Applied Energy 168 (2016) 550–567

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy

A generalized model for wind turbine anomaly identification based on SCADA data

Peng Sun [a], Jian Li [a,*], Caisheng Wang [b], Xiao Lei [a]

[a] State Key Laboratory of Power Transmission Equipment & System and New Technology, Chongqing University, Chongqing 400044, China
[b] Department of Electrical and Computer Engineering, Wayne State University, Detroit, MI 48202, USA

HIGHLIGHTS

- A generalized model is presented for wind turbine anomaly identification.
- Prediction models are developed for the environmentally sensitive SCADA parameters.
- A new index is defined to quantify the abnormal level of wind turbine condition.
- A fuzzy synthetic evaluation method is used to integrate the identification results.
- Two case studies for an onshore wind farm are carried out and analyzed.

ARTICLE INFO

Article history:
Received 15 August 2015
Received in revised form 23 January 2016
Accepted 29 January 2016
Available online 15 February 2016

Keywords:
Anomaly identification
Fuzzy synthetic evaluation
Generalized model
SCADA data
Wind turbine

ABSTRACT

This paper presents a generalized model for wind turbine (WT) anomaly identification based on the data collected from wind farm supervisory control and data acquisition (SCADA) system. Neural networks (NNs) are used to establish prediction models of the WT condition parameters that are dependent on environmental conditions such as ambient temperature and wind speed. Input parameters of the prediction models are selected based on the domain knowledge. Three types of sample data, namely the WT's current SCADA data, the WT's historical SCADA data, and other similar WTs' current SCADA data, are used to train the condition parameter prediction models. Prediction accuracy of the models trained by these sample data is compared and discussed in the paper. Mean absolute error (MAE) index is used to select the prediction models trained by historical and other similar WTs' current SCADA data. Abnormal level index (ALI) is defined to quantify the abnormal level of prediction error of each selected model. To improve the accuracy of anomaly identification, a fuzzy synthetic evaluation method is used to integrate the identification results obtained from the different selected models. The proposed method has been used for real 1.5 MW WTs with doubly fed induction generators. The results show that the proposed method is more effective in WT anomaly identification than traditional methods.

1. Introduction

Wind energy is considered one of the most promising alternatives to traditional electric power and energy development [1]. Rapid developments of wind energy in recent years have drawn attention to issues on operation and maintenance (O&M) of wind farms [2–4]. The O&M costs of wind turbines (WTs) account for approximately 25–30% of the overall energy generation cost [5]. Various condition monitoring and fault diagnosis approaches of WTs, such as vibration analysis [6], acoustic analysis [7], lubrication analysis [8], and strain measurement [9], have been proposed to reduce unscheduled downtime as well as O&M costs. However, most of these approaches have not been widely used in wind farms because of the limitations with data storage capacities and high cost of installing additional equipment or sensors for condition monitoring systems of WTs. As one of the most important parts of the WT condition monitoring systems, the supervisory control and data acquisition (SCADA) system of a wind farm can provide a large amount of measurements such as the temperatures (e.g., bearing temperature, oil temperature), wind parameters (e.g., wind speed, wind direction), and energy conversion parameters (e.g., output power, pitch angle, rotor speed), which are widely used by wind farm operators to monitor the health condition of WTs. The SCADA data have attracted considerable research interest for being used in wind speed and wind power forecasting [10–13], power production assessment [14–17], and WT wake effect

* Corresponding author.

http://dx.doi.org/10.1016/j.apenergy.2016.01.133
0306-2619/

modeling [18–20]. Furthermore, the SCADA system records comprehensive WT condition parameters that could be fault informative [21,22]. Hence, detection of WT faults based on the SCADA data is a cost-effective approach to improving the reliability of WTs and reducing the O&M costs of wind farms.

Values of some WT condition parameters obtained from the wind farm SCADA system, such as output power, rotor speed, and component temperature, fluctuate quite significantly with the environmental conditions such as wind speed and ambient temperature. Those parameters are not appropriate to be used for WT condition monitoring and fault diagnosis directly. In order to solve this problem, various data mining and anomaly identification methods have been developed to mitigate the impacts of environmental conditions upon the real WT condition parameters [23]. The WT power curve, which could provide the relationship between power output and wind speed, is one of the most common tools in anomaly identification and performance analysis of WTs [24]. Researchers have applied different methodologies in estimating and monitoring the WT power curves based on wind farm SCADA data. In [25] three machine learning models, namely a generalized mapping regressor (GMR), a general regression neural network (GRNN) and a feed-forward multi-layer perceptron (MLP), were used to monitor the power curves of a wind farm. It was highlighted that the abnormal conditions of the WTs can be detected by using the residuals between the predicted power and observed power. In [26,27] SCADA data mining approaches were used for modeling the WT power curve and the anomaly identification capabilities of the different models were compared. In [28] the sectorial power curves of WTs were analyzed to investigate the relationship between wind direction and the WT power production. The power curves of two small-size WTs were compared in [29] to study the impacts of the ambient turbulence conditions upon the WT performance. A comprehensive review of the existing WT power curve monitoring techniques can be found in [30].

Modeling the normal behavior of SCADA parameters is another effective method for WT anomaly identification [31]. In comparison to the power curve monitoring technique, it could fully utilize the comprehensive operational information hidden within the SCADA system to identify anomalies in most important WT components. By using advanced SCADA data mining methods, various condition parameter prediction models have been developed to detect the significant changes in WT behavior prior to fault occurrences. In [32], the SCADA parameters were chosen to model the WT performance using multi-regime modeling approach. It was shown that the residuals of the proposed models could effectively predict the change of operational conditions prior to fault occurrences. Different WT performance curves, such as the power curve, rotor curve, and blade pitch curve were modeled in [33] for monitoring performance of WTs. The Mahalanobis distance was calculated to identify anomalies in the performance curves. A nonlinear data-based modeling approach was proposed in [34] to detect faults in WT generator winding and gearbox bearing. The obtained thresholds for WT winding and gearbox bearing temperatures can provide an early warning of WT component faults. In [35] a nonparametric regression model was established to investigate the relationship between the WT responses and a set of weather variables for the condition monitoring of WTs. By considering the influences of varying operational conditions and the control effects, a data preprocess approach was developed in [31] to extract useful information from the raw SCADA data. It was highlighted that the value of SCADA data depends on both the WT health condition and its operational conditions. In [36] normal behavior models of various WT condition parameters were established to detect incipient anomalies in the main components of a WT. The intelligent anomaly identification systems called multi agent system (MAS) [37] and SIMAP [38] were developed using the NN based prediction models of WT condition parameters such as gearbox bearing temperature, gearbox oil temperature and generator winding temperature. In [39,40] a WT condition monitoring method was proposed by using adaptive neuron-fuzzy inference systems (ANFIS) and NN based condition parameter prediction models. It was claimed that the prediction errors coming from successfully trained models are normally distributed with a mean around zero [39]. The existing statistical anomaly identification methods are mainly based on the assumption that normal instances occur in the high probability region of a stochastic model, while abnormal conditions happen in the low probability regions [41]. Thus, the prediction error (i.e. the difference between the measured value and model output) could provide an indication of parameter behavior changes and incipient WT faults. Certain thresholds of prediction errors are usually set to identify the anomalies in WTs. Two aspects should be considered when using these anomaly identification methods. Firstly, the sensitivity of anomaly identification is affected by the accuracy of the condition parameter prediction models [26,36]. Secondly, the training samples should be for the normal condition to assure the effectiveness [32,39].

Various data-driven approaches, such as NN [42,43], support vector machine (SVM) [44], ANFIS [45], and nonlinear state estimate technique (NSET) [46] were used to establish condition parameter prediction models. In [47] three regression models, namely NN model, the full signal reconstruction (FSRC) NN model, and autoregressive NN model, were developed for WT anomaly identification and their performances were compared. It was shown that the NN based models perform better and can give an earlier alarm of the WT faults. Performance of the prediction models established by six data mining algorithms, namely NN, SVM, boosting tree algorithm (BTA), random forest, generalized additive model, and the *k*-nearest neighbors, have been studied in [48]. In [48,49] different algorithms were used to obtain the relevant input parameters for the predicting target parameters. A genetic algorithm (GA) method was used in [39] to select the input parameters for 45 prediction models based on 33 SCADA condition parameters. In [38] three data-mining algorithms, specifically, BTA, wrapper with best first search (WBFS), and wrapper with genetic search (WGS), were used to select the most relevant input parameters for predicting the target condition parameter.

However, little research has been done on the impact of the training samples on the effectiveness of those proposed anomaly identification methods. There are no generalized rules for selecting training samples. Publication [47] pointed out that the period of training samples should be as short as possible, since the components of a WT are subjected to wear and degradation from the very beginning. The operational SCADA data collected in several weeks or several months were generally used to train the prediction models [32,36,47]. Yet, it is difficult to ensure that the training samples are under the normal condition. The accuracy and the effectiveness of anomaly identification may decrease when the training samples contain abnormal condition data. When collecting the training samples from historical SCADA data, the prediction accuracy may be reduced by wear and degradation of WT components. In addition, because significant differences may exist in the values of condition parameters for different WTs in a wind farm, the accuracy of prediction models trained by the sample data obtained on other WTs needs further investigation.

The motivation of the present paper lies in the identification of WT anomalies based on wind farm SCADA data. In this study, different WT condition parameter prediction models and a fuzzy theory method are integrated together to develop a generalized model for WT anomaly identification. The development of prediction models of rotor speed, output power, and component temperature is given in the paper. The WT's current SCADA data, the WT's historical SCADA data and other similar WTs' current SCADA data are

used to train the prediction models respectively and their performances are compared. The prediction models trained by different types of sample data are screened and selected by their performance of prediction accuracy. A new index called abnormal level index (ALI) is proposed to qualify the abnormal level of a WT condition parameter based its corresponding prediction model. Finally, a fuzzy synthetic evaluation method is utilized to integrate the anomaly identification result of each prediction model to detect anomalies in WTs. In comparison with the previous studies, the novelty of the paper can be summarized into two aspects:

Multiple prediction model integration: Instead of using a single prediction model, multiple prediction models trained by different types of sample data are integrated to detect the anomalies in WT condition parameters. The proposed method has demonstrated to be more effective in WT anomaly identification than the traditional single-model-based method.

Abnormal level quantification: Instead of setting a threshold for the prediction error, an innovative index called ALI is proposed to quantify the abnormal level of the WT condition parameters. Compared with the traditional threshold based method, the ALI based method can accurately identify the WT anomalies with less misdiagnosis.

The remainder of the paper is organized as follows: The SCADA data that can be used for WT anomaly identification are discussed and grouped in Section 2. Section 3 proposes the methodology for developing the generalized model. The condition parameter prediction models are developed and their performance is discussed in Section 4. Prediction models selection is illustrated in Section 5. Quantification of abnormal level of prediction errors is given in Section 6. The fuzzy synthetic evaluation method is presented in Section 7. Two cases are investigated to validate the proposed method in Section 8. The conclusion is drawn in Section 9.

2. Parameter description and classification

Table 1 shows typical condition parameters measured and delivered by the SCADA system of a wind farm. The positions of corresponding sensors are shown in Fig. 1. These condition parameters can be grouped into two types:

Type 1: Type 1 parameters include various component temperatures, WT output power and rotor speed, which are strongly influenced by environmental conditions. For example, the relationship between the gearbox input shaft temperature of a variable speed constant-frequency (VSCF) WT and the wind speed is shown in Fig. 2. For the VSCF WT, when the wind speed is below its rated speed, a faster rotational speed due to a higher wind speed will evidently raise the temperature of the mechanical components. When the wind speed is over the rated wind speed, the WT will be kept at its rated output power by variable pitch control and the component temperatures will be less affected by the wind speed. Fig. 3 shows the relationship between the gearbox input shaft temperature and the ambient temperature. The heat dissipation of mechanical components of a WT is also affected by the ambient temperature. At a lower ambient temperature, the heat dissipation is faster and the component temperatures can vary in a larger range. Conversely, at a higher temperature, the component temperature tolerable variation range is smaller with slower heat dissipation. For the parameters of output power and rotor speed, the wind speed is the most influential factor and wind farm operators usually use a power curve to estimate the power generated by a WT at different wind speeds.

Table 1
WT condition parameters studied in this paper.

