind turbine. In this method, XGBoost is used to establish the normal temperature regression prediction model of wind turbine components, and the residual change trend between the predicted value and the actual value is used as the early warning index. In the selection of characteristic quantity, in addition to using PCC to determine the monitoring data, this paper uses the idea of time series prediction to select the temperature value of the previous period of the component as the characteristic quantity, so that the selected characteristic quantity can reflect the temperature characteristics of the component more. This paper use XGBoost algorithm to construct the temperature regression prediction model. Compared with the regression prediction model constructed by SVM algorithm, XGBoost has higher prediction accuracy and is more suitable for the characteristics of large amount of SCADA data. In the process of fault state assessment, this paper uses the control chart based on EWMA principle to control the change trend of residual, and divides the normal, early warning and alarm intervals by setting control lines. However, Since the extreme value of the residual value, using fixed fault threshold may lead to false alarm. For this problem, this paper proposes a dynamic threshold setting method based on adaptive algorithm, which can avoid the false alarm caused by the extreme value of residual error, and can also early warn the fault. The fault warning method proposed in this paper has wide applicability, and has been verified in the fault warning of generator and gearbox. In theory, this method can be applied to other industrial systems with similar multi-sensor data structures.

In the next step, this paper will start from two aspects: feature selection and wind turbine overall fault warning. On the one hand, since the large amount of SCADA data monitoring, in order to improve the speed and accuracy of feature selection, we should consider a method to automatically select the best feature. On the other hand, Though the method in this paper can predict the faults of generator and gearbox of wind turbine, the wind turbine is a typical nonlinear and multi coupling system, and the relationship between different faults is complex and coupling is strong. How to use SCADA data

to accurately predict the overall fault of wind turbine will be a great challenge.

REFERENCES

[1] S. Yu, Y. Zheng, and L. Li, "A comprehensive evaluation of the development and utilization of China's regional renewable energy," *Energy Policy*, vol. 127, no. Nov. 2018, pp. 73–86, 2019.

[2] J. Brozyna, W. Strielkowski, A. Fomina, and N. Nikitina, "Renewable energy and eu 2020 target for energy efficiency in the Czech Republic and Slovakia," *Energies*, vol. 13, no. 4, pp. 1–20, 2020.

[3] B. K. Sahu, "Wind energy developments and policies in China: A short review," *Renew. Sustain. Energy Rev.*, vol. 81, no. Jan. 2016, pp. 1393–1405, 2018.

[4] M. Pettersson, K. Ek, K. Söderholm, and P. Söderholm, "Wind power planning and permitting: Comparative perspectives from the Nordic countries," *Renew. Sustain. Energy Rev.*, vol. 14, no. 9, pp. 3116–3123, 2010.

[5] C. El Mokhi, "Optimization of Wind Turbine Interconnections in an Offshore Wind Farm Using Metaheuristic Algorithms," pp. 1-24, 2020.

[6] *Global Wind Energy Council*, GWEC, Brussels, Belgium, 2015.

[7] H. Szumilas-Kowalczyk, N. Pevzner, and R. Giedych, "Long-term visual impacts of aging infrastructure: Challenges of decommissioning wind power infrastructure and a survey of alternative strategies," *Renew. Energy*, vol. 150, pp. 550–560, 2020.

[8] P. Zhang and D. Lu, "A survey of condition monitoring and fault diagnosis toward integrated O&M for wind turbines," *Energies*, vol. 12, no. 14, p. 151, 2019.

[9] K. Huang, Y. Wu, C. Yang, G. Peng, and W. Shen, "Structure dictionary learning-based multimode process monitoring and its application to aluminum electrolysis process," *IEEE Trans. Autom. Sci. Eng.*, vol. 17, no. 4, pp. 1989–2003, Oct. 2020.

[10] B. C. P. Lau, E. W. M. Ma, and M. Pecht, "Review of offshore wind turbine failures and fault prognostic methods," in *Proc. Prognostics Syst. Health Manage. Conf.*, Beijing, China, May 2012, vol. 15, no. 3, pp. 1–5.

[11] Z. Liu and L. Zhang, "A review of failure modes, condition monitoring and fault diagnosis methods for large-scale wind turbine bearings," *Measurement*, vol. 149, Jan. 2020, Art. no. 107002.

[12] Z. Hameed, Y. S. Hong, Y. M. Cho, S. H. Ahn, and C. K. Song, "Condition monitoring and fault detection of wind turbines and related algorithms: A review," *Renew. Sustain. Energy Rev.*, vol. 13, no. 1, pp. 1–39, Jan. 2009.

[13] J. Son, D. Kang, D. Boo, and K. Ko, "An experimental study on the fault diagnosis of wind turbines through a condition monitoring system," *J. Mech. Sci. Technol.*, vol. 32, no. 12, pp. 5573–5582, Dec. 2018.

[14] D. Strömbergsson, P. Marklund, K. Berglund, and P. Larsson, "Bearing monitoring in the wind turbine drivetrain: A comparative study of the FFT and wavelet transforms," *Wind Energy*, vol. 23, no. 6, pp. 1381–1393, Jun. 2020.

