IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 1 June 2022, accepted 17 June 2022, date of publication 22 June 2022, date of current version 29 June 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3185259

A SCADA-Data-Driven Condition Monitoring Method of Wind Turbine Generators

LISHU WANG[1], SHUHAN JIA[1], XIHUI YAN[2], LIBO MA[3], AND JUNLONG FANG[1]
[1]Institute of Electrical and Information, Northeast Agricultural University, Harbin 150030, China
[2]State Grid Shijiazhuang Electric Power Supply Company, Shijiazhuang 050051, China
[3]Department of Electrical Engineering, North China Electric Power University, Baoding 071003, China

Corresponding author: Lishu Wang (wanglishu@neau.edu.cn)

This work was supported in part by the Heilongjiang Education Department, Harbin, China, under Grant 12521038.

ABSTRACT Changes in sensor measurement parameters of wind turbine SCADA systems usually do not provide reliable early alarms. To detect early faults or abnormal conditions of wind turbine generator components, a wind turbine generator condition monitoring framework based on the fusion of cascaded SAE abnormal condition monitoring and LightGBM abnormal condition classification is proposed. The framework consists of two parts. The first part is a strong anti-interference cascade SAE anomaly condition monitoring method considering that early anomalies are easily flooded. The cascade SAE is trained with polynomial features and original features. The isolated forest is used to determine the alarm threshold of reconstruction error between the input and output of the cascade SAE. The operating condition of the wind turbine is judged by comparing the magnitude between the reconstruction error and this threshold. The second part is the anomaly condition classification based on LightGBM. The optimal parameters of LightGBM are searched by Bayesian optimization to build a LightGBM multi-classification anomaly condition classification model. The results of the case study show that the proposed condition monitoring has high anomaly recognition capability: the cascaded SAE method has strong anti-interference properties and can capture the early abnormal conditions of wind turbine generators; LightGBM has a faster training speed than other classifiers with guaranteed abnormality classification accuracy.

INDEX TERMS Wind turbine, condition monitoring, stacked auto-encoder, isolation forest, LightGBM.

I. INTRODUCTION

The harsh operating environment of wind farms results in wind turbines having a high failure rate [1]. The O&M cost of onshore wind turbines accounts for about 10%-15% of the total wind farm revenue [2]. Generator system failure is one of the main causes of wind turbine downtime and accounts for 37% of all fault downtime [3], [4]. Therefore, it is important to diagnose generator faults as early as possible to reduce downtime and maximize productivity. Condition monitoring is the process of determining if there are any abnormalities in the operating condition of wind turbines and when they occur; abnormality identification determines the type of abnormality or time-varying behavior [5], [6]. The abnormal condition of a wind turbine may develop into a permanent failure or may be able to recover to its original state after some time.

An effective wind turbine condition monitoring system requires the installation of numerous high-frequency sampling sensors. The wind turbine supervisory control and data acquisition (SCADA) system is capable of collecting and remotely or locally monitoring the operating parameters of the entire wind farm generator, and is characterized by fast signal changes and numerous operating parameters. The fault or abnormal characteristics of wind turbines are implicit in the SCADA variables that characterize their operating status, so the abnormal information carried by the SCADA variables can be mined to monitor the status of the wind turbine or alarm the abnormal status.

Fault alarms for a range of subassemblies of wind turbines are performed using wind speed statistics in [7]. A fault detection algorithm based on Gaussian process is proposed based on SCADA data with operational variables (pitch angle and rotor speed) as inputs to an additional model in [8]. The effectiveness of wind turbine fault alarms can be improved by processing SCADA data or extracting certain features, such as processing actual SCADA imbalanced data [9], [10] and using NOFRFs approach to extract damage sensitive features [11].

The associate editor coordinating the review of this manuscript and approving it for publication was Chuan Li.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

The principal component analysis is used to select a set of partial variables containing the variation characteristics of the original data to locate wind turbine faults [12]. Models such as generative adversarial networks, transfer learning, and convolutional auto-encoders can be applied to fault detection scenarios with small samples or the same type of wind turbines [13], [14]. Long short-term memory networks and spatio-temporal multiscale neural networks considering spatio-temporal characteristics can effectively capture the fault information of wind turbines in SCADA data [10], [15], [16]. The method of generating reference space or constructing residuals based on the normal behavior of wind turbines detaches the abnormal conditions from the normal data by measuring the difference between normal and abnormal conditions of wind turbines [17]–[19]. Different variants of auto-encoders and neural networks are widely used for wind turbine fault diagnosis [18], [20].

Many studies can detect anomalies in wind turbines, but no further identification of the detected anomalies is done. Classification algorithms can be used to analyze fault data features to identify specific faults in wind turbines [21], [22]. Optimized support vector machine classifiers can be better implemented for wind turbine fault diagnosis [23], [24]. Various ensemble learning algorithms (e.g., random forest and XGBoost) using decision trees as base learners can provide wind turbine condition monitoring schemes [25]–[29]. The XGBoost algorithm has been further improved in terms of the loss function, regularization, and parallelization processing, and has better classification performance, which can improve the accuracy of fault identification more effectively [30]. A wind turbine condition monitoring method based on multi-feature monitoring parameter information fusion is proposed using an optimization scheme based on the Bayesian optimization algorithm and XGBoost feature weight measurement [28]. Although XGBoost has high classification accuracy, its training speed is not advantageous.