No.	WT condition parameters	Unit	Type
1	Temp. of gearbox input shaft	°C	1
2	Temp. of gearbox output shaft	°C	1
3	Temp. of gearbox oil	°C	1
4	Temp. of gearbox cooling water	°C	1
5	Temp. of main bearing ***a*** (on the rotor side)	°C	1
6	Temp. of main bearing ***b*** (on the gearbox side)	°C	1
7	Temp. of generator winding ph.1 (U)	°C	1
8	Temp. of generator winding ph.2 (V)	°C	1
9	Temp. of generator winding ph.3 (W)	°C	1
10	Temp. of generator bearing ***a*** (front)	°C	1
11	Temp. of generator bearing ***b*** (back)	°C	1
12	Temp. of generator cooling air	°C	1
13	Temp. of control cabinet	°C	1
14	Temp. of converter controller	°C	1
15	Output power	kW	1
16	Rotor speed	rpm	1
17	Yaw angle error	°	2
18	Pitch angle error	°	2
19	Hydraulic oil pressure for yaw	bar	2
20	Hydraulic oil pressure for rotor brake	bar	2

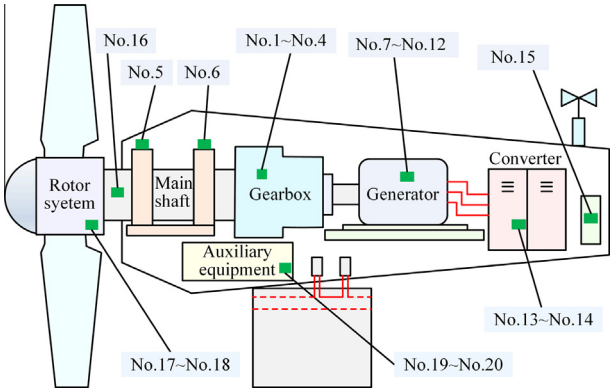


Fig. 1. The main components and sensor positions of the considered WT.

Type 2: Type 2 parameters include yaw angle error, pitch angle error and hydraulic oil pressure. Yaw angle error is the angle between the wind and the nacelle position. Pitch angle error represents the pitch angle that deviates from the set point. Type 2 condition parameters do not have an obvious relationship with environmental conditions. Since the anomaly identification for condition parameters of Type 2 can be easily done by setting certain threshold values, only the WT anomaly identification based on condition parameters of Type 1 will be discussed.

The SCADA data used in this paper are obtained from an onshore wind farm in Northern China. The wind farm has 34 1.5 MW WTs with doubly fed induction generators (DFIG), labeled WT 1 to WT 34. All the WTs are the same type in the wind farm. The SCADA data have been collected since February 15, 2011. The mean values of the parameters are used to reduce the

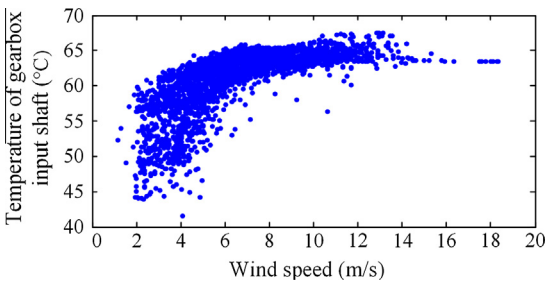


Fig. 2. Relationship between the gearbox input shaft temperature of a VSCF WT and the wind speed.

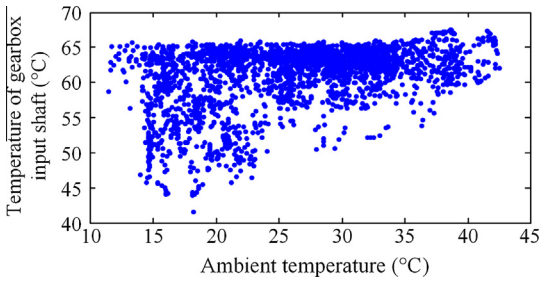


Fig. 3. Relationship between the gearbox input shaft temperature of a VSCF WT and the ambient temperature.

stochastic effects. The sampling rate of the SCADA system is one sample per second, i.e. 1 S/s.

3. Framework of the generalized model

Fig. 4 shows the framework of developing the generalized model for WT anomaly identification based on SCADA data. The generalized model consists of 16 condition parameter prediction models for WT condition parameters of Type 1. The framework has the following four parts:

Part 1 (Prediction model development): WT condition parameter prediction models are developed for those environmentally sensitive SCADA parameters. The input parameters and the training algorithm are determined at this stage. The prediction performance of the models established by four different training algorithms is analyzed. The accuracy of the models trained by different types of sample data is compared as well.

Part 2 (Prediction model selection): The models trained by different types of sample data are screened and selected according to their performance of prediction accuracy. In order to increase the selection efficiency, WT classification and sample similarity calculation are performed based on the distributions of target condition parameters.

Part 3 (Abnormal condition level quantification): Prediction error distribution for each type of condition parameter prediction models based on the corresponding SCADA data is investigated. Abnormal level index (ALI) is proposed to quantify the abnormal level of each condition parameter.

Part 4 (Fuzzy synthetic evaluation): The anomaly identification results obtained by individual prediction models are integrated based on fuzzy synthetic evaluation to detect anomalies in each WT condition parameter.

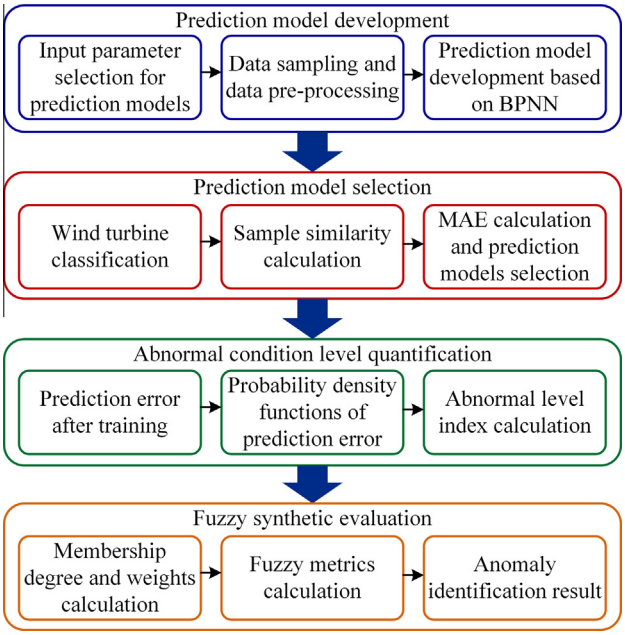


Fig. 4. Schematic of generalized model for WT anomaly identification.

4. Prediction model development

Generally, the current SCADA data collected in several weeks or several months were used to develop the condition parameter prediction models [32,36,47]. However, the current SCADA data used for training the prediction models may be impacted by the WT problems that already existed before and it is difficult to ensure that all the current data are under the normal condition. The accuracy and the effectiveness of anomaly identification may decrease when the training samples contain abnormal condition data. Hence, it is necessary to develop multiple prediction models trained by different types of samples such as the historical SCADA data and the SCADA data obtained from other WTs in the wind farm.

To develop the prediction models based on condition parameters of Type 1, the following three factors should be carefully considered: (1) input parameters, (2) sample data, and (3) training algorithm. The input parameters to each model are determined based on domain knowledge. NNs are used to establish the prediction models for environmentally sensitive SCADA parameters (i.e. Type 1 parameters). The performance of the prediction models trained by different types of sample data is compared.

4.1. Input parameters

The selection of input parameters is a premise in simplifying models and ensuring prediction accuracy. It requires a physical understanding of energy conversion of WTs with DFIGs. The input parameters of different types of condition parameter prediction models are shown in Table 2. The prediction models have two operating modes: normal mode and power limitation mode.

(1) *Normal mode:* Normal mode denotes the situation when the WTs are working in the operating region where the wind speed is lower than the rated value. In the normal mode, a fixed pitch is usually applied at the VSCF WT and the generator torque control is used to maximize the captured power from the wind [50,51]. In the normal mode, the wind speed, yaw error, and the pitch angle are selected as the input parameters to improve the accuracy of the models of output

Table 2
Input parameters of the prediction models based on Type 1 SCADA data.

Target condition parameters	Input parameters (SCADA data)	
	Normal mode	Power limitation mode
Output power (t)	Wind speed (t); yaw angle error (t); pitch angle (t)	Wind speed (t); yaw angle error (t); pitch angle (t)
Rotor speed (t)	Wind speed (t); yaw angle error (t); pitch angle (t)	Wind speed (t); yaw angle error (t); pitch angle (t)
Component temperatures (t)	Wind speed (t); output power (t); ambient temperature (t); component temperatures ($t-1$)	Output power (t); ambient temperature (t); component temperatures ($t-1$)

power and rotor speed. The component temperatures depend on the wind speed, the ambient temperature, the WT's output power and the previous component temperatures.

(2) *Power limitation mode:* Power limitation mode denotes the situation when the WTs are working in the operating region where the wind speed is above the rated value. In the power limitation mode, the pitch control system can keep the output power and the rotor speed at their rated values through changing the blade pitch angle about their longitudinal axis [51]. Hence, in the power limitation mode, when the wind speed exceeds its rated value (i.e. 12 m/s for the studied WTs), the increase of wind speed does not necessarily lead to continuous rise of component temperatures. Therefore, the wind speed is not selected as an input parameter to the prediction models of component temperatures in the power limitation mode.

4.2. Data sampling and data pre-processing

The following three types of training sample data are used for the condition parameter prediction models:

(1) *WT's Current SCADA Data:* The WT's current SCADA data were collected in the last 60 days (i.e. two months).
(2) *WT's Historical SCADA Data:* The WT's historical SCADA data were collected during the same period of time in the previous year.
(3) *Other Similar WTs' Current SCADA Data:* Other similar WTs' current SCADA data were collected in the last 60 days. The similar WTs denote the other WTs in the studied wind farm that have the same type with the target WT.

The testing sample which is used to obtain the prediction error distribution is extracted from the WT's current SCADA data during the most recent two months. To ensure the adequacy of data in analyzing the prediction error, the training/testing data sample ratio is set to 6/4.

When the wind speed varies within its low value region, it may cause frequent WT startup or shutdown. Thus, the data collected during these periods of time have different stochastic characteristics and should be filtered out. To solve this problem, a lower limit of output power is set at 100 kW for the 1.5 MW WTs for sample data selection [52]. In other words, only the data when the output power is greater than 100 kW are used to train the NNs. Furthermore, during certain instances in the operation of a WT, the output power can be curtailed by adjusting the blade pitch angle. This reduction in output power for a particular wind speed is due to change in control by the pitch regulation module and is not related to a degraded condition of the WT [32]. Moreover, the prediction models could not provide an accurate prediction since few SCADA data for this situation can be collected during a period of two months. Consequently, the data instances in which output power is being curtailed are filtered out. Averaging, scaling, and missing data processing are performed before training the NNs. For the component temperature parameters, the average values of 10 min sample data are used. For the output power and the rotor speed which have a shorter time constant, the average values of the sample data in 1 min are used.

4.3. Training algorithm

Back-propagation NN (BPNN), a widely used NN for prediction, is applied in this study for prediction of WT condition parameters. The transfer function used in the hidden layer is tan-sigmoid while the output layer transfer function is log-sigmoid based. In consideration of the size of the training set and the training time, the NN is chosen to have only one hidden layer and the number of neurons in the hidden layer ranges from 2 to 10. The actual number of nodes (i.e. neurons) in the hidden layer is determined by trial-and-error to find the right number at which the NN has the best generalization performance. It's worth noting that the number of hidden nodes of the BPNN model varies for each condition parameter of every WT. The relationship between the BPNN prediction performances of four condition parameters (i.e. gearbox input shaft temperature, generator bearing ***a*** temperature, generator winding ph.1 temperature and main bearing temperature) of a WT and the number of nodes (i.e. neurons) are taken as an example, as shown in Fig. 5. Based on this, the optimal hidden layer node numbers of the four prediction models are chosen as 6, 10, 7, and 6 respectively.

Based on the above chosen hidden node numbers, the BPNN model is trained and tested for 10 times with random weight initializations. Fig. 6 shows the BPNN prediction performances of the four condition parameters over the number of runs. It can be seen that the performance of the BPNN models varies with the different runs. With the suggested number of trials, the BPNN model with a better performance is chosen.

4.4. Performance analysis

After the BPNN models are selected according to Section 4.3, metrics such as mean absolute error (MAE), mean squared error (MSE), and mean absolute percentage error (MAPE) are used to analyze the performance of the condition parameter prediction models, as shown in (1)–(3).

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{1}$$

$$\mathrm{MSE} = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{2}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \times 100\% \tag{3}$$

where n is the number of test samples, $\hat{y}_i$ is the predicted value for a time period i, and y_i is the measured value at the same time.

Firstly, the selection of input parameters is verified. The 16 condition parameter prediction models are trained and tested by using the current SCADA data of WT 9. Fig. 7 shows the boxplot of the 16 condition parameters of the sample data. The median, 25th and 75th percentiles are shown with boxes, while the maximum, minimum and outliers are shown with whiskers and crosses, respectively. Table 3 shows the prediction performance of temperature condition parameters of a WT when the current SCADA sample data is used. Although the prediction accuracy of each component temperature is different, the maximum of MAE is lower than 1 °C. Table 4 shows the prediction performance of output power and rotor speed of a WT when the current SCADA sample data is used. As shown in the table, compared with the prediction models only with wind speed as the input parameter, the proposed prediction models have much higher prediction accuracy.