[15] J. Bi, "Fault diagnosis of wind turbine blades based on wavelet theory and neural network," *Int. J. Performability Eng.*, vol. 1, p. 939, Oct. 2019.

[16] H. Habibi, I. Howard, and S. Simani, "Reliability improvement of wind turbine power generation using model-based fault detection and fault tolerant control: A review," *Renew. Energy*, vol. 135, pp. 877–896, May 2019.

[17] J. Tautz-Weinert and S. J. Watson, "Using SCADA data for wind turbine condition monitoring—A review," *IET Renew. Power Gener.*, vol. 11, no. 4, pp. 382–394, 2017.

[18] H. Wang, H. Wang, G. Jiang, Y. Wang, and S. Ren, "A multiscale spatio-temporal convolutional deep belief network for sensor fault detection of wind turbine," *Sensors*, vol. 20, no. 12, pp. 1–14, Jun. 2020.

[19] D. Zhang and Z. Qian, "Probability warning for wind turbine gearbox incipient faults based on SCADA data," *Proc. Chin. Autom. Congr.*, 2017, pp. 3684–3688.

[20] Y. Lei, B. Yang, X. Jiang, F. Jia, N. Li, and A. K. Nandi, "Applications of machine learning to machine fault diagnosis: A review and roadmap," *Mech. Syst. Signal Process.*, vol. 138, Apr. 2020, Art. no. 106587.

[21] K. Huang, Y. Wu, C. Wang, Y. Xie, C. Yang, and W. Gui, "A projective and discriminative dictionary learning for high-dimensional process monitoring with industrial applications," *IEEE Trans. Ind. Informat.*, early access, May 6, 2020, doi: 10.1109/TII.2020.2992728.

[22] E. Artigao, S. Martín-Martínez, A. Honrubia-Escribano, and E. Gómez-Lázaro, "Wind turbine reliability: A comprehensive review towards effective condition monitoring development," *Appl. Energy*, vol. 228, pp. 1569–1583, Oct. 2018.

[23] Y.-B. Lee, G.-C. Lee, J.-D. Yang, J.-W. Park, and D.-C. Baek, "Failure analysis of a hydraulic power system in the wind turbine," *Eng. Failure Anal.*, vol. 107, Jan. 2020, Art. no. 104218.

[24] P. Marti-Puig, A. M. Blanco, J. J. Cárdenas, J. Cusidó, and J. Solé-Casals, "Feature selection algorithms for wind turbine failure prediction," *Energies*, vol. 12, no. 3, pp. 1–18, 2019.

[25] J. Ye, J.-H. Chow, J. Chen, and Z. Zheng, "Stochastic gradient boosted distributed decision trees," in *Proc. 18th ACM Conf. Inf. Knowl. Manage.*, 2009, pp. 2061–2064.

[26] H. Teng, X. Hu, A. Tang, W. Shu, and Z. Min, "An adaptive EWMA median chart for monitoring the process mean," in *Proc. Chin. Control Decis. Conf. (CCDC)*, Jun. 2019, pp. 3817–3821.

[27] M. Aslam, N. Khan, M. S. Aldosari, and C.-H. Jun, "Mixed control charts using EWMA statistics," *IEEE Access*, vol. 4, pp. 8286–8293, 2016.

[28] X. Wang and V. Makis, "Autoregressive model-based gear shaft fault diagnosis using the Kolmogorov–Smirnov test," *J. Sound Vib .*, vol. 327, nos. 3–5, pp. 413–423, Nov. 2009, doi: 10.1016/j.jsv.2009.07.004.

[29] J. Zhang, F. Liu, Y. Liu, H. Wu, W. Wu, and A. Zhou, "A study of accelerated life test of white OLED based on maximum likelihood estimation using lognormal distribution," *IEEE Trans. Electron Devices*, vol. 59, no. 12, pp. 3401–3404, Dec. 2012.
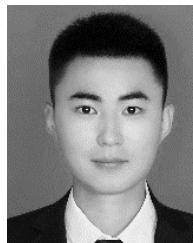
YIRONG LIU received the B.E. degree in engineering from the Shandong University of Science and Technology, in 2018. He is currently pursuing the master's degree with the School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai, China. His research interests include microgrid and wind turbine fault warning and diagnosis.

ZIDONG WU received the B.E. degree in electronic information engineering from the Qilu University of Technology, China, in 2016, and the M.E. degree in electronic and communication engineering from Shandong University, China, in 2020. He currently works with Nari Group Corporation. His current research interests include wind turbine fault warning and diagnosis.

XIAOLI WANG received the B.S. degree in electronic information science and technology and the M.E. degree in circuits and systems from Shandong University, Weihai, China, in 2002 and 2008, respectively. He is currently a Senior Experimentalist with the School of Mechanical, Electrical and Information Engineering, Shandong University. His current research interests include microgrid and embedded system development.

● ● ●