For wind turbine generator abnormal condition monitoring and abnormal condition classification, this paper builds a condition monitoring framework based on cascaded stacked auto-encoder (SAE) and LightGBM to achieve early abnormal condition capture and fast and accurate abnormal condition classification of wind turbines. Firstly, the normal wind turbine SCADA data are used to train the cascaded SAE network and the alarm threshold is determined by the isolation forest. For the wind turbine SCADA data containing abnormal conditions, the magnitude of reconstruction error and alarm threshold are compared to determine whether the wind turbine is abnormal or not. Then Bayesian optimization is used to search the hyperparameters of the LightGBM multi-classification model to identify different anomaly types. Finally, the effectiveness of the proposed method is verified with actual failure cases of wind turbines.

The paper is organized as follows. Section II describes the wind turbine abnormal condition monitoring model. Section III elaborates on the wind turbine abnormal condition classification model. Section IV trains the wind turbine condition monitoring framework. Section V analyzes the effectiveness of the model with examples. Section VI briefly presents the conclusions of this paper.

II. WIND TURBINE ABNORMAL CONDITION MONITORING

To detect abnormal operating conditions of wind turbine generator bearing assemblies promptly and improve the accuracy of abnormality monitoring, a cascaded stacked auto-encoder (SAE)-based abnormality monitoring model is constructed. Auto-Encoder (AE) adjusts the model parameters in an unsupervised learning manner so that the model output reconstructs the input as accurately as possible. An AE consists of an encoder and a decoder, where the encoder extracts abstract features of the data; the decoder is the inverse of the encoder and constructs reconstructed values close to the input. The reconstructed values have the same physical meaning as the input values.

The data under normal conditions are selected as the input for training AE. In this paper, the normalized Supervisory Control and Data Acquisition (SCADA) data under normal operation of wind turbines are selected as the input of AE. The cross-sectional data of the wind turbine SCADA measurement points under a certain moment are $\mathbf{x}$. The encoder and decoder are expressed as

$$\mathbf{h} = \sigma_1(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1) \tag{1}$$

$$\hat{\mathbf{x}} = \sigma_2(\mathbf{W}_2\mathbf{h} + \mathbf{b}_2) \tag{2}$$

where $\mathbf{x}$, $\mathbf{h}$ and $\hat{\mathbf{x}}$ are the input of the input layer, the output of the implied layer and the output of the output layer of AE, respectively. σ_1 and σ_2 are the activation functions of the encoder and decoder, $\mathbf{W}_1$ and $\mathbf{W}_2$ are the weights of the different layers, $\mathbf{b}_1$ and $\mathbf{b}_2$ are the biases of the different layers.

The training process of AE is the process of adjusting the parameter set $\{\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1, \mathbf{b}_2\}$ to minimize the distance metric function $\text{dist}(\mathbf{x}, \hat{\mathbf{x}})$ between the input and output. The SAE is formed by connecting several AEs in series, which can extract the higher-order features of the input data layer by layer and reduce the dimensionality of the input data layer by layer.

The operating condition of the wind turbine is coupled with each SCADA variable. Considering the high-dimensional features of SCADA feature variables and the influence of polynomial features among variables on SAE, a combined model of 2 SAE cascades is constructed in this paper to improve the effectiveness of anomaly monitoring. To obtain the high-dimensional features and interrelated features of SCADA feature variables $\mathbf{x}$, the polynomial and interaction features of $\mathbf{x}$ are generated. For example, for the n-th polynomial feature between two variables a and b as

$$[1, a^n b^0, a^n -^1 b^1, a^n -^2 b^2, \ldots, a^0 b^n] \tag{3}$$

The input to the first SAE is the polynomial feature $\mathbf{x}'$ (which does not contain the constant 1 and the variable $\mathbf{x}$

itself), and the output $\hat{\mathbf{x}}'$ is obtained by training the SAE. The reconstruction error e' between the original input $\mathbf{x}'$ and the output reconstructed value $\hat{\mathbf{x}}'$ is defined as the Euclidean norm of the difference between the two, i.e.

$$e' = \left\| \mathbf{x}' - \hat{\mathbf{x}}' \right\|_2 \tag{4}$$

The input of the second SAE is the feature $[\mathbf{x}, e']$ of the original SCADA feature variable $\mathbf{x}$ combined with the reconstruction error e' of the first SAE. The reconstruction error e of the 2nd SAE input and output is used as the monitoring variable to measure whether the wind turbine is abnormal or not. The structure of the wind turbine abnormal condition monitoring model is shown in Section IV. Cascading SAE enables the reconstruction error to describe the operating state of the wind turbine more accurately, and reduces the situation that the large fluctuation of the wind turbine is mistakenly identified as abnormal.

Under normal operating conditions of wind turbines, there is a stable correlation between their SCADA variables. When an abnormality occurs in the generator components, it is manifested as an abnormal value of one or several SCADA variables, resulting in a large deviation of SAE reconstruction value. With the further aggravation of the abnormality, the reconstruction error value also increases gradually, so the reconstruction error can be used to judge whether the wind turbine is abnormal. When the reconstruction error exceeds the alarm threshold, it can be judged that the wind turbine enters the abnormal alarm condition, which may be the development stage of the early fault of the wind turbine.