Then, the prediction accuracy of the prediction models trained by the aforementioned three types of sample data is compared. The temperature of generator bearing ***a*** (front) of WT 17 in the wind farm is taken as an example. The distributions of the three types of sample data are compared using boxplot. Fig. 8 shows the boxplot of wind speed, ambient temperature and generator bearing ***a*** (front) temperature of the current SCADA sample data and historical SCADA sample data of WT 17, which shows a similar distribution. Fig. 9 shows the boxplot of generator bearing ***a*** (front)

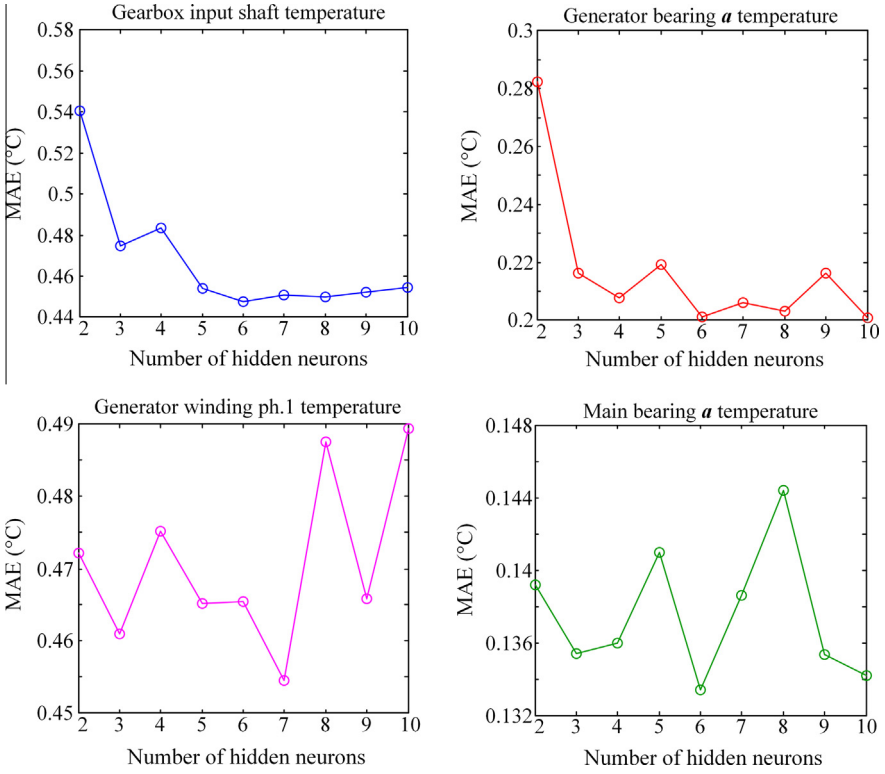


Fig. 5. The BPNN performances over the numbers of hidden nodes.

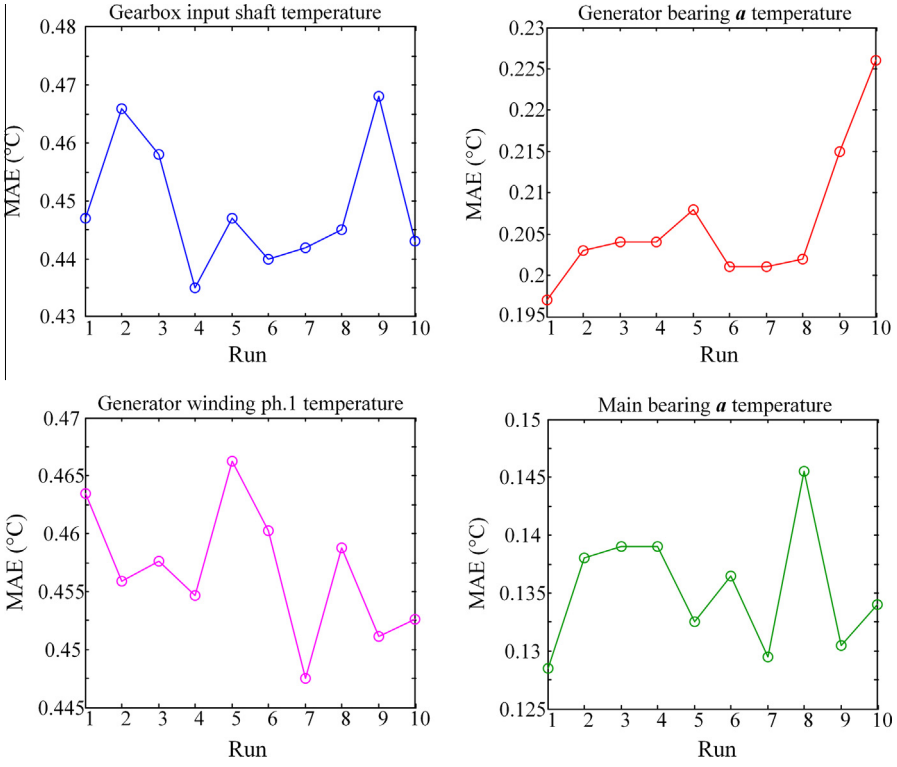


Fig. 6. The BPNN performances over the number of runs.

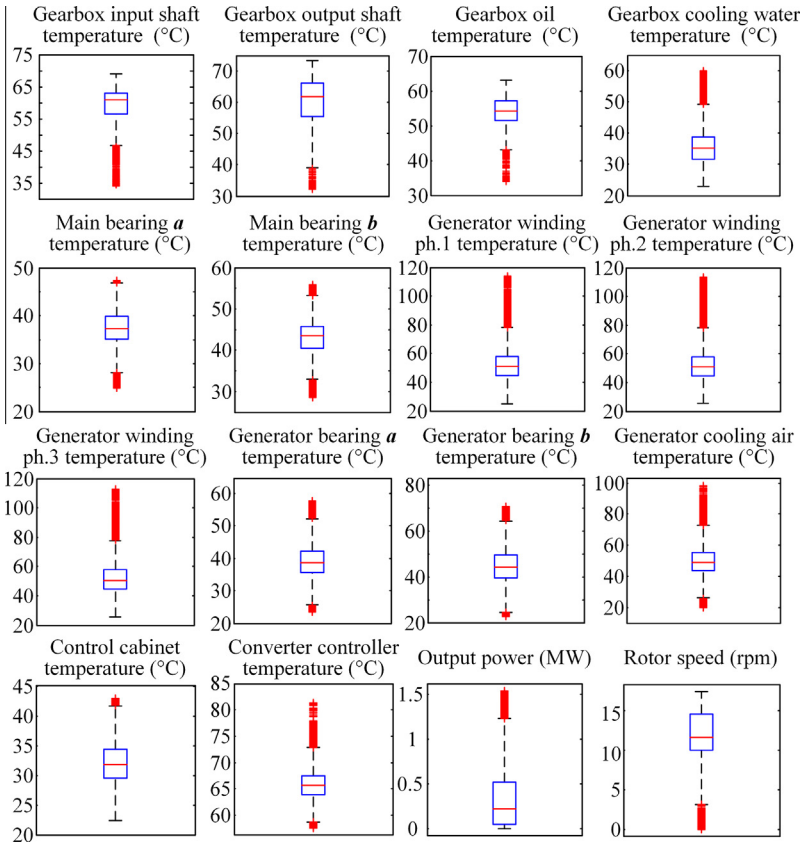


Fig. 7. Boxplot of the 16 condition parameters of the current sample data of WT 9.

Table 3
Prediction performance of component temperatures.

Condition parameters	Index		
	MSE (°C)	MAE (°C)	MAPE (%)
Temp. of gearbox input shaft	0.55	0.38	1.03
Temp. of gearbox output shaft	0.23	0.17	0.35
Temp. of gearbox oil	1.06	0.78	1.55
Temp. of gearbox cooling water	1.34	0.85	1.71
Temp. of main bearing ***a*** (on the rotor side front)	0.16	0.13	0.27
Temp. of main bearing ***b*** (on the gearbox side back)	0.16	0.14	0.29
Temp. of generator winding ph.1	1.93	0.72	1.23
Temp. of generator winding ph.2	1.92	0.71	1.22
Temp. of generator winding ph.3	1.94	0.72	1.25
Temp. of generator bearing ***a*** (front)	0.41	0.17	0.43
Temp. of generator bearing ***b*** (back)	0.45	0.21	0.45
Temp. of generator cooling air	1.52	0.88	1.92
Temp. of control cabinet	0.15	0.12	0.46
Temp. of converter controller	0.71	0.42	1.15

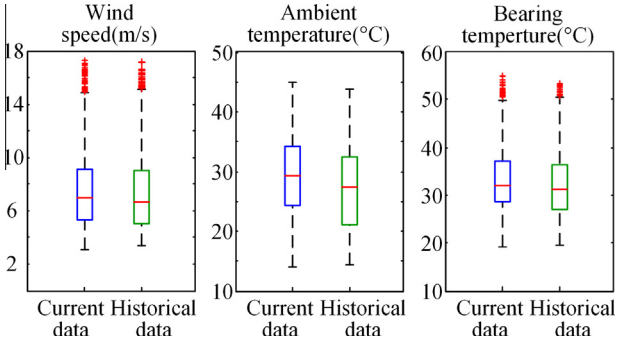


Fig. 8. Boxplot of the current SCADA sample data and historical SCADA sample data of WT 17.

Table 4
Prediction performance of output power and rotor speed.

Condition parameters	Input parameters	Index		
		MSE	MAE	MAPE
Output power	Wind speed; yaw angle error; pitch angle	68.62 kW	39.12 kW	11.52%
	Wind speed	78.2 kW	47.36 kW	17.22%
Rotor speed	Wind speed; yaw angle error; pitch angle	0.37 rpm	0.24 rpm	1.96%
	Wind speed	0.48 rpm	0.36 rpm	2.45%

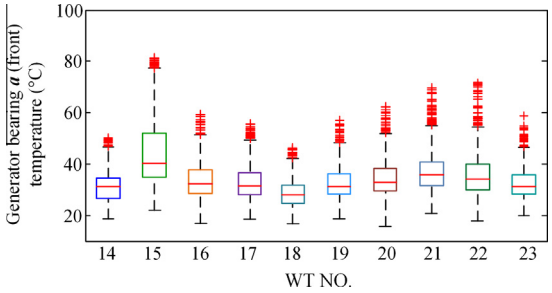


Fig. 9. Boxplot of generator bearing ***a*** (front) temperatures of 10 WTs.

Table 5
Prediction performance of generator bearing ***a*** (front) temperature based on three different types of sample data.

Type	Training sample	Index		
		MSE (°C)	MAE (°C)	MAPE (%)
1	Current data of WT 17	0.36	0.21	0.61
2	Historical data of WT 17	0.60	0.45	1.23
3	Current data of WT 15	2.51	2.22	5.76
	Current data of WT 16	0.45	0.34	0.95
	Current data of WT 18	2.66	1.26	3.22
	Current data of WT 19	0.43	0.31	0.88

temperature of WT 14-WT 23 as an example. It is clear that there are significant differences among the sample distributions of the 10 WTs.

Table 5 shows the prediction accuracy of the prediction models for generator bearing ***a*** (front) temperature of WT 17 trained by the three types of sample data. The prediction model trained by the current SCADA sample data of the WT has the highest accuracy. The accuracy of model trained by historical sample data is obvious lower, which may be caused by the wear and degradation of components. The performance of the four models trained by the current SCADA data of other similar WTs (WT 15, WT 16, WT 18, and WT 19) is different. The prediction accuracy of models trained by the SCADA data of WT 15 and WT 18 is obvious lower because the distributions of the two samples are much different from that of the current data of WT 17 (Fig. 9).

The sample similarity degree (*SSD*) is defined to quantify the degree of similarity between the distributions of the target condition parameters of two WTs. The time series of the target condition parameter of two WTs are obtained for the period which the training sample data belongs to. The normalized correlation coefficient (NCC) is used to calculate the *SSD*.

$$SSD = \frac{\sum_{n=1}^{N} h_1(n) \cdot h_2(n)}{\sqrt{\left(\sum_{n=1}^{N} h_1^2(n)\right) \cdot \left(\sum_{n=1}^{N} h_2^2(n)\right)}} \tag{4}$$

where *SSD* is the sample similarity degree; h_1 and h_2 are the histograms of target condition parameters of two WTs, and N is the number of bins of the histogram. The values of *SSD* is between 0 and 1, with a bigger value indicating a high level of similarity between the two histograms.