In this paper, the alarm threshold for wind turbine anomaly monitoring is determined by the isolated forest algorithm. Isolated forest is similar to random forest, but the selection of features for division and points for segmentation is random each time, rather than based on information gain or Gini index. During the tree building process, if a sample reaches the leaf node quickly (i.e., the distance from the leaf to the root is short), this sample may be anomalous. Anomalous samples can be isolated by fewer times of random feature segmentation compared to normal samples. The reconstruction error of the second SAE is used as the input of the isolated forest, and the alarm threshold is determined by the value of the detected reconstruction error anomalies.

III. WIND TURBINE ABNORMAL CONDITION CLASSIFICATION

To determine the type of abnormal condition of wind turbine found by cascaded SAE, this paper predicts the type of possible early fault or abnormal condition of the wind turbine by LightGBM. LightGBM is an improved optimization algorithm based on Gradient Boosting Decision Tree (GBDT), by building multiple decision trees (base learners) to synthesize the output of the decision tree population to obtain the final result [31].

For given a data set with n examples and m SCADA features $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n} (\mathbf{x}_i \in \mathbf{R}^m, y_i \in \mathbf{R})$, LightGBM predicts the abnormal condition type of wind turbines as $\hat{y}_i$. y_i is the target value corresponding to $\mathbf{x}_i$, i.e., the abnormal condition categories of wind turbines. $\mathbf{x}_i$ denotes the selected SCADA features.

LightGBM uses K additive functions to predict the final classification target, i.e.

$$\hat{y}_i = \phi(\mathbf{x}_i) = \sum_{k=1}^{K} f_k(\mathbf{x}_i), \quad f_k \in \Gamma \tag{5}$$

where $\Gamma = \{f_k(\mathbf{x}) = \omega_{q(\mathbf{X})}\}$ (q: $\mathbf{R}^m \rightarrow T$, $\omega \in \mathbf{R}^T$) is the function space composed of decision trees. q denotes the structure of each tree that maps an example to the corresponding leaf index. T denotes the number of leaves in the tree. Each f_k corresponds to an independent tree structure q and leaf weights ω.

In order to learn the set of functions in the model, the learning objective function of LightGBM is

$$L(\phi) = \sum_{i} l(\hat{y}_i, y_i) + \sum_{k} \Omega(f_k) \tag{6}$$

where l is a differentiable convex loss function that measures the difference between the prediction $\hat{y}_i$ and the target y_i. $\Omega(f) = \gamma T + \frac{1}{2}\lambda \|\omega\|^2$ is the regularization term to penalize the complexity of the model and prevent overfitting. γ and λ are regularization coefficients.

let $\hat{y}_i^{(t)}$ be the prediction of the i-th instance at the t-th iteration, LightGBM adds a new function f_t, i.e., uses a stepwise forward additivity model to maximize the reduction of the following objective function.

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)) + \Omega(f_t) \tag{7}$$

Second-order approximation can be used to quickly optimize this objective function.

$$\tilde{L}^{(t)} \approx \sum_{i=1}^{n} [g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i)] + \Omega(f_t) \tag{8}$$

where $g_i = \frac{\partial l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}}$, $h_i = \frac{\partial^2 l(y_i, \hat{y}_i^{(t-1)})}{\partial (\hat{y}_i^{(t-1)})^2}$.

Define $I_j = \{i \,|q\,(\mathbf{x_i}) = j\}$ as the instance set of leaf j. The model complexity can be written as $\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2$. The model objective function is

$$\tilde{L}^{(t)} = \sum_{j=1}^{T} [\omega_j \sum_{i \in I_j} g_i + \frac{\omega_j^2}{2} (\sum_{i \in I_j} h_i + \lambda)] + \gamma T \tag{9}$$

For a fixed structure $q(\mathbf{x})$, the optimal weight of leaf j and the corresponding optimal objective function value are

$$\omega_j^* = -(\sum\nolimits_{i \in I_j} g_i)/(\sum\nolimits_{i \in I_j} h_i + \lambda) \tag{10}$$

$$\tilde{L}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^{T} [(\sum_{i \in I_j} g_i)^2 / \sum_{i \in I_j} h_i + \lambda] + \gamma T \tag{11}$$

Assume that I_L and I_R are the instance sets of left and right nodes after the split. Letting $I = I_L \cup I_R$, the structure loss

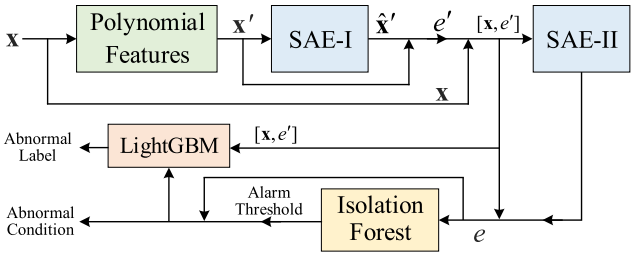

FIGURE 1. Structural framework of wind turbine condition monitoring.

reduction after the split can be used to determine whether to divide and to identify the division candidates.

LightGBM's base classifier is the classification and regression tree based on histogram algorithm, which divides feature values into many bins and searches for split points on the bins. LightGBM abandons the level-wise decision tree growth strategy and uses the leaf-wise algorithm with depth restrictions. The LightGBM algorithm contains two innovative techniques, which are the gradient-based one-side sampling and the exclusive feature bundling, respectively, that enable the model to handle large-scale data and features more efficiently.

IV. WIND TURBINE CONDITION MONITORING

A. MODEL FRAMEWORK

In order to detect abnormal conditions during wind turbine operation and to identify the type of that abnormality, a wind turbine condition monitoring framework that incorporates cascaded SAE and LightGBM is designed, as shown in Figure 1.

The anomaly monitoring model based on two cascaded SAEs is trained using wind turbine normal operation data. The isolated forest algorithm is used to mine outliers for the reconstruction error constructed by the second SAE, and the alarm threshold is determined by the outliers. If the reconstruction error is greater than the alarm threshold, the operating status of the wind turbine is abnormal. The wind turbine anomaly data are used as input to the cascaded SAE model to calculate the reconstruction error. The input SCADA data samples are labeled as normal or abnormal based on the magnitude between the reconstruction error and the alarm threshold. By adding labels to SCADA data with known abnormal types, the LightGBM classification model can be trained to further determine the fault category for the abnormal data screened by the cascade SAE.

The idea of the wind turbine condition monitoring framework is: based on the wind turbine SCADA data, the cascade SAE monitors whether the wind turbine is abnormal in a certain period, and LightGBM further identifies the specific type of the abnormality.

B. MODEL TRAINING

The inputs to the training cascade SAE are multiple SCADA characteristic variables of wind turbines under normal conditions. In this paper, three types of faults in

TABLE 1. Selected SCADA feature variables.

No.	Parameter Name	Notation
1	wind speed (m/s)	v
2	grid side power	P
3	generator front bearing temperature	T_a
4	generator rear bearing temperature	T_b
5	grid U_1 voltage	U_1
6	grid U_2 voltage	U_2
7	grid U_3 voltage	U_3
8	grid I_1 current	I_1
9	grid I_2 current	I_2
10	grid I_3 current	I_3

generator bearing components of wind turbines are collected and 10 important SCADA variables are selected, namely, wind speed v, grid side power P, generator front bearing temperature T_a, generator rear bearing temperature T_b, grid-side three-phase voltage $U_1 \sim U_3$, and grid-side three-phase current $I_1 \sim I_3$ as shown in Table 1.

The cascaded SAE-based abnormal condition monitoring model for wind turbines requires training two SAE networks. Different models of wind turbines need to be trained separately for their respective anomaly monitoring models and to determine the alarm thresholds. Before training the first SAE in the anomaly monitoring model, polynomial features need to be constructed.

The theoretical maximum power output of the wind turbine is

$$P = \frac{1}{2}\rho A v^3 C_p \tag{12}$$

where C_p is called wind energy utilization coefficient, its maximum value is 0.59. In the actual wind turbine power limit is smaller than Baez's law, usually take the value of 0.35∼0.45.

Since the theoretical maximum power output of the wind turbine is proportional to the third power of wind speed, the third power of wind speed v^3 is added as the characteristic variable. The above ten feature variables (removing the power generation P and adding v^3) are the base feature variables. Since wind turbine power is related to a variety of parameters, polynomial features of power 2 are generated with the wind turbine SCADA base features. The constructed polynomial features retain only the combined features between each feature, i.e., they do not contain 0-th power features (i.e., 1), features themselves (i.e., a, b), and combinations of features themselves and themselves (i.e., a^2, b^2). After generating the polynomial features, the correlation coefficient between the polynomial features and P is found using wind turbine power generation as the reference variable. The feature variables with correlation coefficients greater than 0.8 are selected, and finally 35 polynomial features are retained.

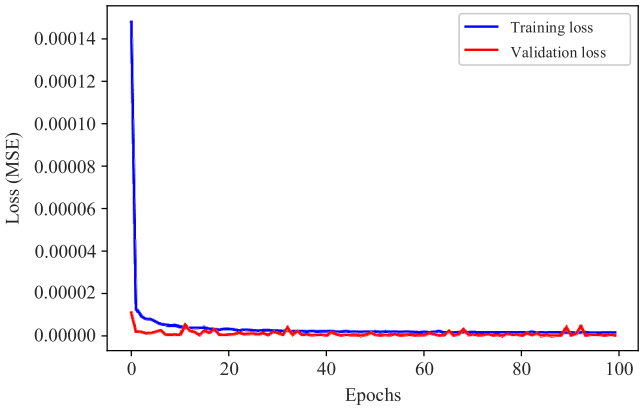

FIGURE 2. Loss function curves of the first SAE.