The relationship between the *SSD* and the MAE of the prediction models trained by the other similar WTs' SCADA data is studied in this paper. The prediction models of generator bearing ***a*** (front) temperature of the 3 target WTs (i.e. WT 21, WT 22, WT 23) are developed by using the current data of nine similar WTs' SCADA data (i.e. WT 12 to WT 20). The *SSD* between the histograms of generator bearing ***a*** (front) temperature of the three target WTs and the nine similar WTs are calculated according to (4). Fig. 10 shows the histograms of prediction error of prediction models for

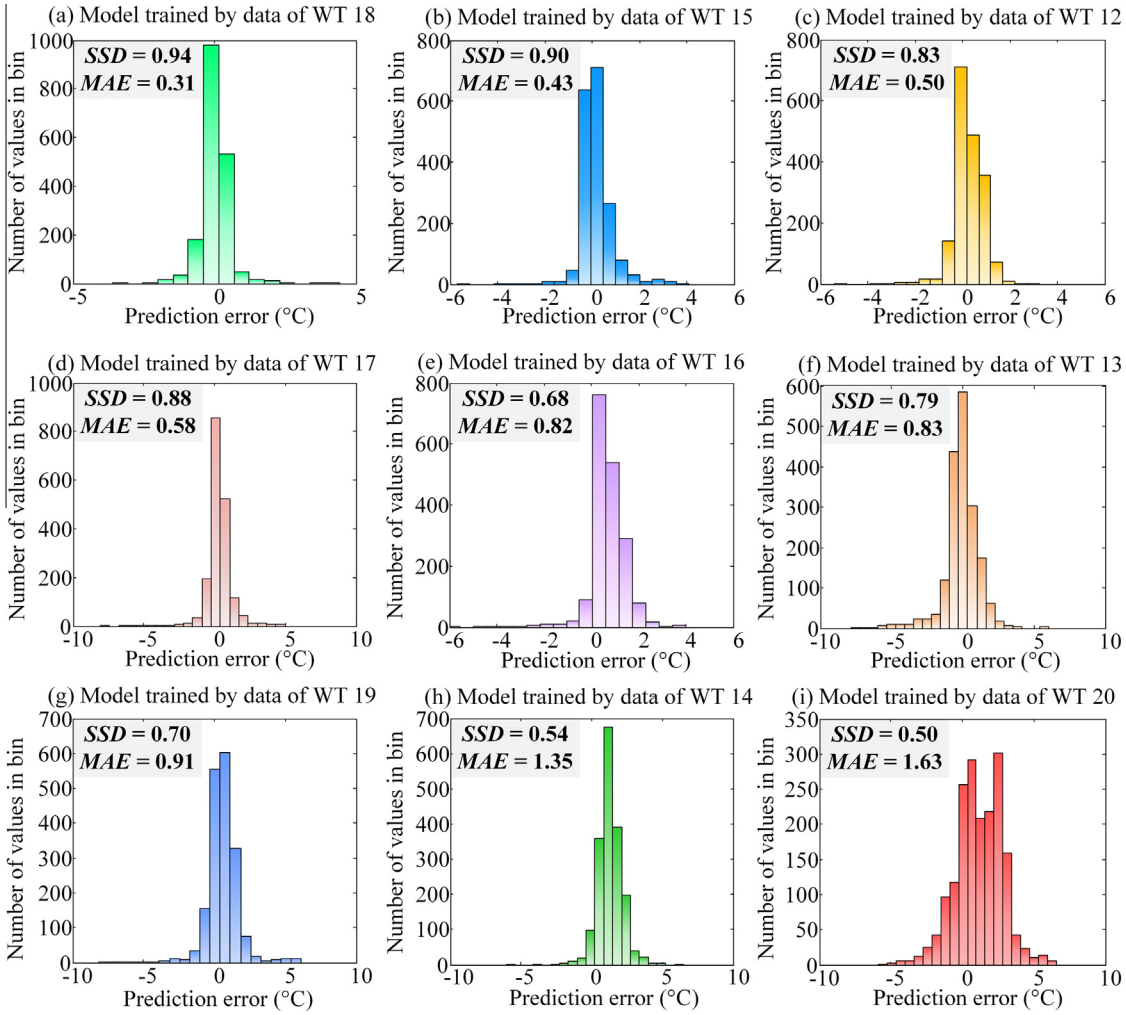


Fig. 10. The histograms of prediction error of prediction models for the generator bearing ***a*** (front) temperature of WT 21 trained by the nine similar WTs' SCADA data.

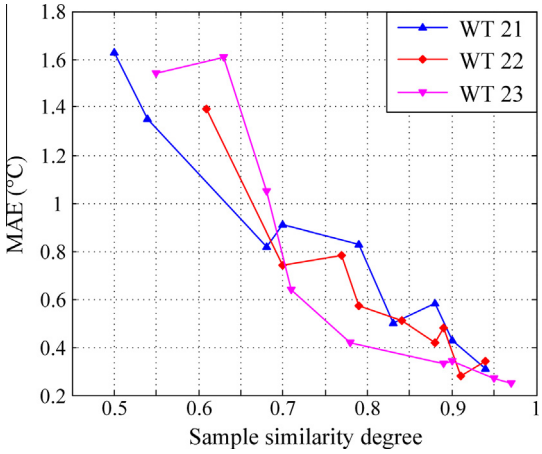


Fig. 11. The relationship between *SSD* and the MAE of prediction models for the generator bearing ***a*** (front) temperature trained by the nine similar WTs' SCADA data.

generator bearing ***a*** (front) temperature of WT 21 trained by the nine similar WTs' SCADA data. The relationship between the *SSD* and the MAE of prediction models for generator bearing ***a*** (front) temperature of the 3 target WTs trained by the nine similar WTs' SCADA data is obtained, as shown in Fig. 11. It can be seen that the MAE will decrease as the *SSD* increases.

5. Prediction model selection

Prediction model selection is necessary because the models' ability in WT anomaly identification has a close relationship with the prediction accuracy. In this research, the prediction model selection method is proposed to select the condition parameter prediction models trained by the WT's historical SCADA data and the other similar WTs' SCADA data. MAE is used to select the prediction models trained by the WT's historical SCADA data. A fast selection method is proposed to increase the selection efficiency for prediction models trained by the other similar WTs' current SCADA data.

The MAE of each model with its corresponding sample data is calculated. For the prediction models trained by the WT's historical SCADA data, if the MAE satisfies the following criterion, the corresponding model is selected.

$$MAE_i \leqslant \lambda MAE_{ref} \quad (5)$$

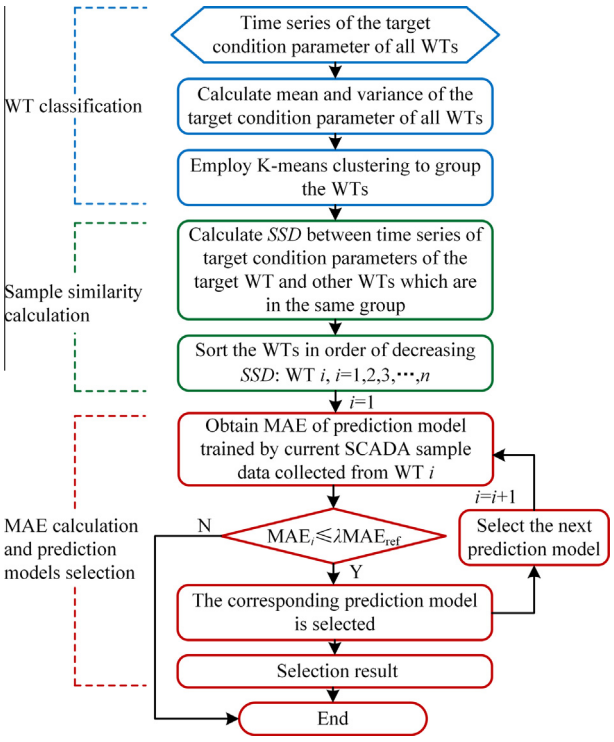


Fig. 12. Flowchart of the selection method for the prediction models trained by other similar WTs' current SCADA data.

where MAE_i is the MAE of *i*th prediction model trained by the WT's historical SCADA data and the other similar WTs' current SCADA data, MAE_{ref} is the MAE of the prediction model trained by the WT's current SCADA data, and λ is the weighting factor.

The SCADA data of 15 random selected WTs are used to select the λ value in (5) among some candidate values (i.e. 1, 2, 3, and 4). Two months of SCADA data from August 1, 2012 to September 30, 2012 are collected when the WTs are under healthy condition, which are used as the current sample data. Three of the 15 WTs are selected as the target WTs and a total of 16 prediction models, such as the model trained by the current sample data, the model trained by the historical sample data, and 14 models trained by the other similar WT's current sample data, are developed for each condition parameter of a target WT. The prediction error probability

Table 6
Average percentage of invalid predictions of the selected prediction models over different λ value.

No.	Condition parameters	MAE_{ref} of the prediction models trained by the current data of the 3 target WTs			The average percentage of invalid predictions (%)			
					$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 4$
1	Temp. of gearbox input shaft	0.44 °C	0.52 °C	0.58 °C	0.46	0.96	5.01	5.99
2	Temp. of gearbox output shaft	0.20 °C	0.26 °C	0.30 °C	0.23	0.81	2.96	3.9
3	Temp. of gearbox oil	0.51 °C	0.68 °C	0.68 °C	0	0.57	3.07	3.07
4	Temp. of gearbox cooling water	0.69 °C	0.66 °C	0.77 °C	0.93	1.73	5.03	6.2
5	Temp. of main bearing ***a*** (on the rotor side front)	0.11 °C	0.11 °C	0.12 °C	0.53	0.76	2.32	3.25
6	Temp. of main bearing ***b*** (on the gearbox side back)	0.13 °C	0.17 °C	0.19 °C	0.46	0.98	1.67	1.67
7	Temp. of generator winding ph.1	0.68 °C	0.53 °C	0.42 °C	0.69	1.03	5.66	5.66
8	Temp. of generator winding ph.2	0.72 °C	0.64 °C	0.45 °C	0.83	1.38	4.51	4.51
9	Temp. of generator winding ph.3	0.66 °C	0.59 °C	0.51 °C	0.69	1.36	4.32	4.97
10	Temp. of generator bearing ***a*** (front)	0.19 °C	0.21 °C	0.19 °C	0.87	1.18	4.85	6.93
11	Temp. of generator bearing ***b*** (back)	0.28 °C	0.31 °C	0.35 °C	0.93	1.67	3.19	3.81
12	Temp. of generator cooling air	0.63 °C	0.47 °C	0.52 °C	1.39	1.62	3.22	5.33
13	Temp. of control cabinet	0.14 °C	0.14 °C	0.16 °C	1.04	1.83	1.83	1.83
14	Temp. of converter controller	0.37 °C	0.45 °C	0.49 °C	0.93	1.32	5.77	7.3
15	Output power	31.09 kW	37.25 kW	34.63 kW	0.69	1.25	3.91	6.29
16	Rotor speed	0.32 rpm	0.44 rpm	0.48 rpm	1.16	1.76	4.98	4.98

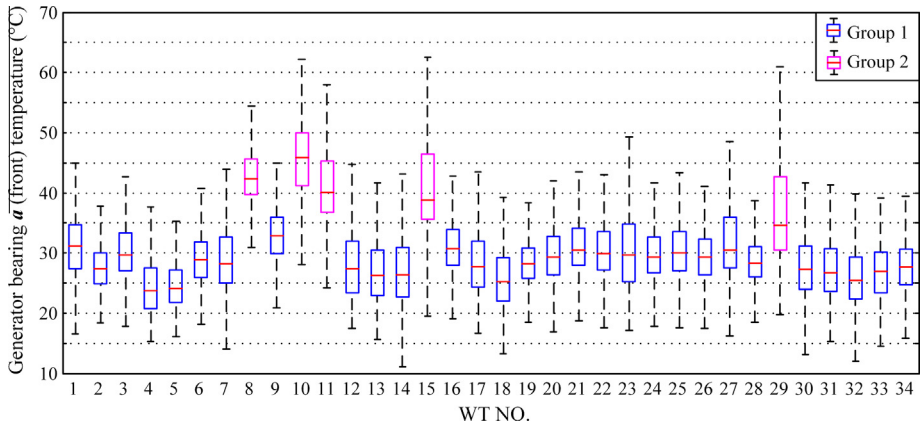


Fig. 13. The WT classification results for generator bearing ***a*** (front) temperature of 34 WTs in the studied wind farm.

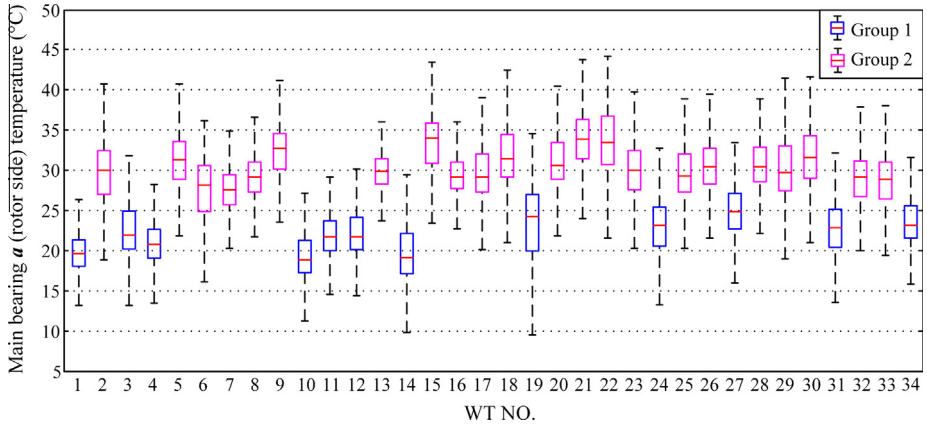


Fig. 14. The WT classification results for main bearing ***a*** (on the rotor side) temperature of 34 WTs in the studied wind farm.