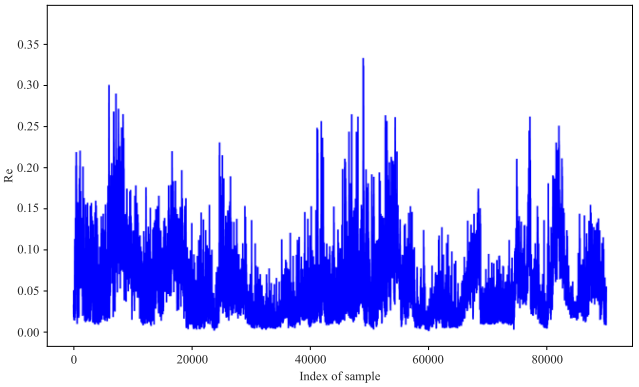

FIGURE 3. Reconstruction errors between inputs and outputs of cascaded SAEs.

These 35 polynomial features are used as the input of the first SAE, and the output dimensions of each layer in this SAE encoding process are set to 1000, 500, 250, and 50, respectively. The activation function is chosen as ReLU, the loss function is chosen as MSE. The model parameters are adjusted with the Adam optimizer. The variation of the loss function curves in the training SAE process is shown in Figure 2, which indicates that the fitting effect of both the training and validation sets is good.

Letting the reconstruction error between the input and output of the 1st SAE be the feature variable e', the 10 features in Table 1 with e' (i.e., $[\mathbf{x}, e']$) are used as the input of the 2nd SAE. The output dimensions of each level of the 2nd SAE coding process are set to 500, 250, and 50, respectively, and other parameters are the same as the 1st SAE. The reconstruction error (denoted by Re) constructed based on the anomaly monitoring model of the two SAE cascades is shown in Figure 3.

The alarm threshold for cascaded SAE anomaly monitoring is determined by the isolated forest algorithm, and the proportion of anomalies in the sample is set to 0.02. The outlier points of the isolated forest detection reconstruction error data distribution are shown in Figure 4. The average value of all outliers detected by the isolated forest is taken. In order to reduce the misjudgment of the condition monitoring caused by the large fluctuation of the wind turbine

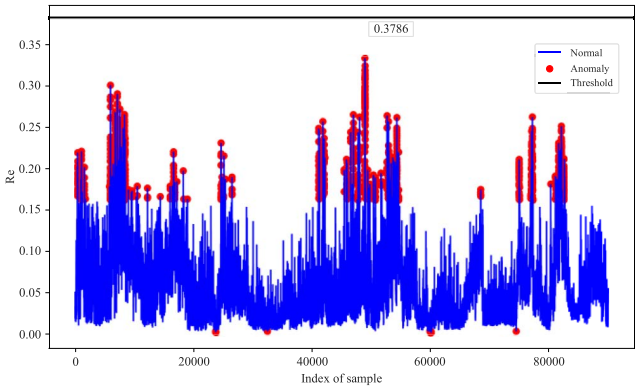

FIGURE 4. Distribution of outlier points of reconstruction error.

TABLE 2. Tuning parameters of LightGBM.

No.	Parameter Name	The parameter value
1	max_depth	6
2	num_leaves	11
3	colsample_bytree	0.8436
4	min_child_samples	245
5	min_child_weight	0.1
6	reg_alpha	2
7	reg_lambda	5
8	subsample	0.823

during normal operation, the reconstruction error is set to be twice this average value. In this paper, if the reconstruction error is greater than the alarm threshold and its duration exceeds 30s (for wind turbines with a 1s acquisition interval), these wind turbines are considered to have an abnormal operation after that time.

The actual anomaly labels are added to the anomaly feature samples $[\mathbf{x}, e']$ detected by the cascaded SAE as the input for training LightGBM. Because the proportion of data corresponding to each anomaly label in the wind turbine SCADA data is uneven, this paper uses stratified sampling to ensure that the proportion of label samples used for LightGBM training is the same as the original data set. Bayesian optimization search and k-fold cross-validation are used. The number of cross-validation is chosen as 5-fold to determine the optimal LightGBM hyperparameters at one time, as shown in Table 2. The trained LightGBM can output the predicted anomaly types.

V. CASE STUDY

A. CASCADE SAE ABNORMAL CONDITION MONITORING

The reconstruction error calculated by the cascaded SAE anomaly monitoring model is small under the normal operation of the wind turbine. During abnormal hours or early fault development, the reconstruction error increases suddenly or has a creeping process. Comparing the magnitude of the reconstruction error value with the alarm threshold can identify whether an early failure or anomaly exists in the wind

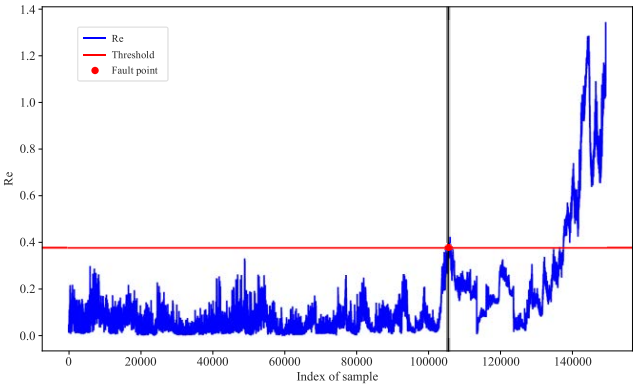

FIGURE 5. Reconstruction error of the wind turbine in Case 1.

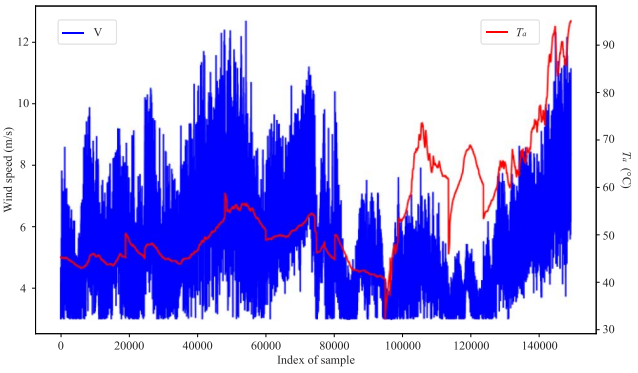

FIGURE 6. Curves of wind speed and T_a in Case 1.

turbine. The cascade SAE abnormality monitoring capability is analyzed with actual cases of 3 different wind turbines in wind farms.

A wind turbine in Case 1, which collects SCADA cross-section data every 1s, was shut down at the site on 2015-10-17 01:11:15 for the generator front bearing temperature overrun fault. The cascade SAE found the anomaly at the red dot in Figure 5 (i.e., 2015-10-16 13:03:21), 12:07:54 ahead of the site time (the alarm threshold set by the isolated forest is 0.3786). The generator front bearing temperature overrun fault is triggered by the generator front bearing temperature greater than 95 degrees Celsius and lasts for 5 seconds. In this paper, in order to reduce the false alarm generated by operating fluctuations, the moment after the reconstruction error is greater than the alarm threshold and lasts for more than 30s is taken as the moment when the hidden trouble occurs in the wind turbine.

As can be seen from Figure 5, the reconstruction error of this wind turbine in the sample index interval [110000, 130000] after the red point falls back to normal. By analyzing the curves of wind speed v and generator front bearing temperature T_a (as shown in Figure 6), the wind turbine was found abnormal near the red point moment. Larger increases and decreases in T_a occurred, but T_a did not meet the conditions for this wind turbine to trigger the temperature overrun fault. In the sample index interval [110000, 130000], the cascaded SAE judged that the abnormality of the wind turbine has disappeared.

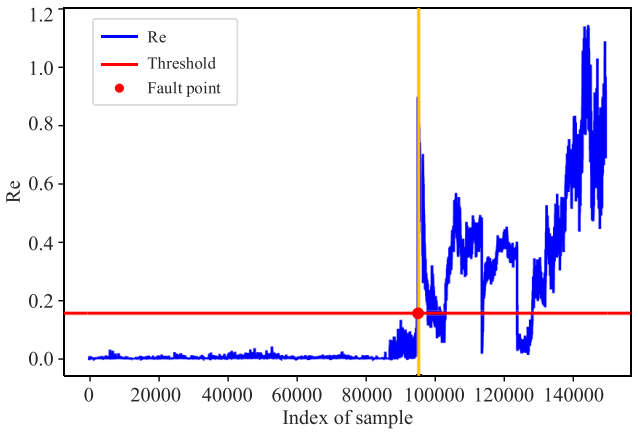

FIGURE 7. Reconstruction error of a single SAE in Case 1.

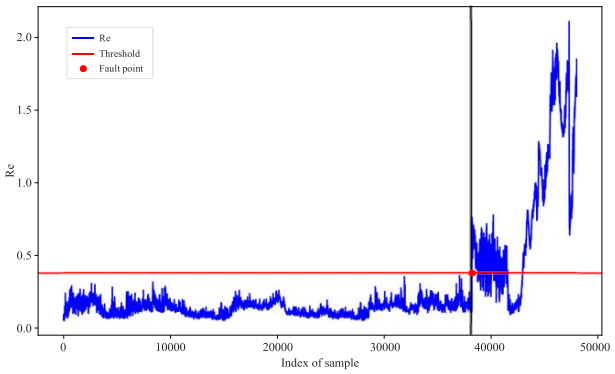

FIGURE 8. Reconstruction error of the wind turbine in Case 2.

A single SAE was constructed as a comparison to the cascaded SAE model with the coding process with output dimensions of 500, 250, and 50 for each layer, and the inputs were the variables in Table 1. The wind turbine anomaly monitoring of the single SAE is shown in Figure 7, which amplifies the larger fluctuations of this wind turbine in the sample index interval [90000, 130000]. At this point, the variable T_a physically directly reflects the fluctuations of this wind turbine and dominates the trend of the reconstruction error. Although a single SAE can detect the abnormal condition of wind turbines earlier than the cascaded SAE, it is less robust and more affected by the larger fluctuations.

A wind turbine in Case 2 records a section data every 1s and fails to shut down at 2015-07-24 04:50:00 in the field due to generator rear bearing temperature overrun fault. The cascade SAE model of this wind turbine was constructed, and the isolated forest determined its alarm threshold to be 0.4218. The reconstruction error of the cascade SAE at 2015-07-24 02:06:42 was greater than the alarm threshold, and the wind turbine was judged to have an abnormality. This fault was found earlier than the site at 02:43:18, as shown in Figure 8. The trigger condition for the generator rear bearing temperature overrun fault is after the generator rear bearing temperature is greater than 95 degrees Celsius and lasts for 5 seconds.