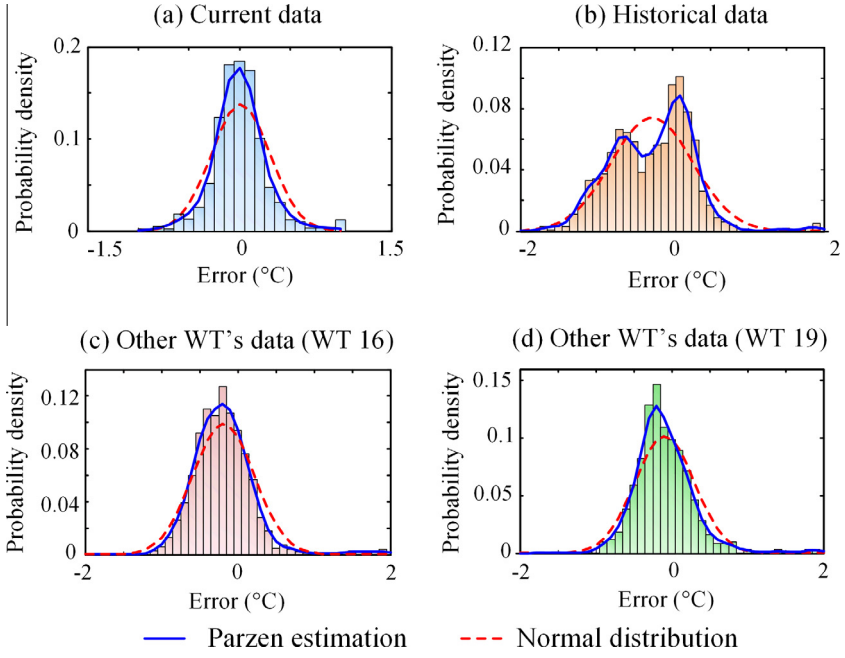


Fig. 15. Normalized histogram and two PDFs of prediction error for prediction models in Table 5.

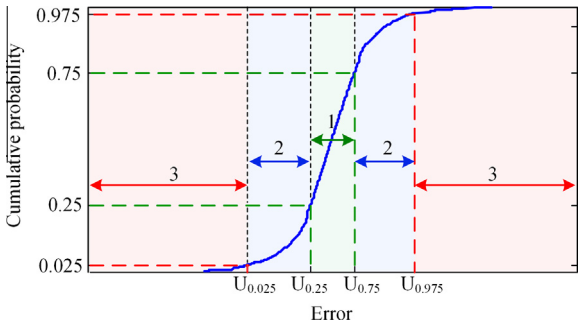


Fig. 16. Three regions of prediction error.

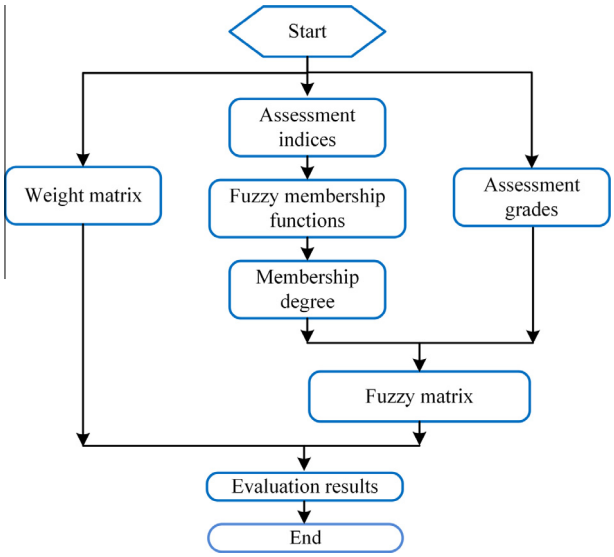


Fig. 17. Flowchart of the general fuzzy synthetic evaluation method.

distributions and the MAE of each prediction model are obtained by using the test data extracted from the target WT's current SCADA data. The prediction models trained by the historical data and the other similar WTs' data are selected based on (5).

The selected models are used to predict the 16 condition parameters of the 3 target WTs on October 1, 2012 (the 3 target WTs are healthy on that day). Based on the prediction error probability distributions obtained before, if the prediction error is not within its 97.5% confidence interval, the prediction is considered invalid. These invalid predictions may cause the misdiagnosis of WT anomalies according to the principle of the statistical anomaly detection techniques [41]. Table 6 shows the average percentage of invalid predictions of the prediction models that are selected over different λ values. The percentage of invalid predictions is very low

Table 8
The membership degrees of the discretized ALIs related to the three assessment grades.

No.	ALI range	Membership degrees related to the assessment grades		
		Normal	Moderate	Abnormal
1	0–0.1	1 (34/34)	0 (0/34)	0 (0/34)
2	0.1–0.2	0.98 (56/57)	0.02 (1/57)	0 (0/57)
3	0.2–0.3	0.97 (73/75)	0.03 (2/75)	0 (0/75)
4	0.3–0.4	0.98 (49/50)	0.02 (1/50)	0 (0/50)
5	0.4–0.5	0.58 (26/45)	0.36 (16/45)	0.07 (3/45)
6	0.5–0.6	0.1 (6/52)	0.69 (36/52)	0.21 (11/52)
7	0.6–0.7	0 (0/46)	0.41 (19/46)	0.59 (27/46)
8	0.7–0.8	0 (0/55)	0.07 (4/55)	0.93 (51/55)
9	0.8–0.9	0 (0/32)	0 (0//32)	1 (32/32)
10	0.9–1	0 (0/48)	0.02 (1/48)	0.98 (47/48)

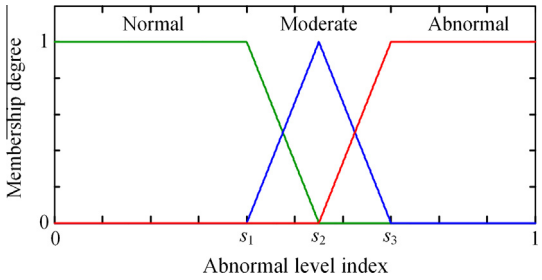


Fig. 18. Membership functions of ALI.

when λ is 1, but few prediction models trained by the historical SCADA data and the other similar WT's current SCADA data can be selected according to (5). When λ is 2, the maximum of the average percentage of invalid predictions is lower than 2%. However, when λ is 3 and 4, the percentage of invalid prediction increase significantly for the majority of the condition parameters. Therefore, in order to have less invalid predictions, λ is chosen as 2 in this study.

A wind farm usually has dozens to hundreds of WTs. When selecting prediction models trained by other similar WTs' current SCADA data, it is difficult to perform the model selection on each condition parameter of each WT by using (5) directly. It is necessary to improve the selection efficiency. According to the discussion in Section 4, the prediction accuracy of the models trained by the sample data whose distributions of the target condition parameter are obvious different from that of the target WT is much lower. These models with bad performance can be directly removed. Based on this rule, the flowchart of selection method for the condition parameter prediction models of a target WT trained by other similar WTs' current SCADA data is established, shown in Fig. 12. The procedure can be summarized into the three steps as follows:

Table 7
The WT fault cases and the corresponding number of calculated ALIs for the prediction errors of the condition parameters.

No.	WT fault cases	Target condition parameters	Number of calculated ALIs
1	Converter fan malfunction of WT 6	Converter controller temperature	50
2	Over-speed of WT 9	Rotor speed	52
3	Gearbox output shaft overheating of WT 9	Gearbox output shaft temperature	46
4	Generator bearing overheating of WT 14	Generator bearing ***b*** temperature	37
5	Carbon brush oxidation of WT 18	Generator bearing ***a*** temperature	62
6	Gearbox input shaft overheating of WT 18	Gearbox output shaft temperature	53
7	Gearbox oil over-temperature of WT 20	Generator oil temperature	71
8	Control system fault of WT 24	Output power	74
9	Generator bearing overheating of WT 26	Generator bearing ***b*** temperature	49

5.1. WT classification

The WT classification aims to find the WTs whose sample distribution of the condition parameter is obvious different from that of the target WT. The time series of the target condition parameter are obtained for all the WTs during the period which the training sample data belongs to. The mean and variance of the condition parameters are calculated and a *k*-means clustering method, which is one of the most widely used unsupervised learning algorithms to solve the clustering problem, is used to group the WTs. Figs. 13 and 14 show two examples of the boxplots and the classification results of two condition parameters (i.e. generator bearing ***a*** (front) temperature and main bearing ***a*** (on the rotor side) temperature) of the 34 WTs in the studied wind farm. It can be seen that the sample distribution of the target condition parameter in the same group is similar.

5.2. Sample similarity calculation

In each group, the histograms of each time series of target condition parameter are obtained. The *SSD* between each histogram is calculated according to (4).

5.3. MAE calculation and prediction models selection

For a target condition parameter of one certain WT *i*, the prediction model trained by the WT's current SCADA data is established and the MAE_{ref} is obtained. The selection is only performed on the prediction models trained by the current SCADA of the other WTs which are in the same group with WT *i*. Following the *SSD* decreasing order, the MAE of the models trained by these similar WTs' current SCADA data is obtained and used to select the prediction models. If the MAE satisfies (4), the corresponding model is selected.

6. Abnormal level quantification

Parzen estimation, a commonly used nonparametric estimation method, is used to describe the prediction error distributions of condition parameters. Probability density functions (PDFs) of the parameter values can be expressed as follows [53]:

$$P(x) = \frac{1}{hn}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{6}$$

where $\{x_1, \ldots, x_n\}$ is a series of sample data, K is the kernel function, and h is the window width. The Gaussian kernel is used and the approximate mean integrated square error (AMISE) based approach is applied to obtain the optimal window width [54].

The normalized histogram and Parzen estimation based PDF of prediction error for the prediction models in Table 5 trained by the WT's current SCADA data, the WT's historical SCADA data and the other similar WTs' current SCADA data (WT 16 and WT 19) are given as an example, as shown in Fig. 15. In the previous studies, the normal distribution based PDF is generally used to describe the prediction error distributions of WT condition parameters [39,40,52]. The normal distribution PDF is also shown in Fig. 15 for comparison to the Parzen based PDF. It is obvious that the prediction error distribution of model trained by historical sample data cannot be considered as normally distributed. Parzen estimation based PDF is more suitable than normal distribution based PDF for the description of prediction error distributions of condition parameters.

Abnormal level index (ALI) is defined to quantify the abnormal level of prediction errors. The calculation of the ALI is based on the principle that normal (no-fault) data occur in high-probability regions and abnormal conditions occur in low-probability regions. The prediction errors are divided into three regions according to the quantile 0.025, 0.25, 0.75 and 0.975, as shown in Fig. 16. A fixed-size moving time window, which contains the data in 24 h, is applied to obtain the error residual series. The moving step of the time window is set to be 2 h and it can be adjusted if required. ALI is calculated as follows:

$$\mathrm{ALI} = 1 - \frac{N_1 C_1}{\sum_{k=1}^{3} N_k C_k} \tag{7}$$

where N_k is the number of data in the kth probability region and C_k ($k = 1, 2, 3$) is the corresponding penalty factor, which is set to be {1, 3, 5}, respectively. ALI has a value in the range between 0 and 1, with a bigger value indicating a higher level of abnormal condition.

7. Fuzzy synthetic evaluation

Although each prediction model can be used for anomaly identification, the sensitivity and the accuracy is different due to the training sample data. The fuzzy synthetic evaluation, which has been widely used for fuzzy-based decision making [55–57], is employed in this research to integrate the results of each outcome of the anomaly identification. Fig. 17 shows the flowchart of the general synthetic evaluation, which primarily consists of the weight matrix calculation of the different layer indices, fuzzy membership functions, assessment grades, fuzzy matrix calculations and defuzzification.

The index in the index layer is the ALI obtained from each prediction model. The distinct assessment grades are defined as $L = \{l_1, l_2, l_3\}$ = {Normal, Moderate, Abnormal}.