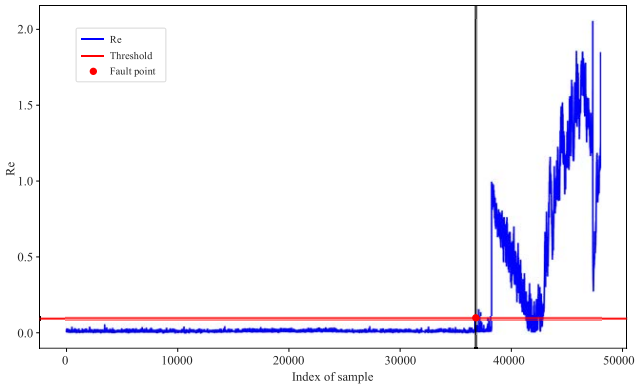

FIGURE 9. Reconstruction error of a single SAE in Case 2.

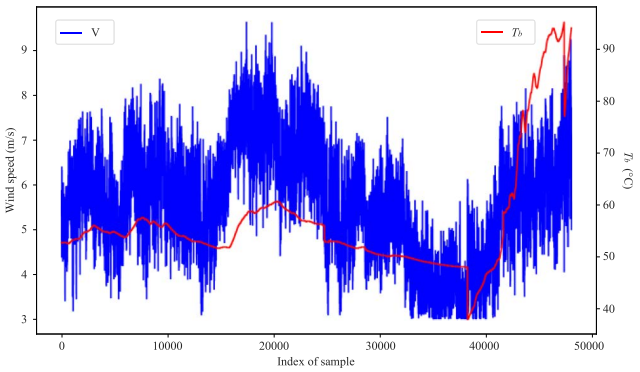

FIGURE 10. Curves of wind speed and T_b in Case 2.

Comparing the single SAE model (the model structure is the same as the setup of the single SAE in Case 1), as shown in Figure 9, the time to detect anomalies is similar to that of the cascaded SAE. The single SAE is less resistant to interference than the cascaded SAE, and both have roughly the same reconstruction error variation trend.

The variable T_b physically directly reflects this fault of this wind turbine. After the sample index value of 45000, the trend of the reconstruction error is similar to the trend of T_b, as shown in Figure 10. The generator bearing temperature overrun fault needs to focus on the trend of bearing temperature.

A wind turbine in Case 3 records a section data every 10min and fails to shut down at 2015-10-23 17:00:00 due to damage to the front and rear bearings of the generator. The cascade SAE model of this wind turbine was constructed, and the isolated forest determined its alarm threshold to be 0.3702. The reconstruction error of the cascade SAE at 2015-10-21 07:50:00 was greater than the alarm threshold, and the wind turbine was judged to have an early fault. Its fault was found at 09:10:00 earlier than the site, as shown in Figure 11.

Compared with the single SAE model, as shown in Figure 12, the single SAE model basically cannot detect the early failure of this wind turbine, it only detects anomalies in very short intervals. Its abnormality monitoring capability

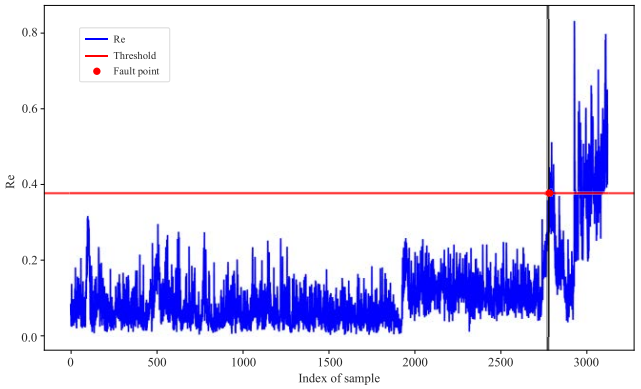

FIGURE 11. Reconstruction error of the wind turbine in Case 3.

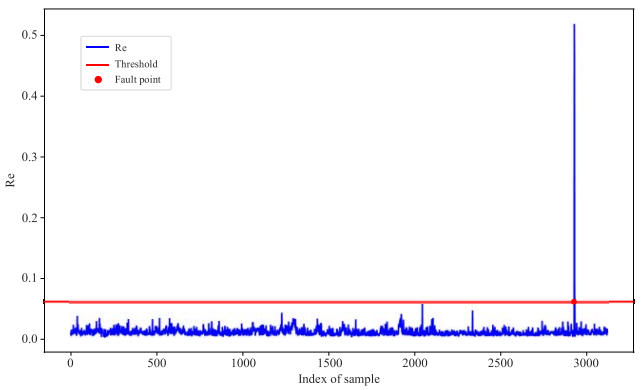

FIGURE 12. Reconstruction error of a single SAE in Case 3.