The fuzzy assessing matrix could give the relation between the assessing parameter and assessing result; it is the key part when making a fuzzy assessment [55]. The fuzzy matrix V is established by substituting the ALIs for prediction errors of each selected prediction model into the membership functions. Fuzzy matrix can be expressed as follows:

$$V = \begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_n \end{bmatrix} = \begin{bmatrix} v_{11} & v_{12} & v_{13} \\ v_{21} & v_{22} & v_{23} \\ \vdots & \vdots & \vdots \\ v_{n1} & v_{n2} & v_{n3} \end{bmatrix} \tag{8}$$

where v_{ij} ($j = 1, 2, 3$) is the membership degree.

In order to develop the membership functions of ALI, the fuzzy statistics method is utilized [58]. A total of 494 ALIs for the prediction errors of the selected prediction models are calculated by using the SCADA data collected from 9 WT fault cases, as shown in Table 7. The ALIs are discretized into 10 ranges (i.e. 0–0.1, 0.1–0.2, …, 0.9–1). The membership degrees of the discretized ALIs related to each assessment grade then can be characterized by:

$$F_i = \frac{N_k^{(i)}}{N_k} \tag{9}$$

where N_k is the number of the ALI data in the kth ALI range and $N_k^{(i)}$ is the number of the ALI data related to the ith assessment grade (i.e. normal, moderate and abnormal).

Table 8 shows the statistical results of the possibilities of the discrete ranges of ALI related to the three assessment grades. Based on the statistical results, the membership functions with trapezoidal and triangular shapes are used in this research, where $s_1 = 0.4$, $s_2 = 0.55$, $s_3 = 0.7$, as shown in Fig. 18. A linear variation of the membership values is reasonable as long as no other dependency is known. Publication [59] proposed a diagnosis method for detecting stator winding faults in induction motors based on fuzzy logic. It was shown that the combination of triangular and trape-

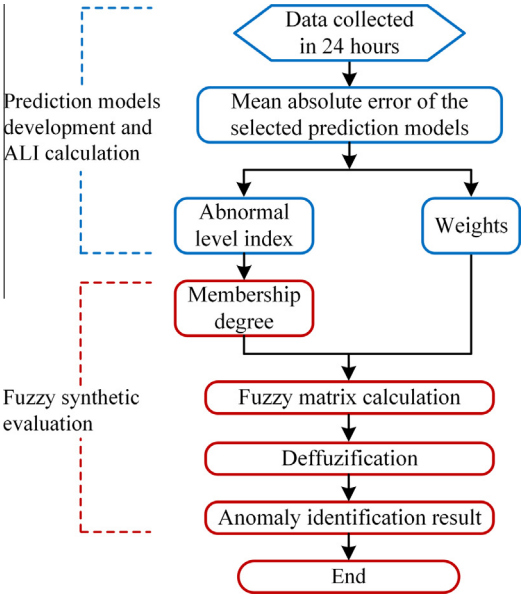


Fig. 19. Flowchart of the anomaly identification method.

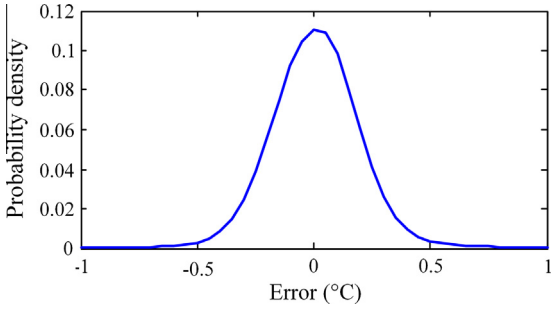
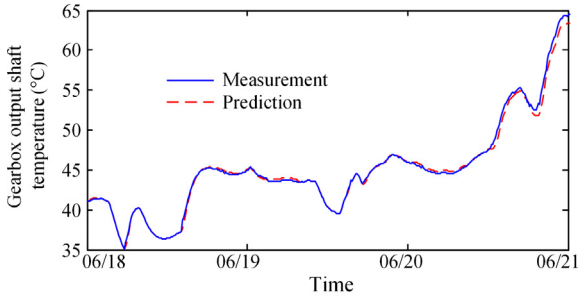
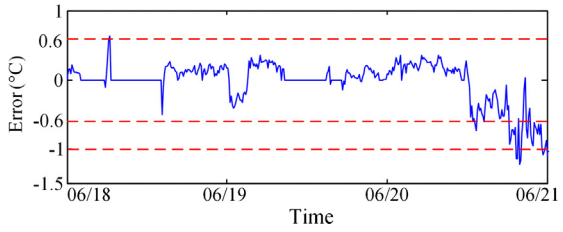


Fig. 20. Prediction error PDF for prediction model of gearbox output shaft temperature.



(a) Measured and predicted values



(b) Prediction error

Fig. 21. Prediction result of gearbox output shaft temperature (WT 24) during June 18 to June 20, 2011.

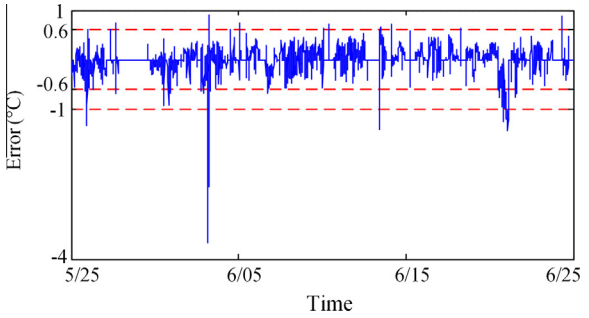


Fig. 22. Prediction error of gearbox output shaft temperature of WT 24 during May 25 to June 25, 2011.

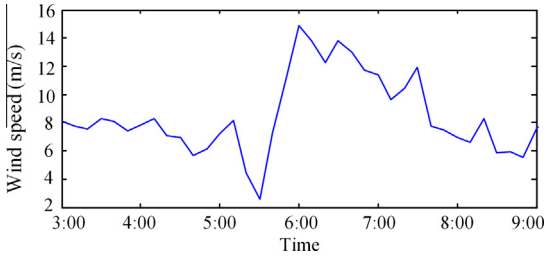
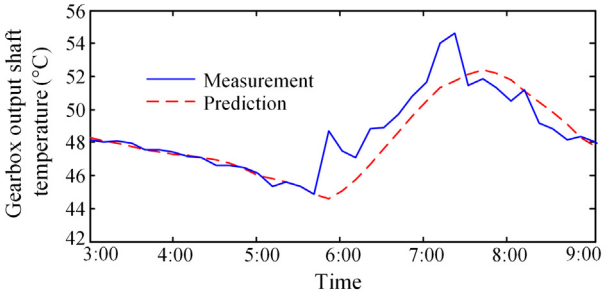
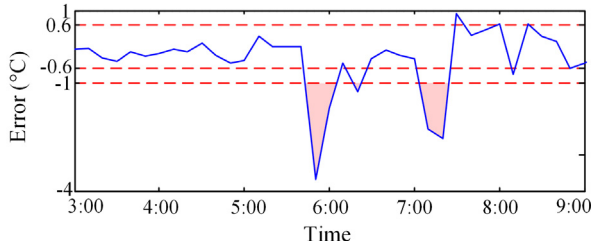


Fig. 23. Time series of wind speed in June 4, 2011.



(a) Measured and predicted values



(b) Prediction error

Fig. 24. Prediction result of gearbox output shaft temperature of WT 24 in June 4, 2011.

zoidal membership functions is the most appropriate for fault diagnosis in induction motors [59].

With the ALI as basis, the value of membership degree related to the three condition assessment grades can be calculated as follows:

$$v_{i1} = \begin{cases} 1 & x_i < 0.4 \\ 3.67 - 6.67x_i & 0.4 \leqslant x_i < 0.55 \\ 0 & x_i \geqslant 0.55 \end{cases} \quad (10)$$

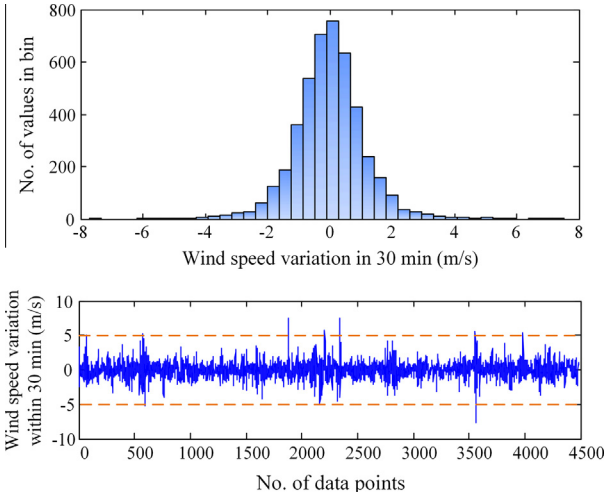


Fig. 25. The distribution of wind speed variation within 30 min during the period which the training sample data belongs to.

$$v_{i2} = \begin{cases} 0 & x_i < 0.4 \text{ or } x_i \geqslant 0.7 \\ 6.67x_i - 2.67 & 0.4 \leqslant x_i < 0.55 \\ 4.67 - 6.67x_i & 0.55 \leqslant x_i < 0.7 \end{cases} \quad (11)$$

$$v_{i3} = \begin{cases} 0 & x_i < 0.55 \\ 6.67x_i - 3.67 & 0.55 \leqslant x_i < 0.7 \\ 1 & x_i \geqslant 0.7 \end{cases} \quad (12)$$

where x_i is the ALI for prediction error of the ith condition parameter prediction model.

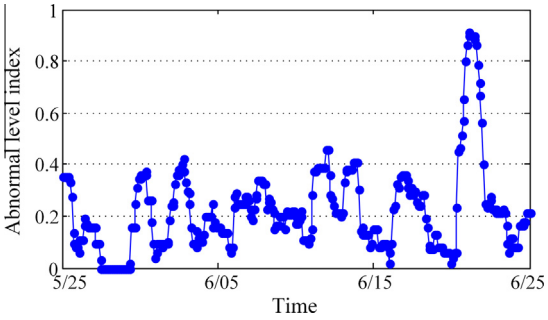


Fig. 27. The ALIs of prediction error of gearbox output shaft temperature of WT 24 from May 25 to June 25, 2011.

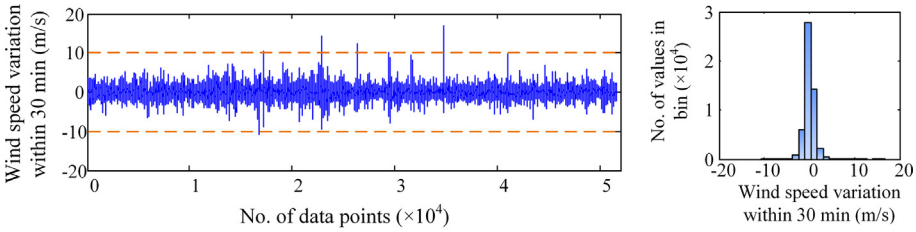
The weight of each model should be related to the prediction accuracy. In this study, the ith prediction model's weight is defined as follows:

$$\omega_i = \left(1 - \frac{r_i}{\sum_{i=1}^{N} r_i}\right) \Big/ \sum_{i=1}^{N} \left(1 - \frac{r_i}{\sum_{i=1}^{N} r_i}\right) = \left(1 - \frac{r_i}{\sum_{i=1}^{N} r_i}\right) \Big/ (N-1) \quad (13)$$

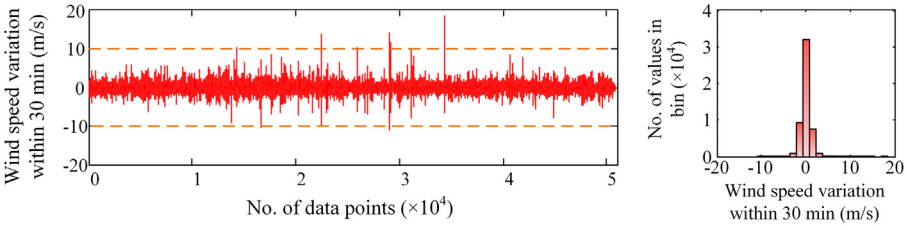
where r_i is the MAE of the ith selected prediction model and N is the number of the total selected prediction models.

Following the discussion in the previous sections, a procedure for dealing with the WT anomaly identification is established, as shown in Fig. 19. The procedure can be summarized into the following two steps:

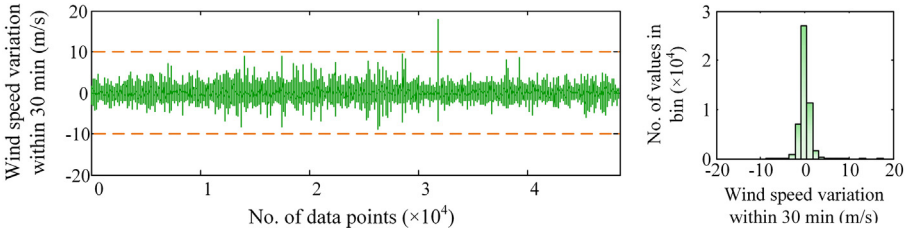
Step 1 (Prediction model development and ALI calculation): The BPNN based condition parameter prediction models are trained



(a) The distribution of wind speed variation within 30 min of WT 24



(b) The distribution of wind speed variation within 30 min of WT 5



(c) The distribution of wind speed variation within 30 min of WT 12

Fig. 26. The distribution of wind speed variation within 30 min of 3 WTs in the studied wind farm from February 15, 2011 to February 15, 2012.

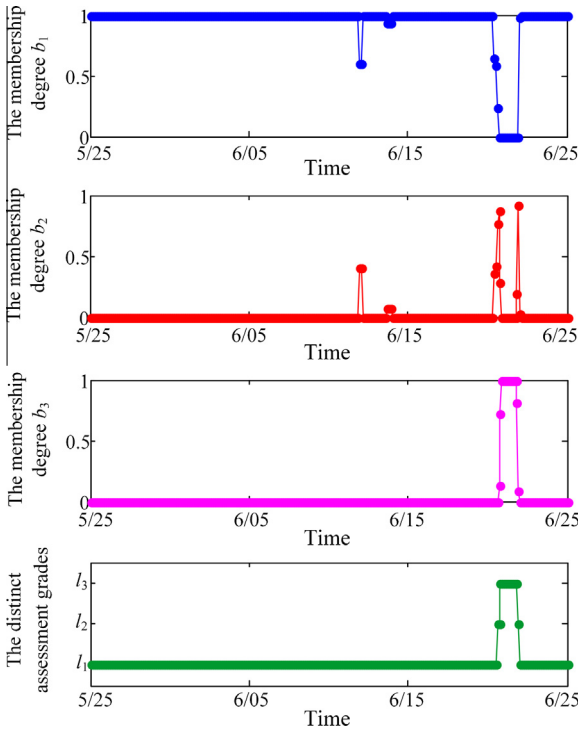


Fig. 28. The anomaly identification results for gearbox output shaft temperature of WT 24 from May 25 to June 25, 2011.

by the WT's current SCADA data, the WT's historical SCADA data, and the other similar WTs' current SCADA data, respectively. The MAE of each model is obtained with the testing sample data. The prediction models are selected according to what was discussed in Section 5. The weight ω_i can be calculated by using (13). A fixed-size moving window, which contains the data in 24 h, is used to obtain the error residuals. The ALI of each model can be calculated based on (7).

Step 2 (Fuzzy synthetic evaluation): According to (10)–(12), the membership degrees of each ALI for different conditions can be obtained. The membership degree of the outcome in the objective layer can be calculated as follows:

$$B = [b_1 \quad b_2 \quad b_3] = [\omega_1 \quad \omega_2 \quad \cdots \quad \omega_n] \begin{bmatrix} v_{11} & v_{12} & v_{13} \\ v_{21} & v_{22} & v_{23} \\ \vdots & \vdots & \vdots \\ v_{n1} & v_{n2} & v_{n3} \end{bmatrix} \tag{14}$$

The maximum membership principle of the defuzzification is used. Therefore, the calculated l_i $(i = 1, 2, 3)$ corresponds to the $b_{\max} = \max(b_i)$.

8. Case study and verification

Two cases are selected to validate the proposed anomaly identification method. In case 1, the proposed method is compared with the traditional threshold method. Generally, the threshold value depends on the confidence interval of prediction error. The upper and low bounds of 97.5% and 99% confidence interval are used as the thresholds. The proposed method is compared with a single-model-based method in case 2. Finally, the statistical results are given to validate the generalization accuracy of the proposed WT anomaly identification method.

8.1. Case 1

WT 24 suffered a sudden breakdown on June 25, 2011 due to the overheating of the gearbox output shaft temperature. According to the maintenance records, this WT has been in operation for two months before the fault.

The process of anomaly identification of the gearbox output shaft temperature is given in this case study. Because the SCADA data collected since February 15, 2011 are available, the prediction model trained by historical data is not considered. According to the criteria of the prediction model selection (Section 5), only the models trained by the current SCADA data can be used. Fig. 20 shows the prediction error PDF for the prediction model of gearbox output shaft temperature. The model's MAE with the testing sample data is 0.12 °C. The upper and low bounds of 97.5% confidence interval is ±0.6 °C. The upper and low bounds of 99% confidence interval is ±1 °C.

Fig. 21 shows the prediction results during June 18 to June 20, 2011. According to the proposed method of pre-processing, the prediction results when output power is less than 100 kW are not included by setting the corresponding prediction errors to zero. It is noted that the prediction value is lower than the measured value in the later part of June 20. The ALI over the time interval from 0:00 am, June 20 to 24:00 pm, June 20, 2011 is calculated to be 0.91 and the membership degree of the outcome in the objective layer is [0,0,1]. According to the maximum membership principle, this condition parameter is considered as abnormal.

As seen in Fig. 21(b), the abnormal condition can also be detected by a threshold-based method. However, the threshold-based method can also cause misdiagnosis. Fig. 22 shows the prediction error during May 25 to June 25, 2011. Misdiagnosis occurs several times when using the threshold-based method, which can be analyzed by using the data of June 4. The wind speed increased very fast in a short time on that day, as shown in Fig. 23. The wind speed variation within a 30 min interval between 5:30 am and 6:00 am on June 4, 2011, is 12.3 m/s. Fig. 24 shows the prediction error during a period of six hours, i.e., from 3:00 am to 9:00 am. The result shows that the rapid increase of wind speed is the root cause of the significant increase in the prediction error. Fig. 25 presents the distribution of wind speed variations within 30 min during the period which the training sample of the prediction model of gearbox output shaft temperature of WT 24 belongs to. It can be seen that the wind speed variations within 30 min are mostly distributed in the range of −5 m/s to 5 m/s. The wind speed variations within 30 min of WT 24 and two other similar WTs (i.e. WT 5 and WT 12) from February 15, 2011 to February 15, 2012 are shown in

Table 9
The distribution of the prediction error of gearbox output shaft temperature.

Three regions of prediction error	The number of data
Region 1	120
Region 2	16
Region 3	8
ALI	0.42

Table 10
Prediction performance of generator bearing ***a*** (front) of four prediction models.

Model	Training sample	Index		
		MSE (°C)	MAE (°C)	MAPE (%)
1	Current data of WT 22	0.41	0.17	0.43
2	Historical data of WT 22	0.41	0.18	0.47
3	Current data of WT 20	0.50	0.29	0.71
4	Current data of WT 25	0.53	0.33	0.82

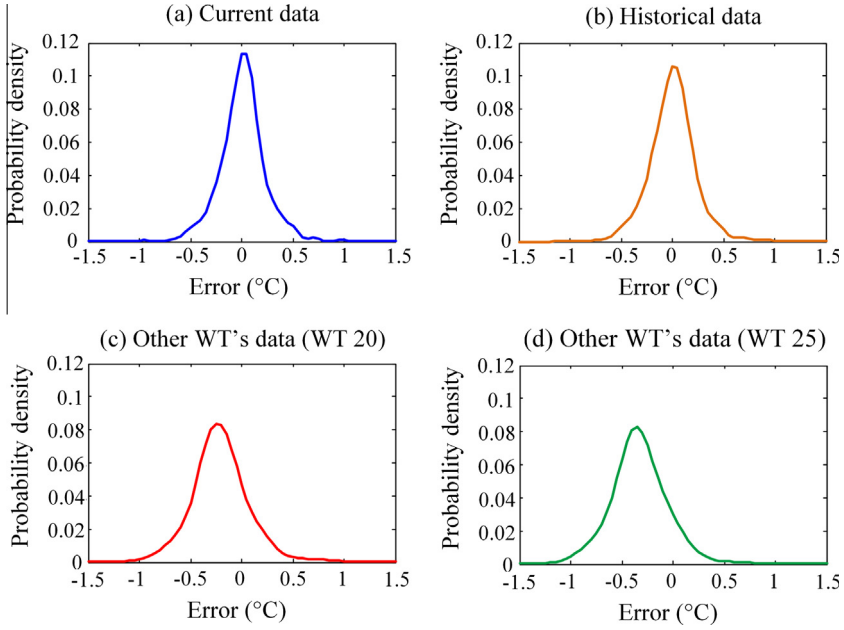


Fig. 29. Prediction error PDFs for prediction models in Table 10.

Fig. 26. Few wind speed variations are larger than 10 m/s in a year. The prediction models could hardly provide accurate predictions of condition parameters when the wind speed increases very fast due to the lack of training data. It is unreasonable to consider gearbox output shaft temperature abnormal by using the threshold-based method.

By using the proposed method, the ALIs from May 25 to June 25, 2011 are calculated, as shown in Fig. 27. The anomaly identification results for gearbox output shaft temperature of WT 24 from May 25 to June 25, 2011 are shown in Fig. 28. Table 9 shows the distribution of the prediction error of gearbox output shaft temperature of WT 24 from 1:00 am, June 4 to 1:00 am, June 5, 2011. Although the prediction errors significantly increase during the period of fast change in wind speed, most of the errors are still distributed in the high-probability region (i.e. region 1). According to (7), the ALI over the time interval from 1:00 am, June 4 to 1:00 am, June 5, 2011 is calculated to be 0.42 and the condition parameter can be considered normal. Therefore, the proposed method is more effective than the traditional threshold method.

8.2. Case 2

WT 22 suffered a sudden breakdown on June 11, 2012 due to the severe oxidation of its carbon brush.

The process of anomaly identification of the temperature of generator bearing ***a*** (front) is given in this case study. According to the prediction model selection result, four prediction models, including the model trained by the current SCADA data of the WT, the model trained by the historical data of the WT, the model trained by the SCADA data of WT 20 (a similar WT), and the model trained by the SCADA data of WT 25 (another similar WT), can be used to predict the temperature of generator bearing ***a*** (front). The prediction performance of the four prediction models is shown in Table 10. The prediction error PDFs for the prediction models in Table 10 are shown in Fig. 29.

The prediction error profiles between June 7 and June 10, 2012 are shown in Fig. 29. The anomalies can be observed from the prediction result of each model. It can be seen from Fig. 30 that Model 1 is less effective in detecting abnormal conditions. The reason may be that the current SCADA data (in the previous 60 days before

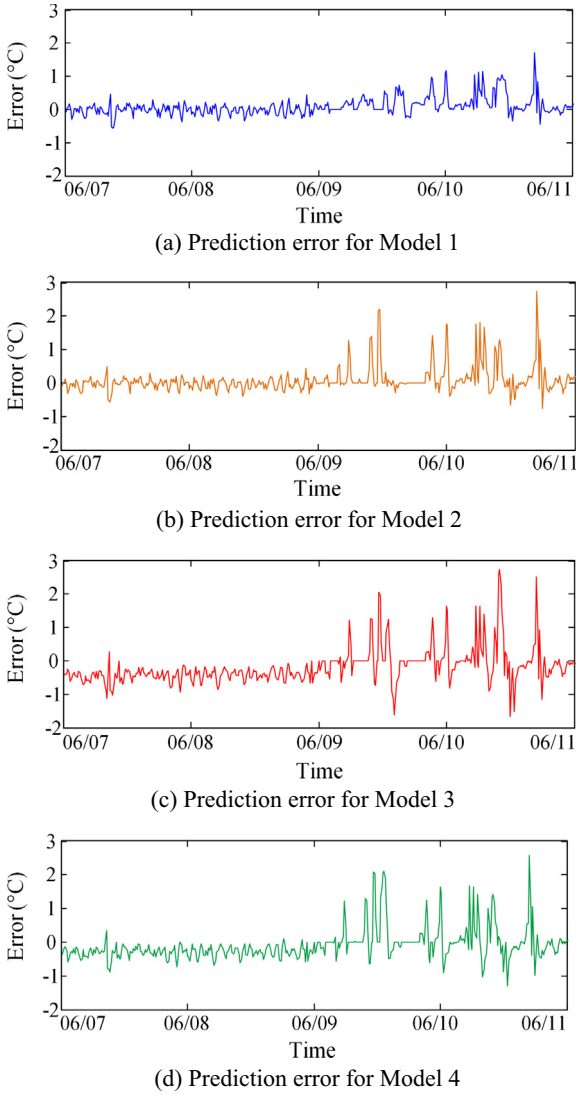


Fig. 30. Prediction error obtained by four prediction models.

Table 11
ALI of four prediction models.

Anomaly identification indices		Model 1	Model 2	Model 3	Model 4
Quantile	0.025	−0.48	−0.48	−0.75	−0.87
	0.25	−0.11	−0.13	−0.35	−0.48
	0.75	0.12	0.13	−0.07	−0.17
	0.975	0.49	0.47	0.36	0.24
Statistics distribution	Region 1	110	80	25	23
	Region 2	12	31	76	81
	Region 3	22	33	43	40
ALI		0.57	0.76	0.95	0.95

Table 12
Anomaly identification results of WTs.

Statistical results	Sample condition		
	Normal	Moderate	Abnormal
Number of sample groups	36	43	49
Correct identification results	35	39	45
Accuracy (%)	97.22	90.7	91.84
Total accuracy (%)	93.25		

June 11) used for training Model 1 have been impacted by the problem that existed before and caused the breakdown of the WT on June 11, 2012.

The three regions of prediction error are divided according to Fig. 16. The statistics distributions of prediction error for the four models are obtained. The ALI of each model over the time interval from 18:00 pm, June 9, 2012 to 18:00 pm, June 10, 2012 is calculated according to (7), as shown in Table 11. The membership degree in the index layer can be calculated by (10)–(12) and the fuzzy matrix is obtained as follows:

$$V = \begin{bmatrix} 0 & 0.87 & 0.13 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

According to (13), the model's weight factors are {0.28, 0.27, 0.23, 0.22}. The membership degrees in the objective layer are obtained by using (14), which is {0, 0.24, 0.76}. According to the maximum membership principle, this condition of the WT is considered abnormal. However, when only the model trained by the current SCADA data (i.e. Model 1) is used, the identification result would be "Moderate". Therefore, the proposed method by combing the results of different models is more accurate than a single-model-based method.

8.3. Expanded cases for verification

Another 128 groups of anomaly identification samples (each group contains the SCADA data in 24 h) are selected from 14 WT fault cases. The anomaly identification procedure is then implemented to further validate the proposed method, and the results are shown in Table 12. The total accuracy of the proposed anomaly identification method is 93.25%, which demonstrates that the effectiveness of the anomaly identification method.

9. Conclusions

A generalized model for anomaly identification of a WT based on rotor speed, output power, and component temperature was presented in this paper. Different WT condition parameter prediction models have been developed and trained by three types of data, namely the current SCADA data of the WT, its historical SCADA data and other similar WTs' current SCADA data. The prediction performance of the three types of models has been compared. The prediction model trained by the current SCADA data has the highest accuracy while the accuracy of the models trained by historical sample data is lower. The accuracy of the models trained by the other similar WTs' current SCADA data varies depending on the training sample distributions. The WT condition parameter prediction models trained by different types of sample data are selected by their performance of prediction accuracy.

A new index called ALI is proposed to quantify the abnormal level of WT condition. Fuzzy synthetic evaluation is used to integrate the results obtained from the selected WT condition parameter prediction models. Two case studies for an onshore wind farm in Northern China have been carried out and analyzed. The results show that the proposed method is more effective in anomaly identification of WTs than traditional methods such as a threshold based method and a single-model-based method.

Acknowledgments

This work was supported in part by National Key Basic Research Program of China (973 Program) (2012CB215205), the Specialized Research Fund for the Doctoral Program of Higher Education of China (SRFDP) (20110191130004), the visiting scholar fund of State Key Lab of Power Transmission Equipment and System Security (SKLPES), Chongqing University, China, the National Natural Science Foundation of China (No. 51321063), and the 111 Project of the Ministry of Education, China (B08036). The work of C. Wang was partially supported by the National Science Foundation of USA under Grant ECCS-1202133.

References

[1] Chehouri A, Younes R. Review of performance optimization techniques applied to wind turbines. Appl Energy 2015;142:361–88.
[2] Gil MDP, Gomis-bellmunt O, Sumper A. Technical and economic assessment of offshore wind power plants based on variable frequency operation of clusters with a single power converter. Appl Energy 2014;125:218–29.
[3] Yang WX, Tavner PJ, Crabtree CJ, Feng Y, Qiu Y. Wind turbine condition monitoring: technical and commercial challenges. Wind Energy 2014;17:673–93.
[4] Johan R, Margareta BL. Survey of failures in wind Power Systems with focus on Swedish wind power plants during 1997–2005. IEEE Trans Energy Convers 2007;22:167–73.
[5] Milborrow D. Operation and maintenance costs compared and revealed. Wind Stats 2006;19(3):3.
[6] Caselitz P, Bussel GW, Spinato F. Rotor condition monitoring for improved operational safety of offshore wind energy converters. J Solar Energy Eng 2005;127:253–61.
[7] Soua S, Lieshout PV, Perera A, Gan TH, Bridge B. Determination of the combined vibrational and acoustic emission signature of a wind turbine gearbox and generator shaft in service as a pre-requisite for effective condition monitoring. Renew Energy 2013;51:175–81.
[8] Becker E, Posta P. Keeping the blades tunning: condition monitoring of wind turbine gearbox. Refocus 2006;7:26–32.
[9] Verbruggen TW. Wind turbine operation and maintenance based on condition monitoring, WT-Ω. The Netherlands: Final Report Energy Research Center; 2003. p. 1–9 [ECN-C-03-047].
[10] Song Z, Jiang Y, Zhang ZJ. Short-term wind speed forecasting with Markov-switching model. Appl Energy 2014;130:103–12.
[11] Zhang WY, Wu J, Wang JZ, Zhao WG, Shen L. Performance analysis of four modified approaches for wind speed forecasting. Appl Energy 2012;99:324–33.
[12] Chen KL, Yu J. Short-term wind speed prediction using an unscented Kalman filter based state-space support vector regression approach. Appl Energy 2014;113:690–705.
[13] Wang J, Botterud A, Bessa R, Keko H, Carvalho L, Issicaba D, et al. Wind power forecasting uncertainty and unit commitment. Appl Energy 2011;88:4014–23.
[14] Zhang H, Yu YJ, Liu ZY. Study on the maximum entropy principle applied to the annual wind speed probability distribution: a case study for observations of intertidal zone anemometer towers of Rudong in East China Sea. Appl Energy 2014;114:931–8.
[15] Kou P, Gao F, Guan XH. Sparse online warped Gaussian process for wind power probabilistic forecasting. Appl Energy 2013;108:410–28.

[16] Lin J, Sun YZ, Chen L, Gao WZ. Assessment of the power reduction of wind farms under extreme wind condition by a high resolution simulation model. Appl Energy 2012;96:21–32.
[17] D'Amico G, Petroni F, Prattico F. Economic performance indicators of wind energy based on wind speed stochastic modeling. Appl Energy 2015;154:290–7.
[18] Barthelmie RJ, Pryor SC. An overview of data for wake model evaluation in the Virtual Wakes Laboratory. Appl Energy 2013;104:834–44.
[19] McKay P, Carriveau R, Ting DS-K. Wake impacts on downstream wind turbine performance and yaw alignment. Wind Energy 2013;16:221–34.
[20] Hansen K, Barththelmie R, Jensen L, Sommer A. The impact of turbulence intensity and atmospheric stability on power deficits due to wind turbine wakes at horns rev wind farm. Wind Energy 2012;15:183–96.
[21] Kusiak A, Verma A. The prediction and diagnosis of wind turbine faults. IEEE Trans Sustain Energy 2011;2:87–96.
[22] Kusiak A, Verma A. A data-mining approach to monitoring wind turbines. IEEE Trans Sustain Energy 2012;3:150–7.
[23] Kusiak A, Zhang Z, Verma A. Prediction, operations, and condition monitoring in wind energy. Energy 2013;60:1–12.
[24] Hameed Z, Hong YS, Cho YM, Ahn SH, Song CK. Condition monitoring and fault detection of wind turbines and related algorithms: a review. Wind Energy 2009;13:1–39.
[25] Marvuglia A, Messineo A. Monitoring of wind farms power curves using machine learning techniques. Appl Energy 2012;98:574–83.
[26] Schlechtingen M, Santos IF, Achiche S. Using data-mining approaches for wind turbine power curve monitoring: a comparative study. IEEE Trans Sustain Energy 2013;4:671–9.
[27] Kusiak A, Zheng HY, Song Z. Models for monitoring wind farm power. Renew Energy 2009;34:583–90.
[28] Astolfi D, Castellani F, Garinei A, Terzi L. Data mining techniques for performance analysis of onshore wind farms. Appl Energy 2015;148:220–33.
[29] Pagnini LC, Burlando M, Repetto MP. Experimental power curve of small-size wind turbines in turbulent urban environment. Appl Energy 2015;154:112–21.
[30] Lydia M, Kumar SS, Selvakumar AI, Kumar GEP. A comprehensive review on wind turbine power curve modeling techniques. Renew Sustain Energy Rev 2014;30:452–60.
[31] Yang WX, Court R, Jiang JS. Wind turbine condition monitoring by the approach of SCADA data analysis. Renew Energy 2013;53:365–76.
[32] Lapira E, Brisset D, Ardakani HD, Siegel D, Lee J. Wind turbine performance assessment using multi-regime modeling approach. Renew Energy 2012;45:86–95.
[33] Kusiak A, Verma A. Monitoring wind farms with performance curves. IEEE Trans Sustain Energy 2013;4:192–9.
[34] Cross P, Ma XD. Nonlinear system identification for model-based condition monitoring of wind turbines. Renew Energy 2014;71:166–75.
[35] Yampikulsakul N, Byon E, Huang S, Sheng SW, You MD. Condition monitoring of wind power system with nonparametric regression analysis. IEEE Trans Energy Convers 2014;29:288–99.
[36] Zaher A, McArthur SDJ, Infield DG. Online wind turbine fault detection through automated SCADA data analysis. Wind Energy 2009;12:574–93.
[37] Zaher A, McArthur SDJ. A multi-agent fault detection system for wind turbine defect recognition and diagnosis. In: Proc IEEE Lausanne POWERTECH; 2007. p. 22–7.
[38] Sanz-Bobi MA, Pico JD, Garcia MC. SIMAP: intelligent system for predictive maintenance application to the health condition monitoring of a windturbine gearbox. Comput Indust 2006;57:552–68.
[39] Schlechtingen M, Santos IF, Achiche S. Wind turbine condition monitoring based on SCADA data using normal behavior models. Part 1: System description. Appl Soft Comput J 2013;13:259–70.
[40] Schlechtingen M, Santos IF. Wind turbine condition monitoring based on SCADA data using normal behavior models. Part 2: Application examples. Appl Soft Comput J 2014;14:447–60.
[41] Chandola V, Banerjee A, Kumar V. Anomaly detection: a survey. ACM Comput Surv 2009;41:1–58.
[42] Ata R. Artificial neural networks applications in wind energy systems: a review. Renew Sustain Energy Rev 2015;49:534–62.
[43] Thiae L, Fall SS, Kasse M, Sylla E, Thioye S. A neural network based approach for wind resource and wind generators production assessment. Appl Energy 2010;87:1744–8.
[44] Liu YQ, Shi J, Yang YP, Lee W. Short-term wind-power prediction based on wavelet transform-support vector machine and statistic-characteristics analysis. IEEE Trans Ind Appl 2012;48:1136–41.
[45] Chen B, Peter CM, Peter JT. Automated on-line fault prognosis for wind turbine pitch systems using supervisory control and data acquisition. IET Renew Power Gener 2015;9:503–13.
[46] Guo P, Infield D, Yang XY. Wind turbine generator condition-monitoring using temperature trend analysis. IEEE Trans Sustain Energy 2012;3:124–33.
[47] Schlechtingen M, Santos IF. Comparative analysis of neural network and regression based condition monitoring approaches for wind turbine fault detection. Mech Syst Signal Process 2011;25:1849–75.
[48] Kusiak A, Li WY. Virtual models for prediction of wind turbine parameters. IEEE Trans Energy Convers 2010;25:245–52.
[49] Kusiak A, Verma A. Analyzing bearing faults in wind turbines: a data-mining approach. Renew Energy 2012;48:110–6.
[50] Spera DA. Wind turbine technology: fundamental concepts of wind turbine engineering. New York: ASME; 1994.
[51] Assareh E, Biglari M. A novel approach to capture the maximum power from variable speed wind turbines using PI controller, RBF neural network and GSA evolutionary algorithm. Renew Sustain Energy Rev 2015;51:1023–37.
[52] Li J, Lei X, Li X, Ran L. Normal behavior models for the condition assessment of wind turbine generator systems. Electr Power Comp Syst 2014;42:1201–12.
[53] Parzen E. On the estimation of a probability density function and the mode. Ann Math Stat 1962;33:1065–76.
[54] Silverman BW. Density estimation for statistics and data analysis. New York: Chapman and Hall; 1986.
[55] Qian Z, Yan Z. Fuzzy synthetic method for life assessment of power transformer. IEE Proc Sci Measur Technol 2004;151(3):175–80.
[56] Li H, Hu YG, Yang C, Chen Z, Ji HB, Zhao B. An improved fuzzy synthetic condition assessment of a wind turbine generator system. Int J Electr Power Energy Syst 2013;45:468–76.
[57] Lu RS, Lo SL, Hu JY. Analysis of reservoir water quality using fuzzy synthetic evaluation. Stoch Env Res Risk A 1999;13:327–36.
[58] Wang PZ. Fuzzy set theory and its applications. Shanghai: Shanghai Science and Technology Press; 1983 [in Chinese].
[59] Rodrigures RVJ, Arkkio A. Detection of stator winding fault in induction motor using fuzzy logic. Appl Soft Comput J 2008;8:1112–20.