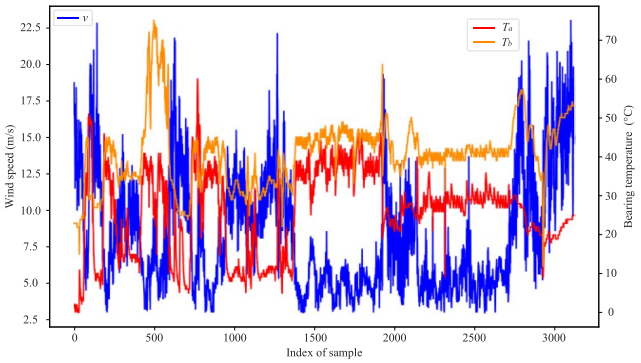

FIGURE 13. Curves of wind speed, T_a and T_b in Case 3.

is worse than the cascaded SAE. The trends of wind speed, T_a and T_b of this wind turbine are shown in Figure 13. The values of T_a and T_b fluctuate within the normal range before the failure shutdown of this wind turbine.

When early faults occur in wind turbine operation, the correlation relationship between variables is destroyed. But the cascaded SAE model still outputs the corresponding reconstructed variables for abnormal data according to the correlation relationship in normal time, which leads to an increase in the reconstruction error value, so the abnormal condition or early faults of wind turbines can be monitored. The samples with reconstruction error greater than the alarm

TABLE 3. Confusion matrix for LightGBM anomaly identification results.

Conditions	Predicted F_1	Predicted F_2	Predicted F_3
Actual F_1	2964	0	0
Actual F_2	0	1766	0
Actual F_3	0	2	29

TABLE 4. Comparison of different classifiers.

Classifiers	Training speed (s)	F-score	Accuracy score
SVM	152.1771	0.98	0.9785
decision tree	1.5786	0.97	0.9695
random forest	23.9611	1.00	0.9989
XGBoost	11.4288	1.00	0.9992
LightGBM	1.8606	1.00	0.9995

threshold are added with normal or fault labels to practice the LightGBM algorithm for further abnormal condition categories.

B. LIGHTGBM ABNORMAL CONDITION CLASSIFICATION

After detecting the sample data of wind turbine reconstruction errors greater than the alarm threshold, the LightGBM model is trained with the above three types of faults to identify the corresponding anomaly categories. The LightGBM is trained with the anomaly data detected by the cascaded SAE for these three types of wind turbines, including three types of fault conditions: generator front bearing temperature overrun fault F_1, generator rear bearing temperature overrun fault F_2, and generator front and rear bearing damage F_3.

The wind turbine anomaly identification results are shown in the confusion matrix in Table 3. The values in the table indicate the number of entries of wind turbine SCADA data. The diagonal data of the confusion matrix is the number of correct classifications, and the off-diagonal is the number of incorrect classifications in the corresponding row or column.

Table 3 shows that LightGBM was able to identify faults F_1 and F_2 100% of the time, and for fault F_3, there were very few times when it was incorrectly classified as fault F_2. The reason for this could be the similar fluctuations in the changes in the SCADA characteristics of the wind turbines for both early faults after the reconstruction error was greater than the threshold.

The cascaded SAE anomaly monitoring algorithms are constructed in the same way for these three fault types of wind turbines. Different types of classifiers are trained based on SCADA anomaly data with added anomaly labels. LightGBM was compared with support vector machine (SVM), decision tree, random forest, and XGBoost, as shown in Table 4.

From Table 4, LightGBM has better classification accuracy than other classifiers. Random forest and XGBoost are both decision tree based algorithms and their accuracy is close to LightGBM. The Decision tree training speed is faster than LightGBM because of the low complexity of the decision tree model, so its training parameters take a short time, but the decision tree accuracy is lower. XGBoost accuracy is closest to LightGBM, and its training speed is faster than other classifiers, but it is 6 times longer than the training time of LightGBM. LightGBM has certain advantages in training time and classification accuracy. The effectiveness of LightGBM for wind turbine abnormal condition classification is verified.

VI. CONCLUSION

In order to detect early faults or abnormal conditions of wind turbine generator components in a timely manner, this paper designs a framework for wind turbine generator condition monitoring based on cascaded SAE and LightGBM, and verifies the effectiveness of this framework through case studies. By training the cascade SAE with normal operation data of wind turbines and setting appropriate alarm thresholds for different wind turbines, the abnormal data capture of early faults of generator components is achieved. The problem of difficulty in obtaining early fault samples in the field is also solved. The cascaded SAE is less affected by the fluctuation of wind turbine operation than the single SAE and has a certain anti-interference capability. The parameters of the LightGBM anomaly classification model are determined by Bayesian optimization search and combined with stratified sampling and 5-fold cross-validation. Compared with other classifiers, LightGBM has higher classification accuracy and faster training speed, and can accurately identify wind turbine anomaly categories.

REFERENCES

[1] W. Qiao and D. Lu, "A survey on wind turbine condition monitoring and fault diagnosis—Part I: Components and subsystems," *IEEE Trans. Ind. Electron.*, vol. 62, no. 10, pp. 6536–6545, Oct. 2015.
[2] B. Lu, Y. Li, X. Wu, and Z. Yang, "A review of recent advances in wind turbine condition monitoring and fault diagnosis," in *Proc. IEEE Power Electron. Mach. Wind Appl.*, Jun. 2009, pp. 1–7.
[3] J. M. P. Pérez, F. P. G. Márquez, A. Tobias, and M. Papaelias, "Wind turbine reliability analysis," *Renew. Sustain. Energy Rev.*, vol. 23, pp. 463–472, Jul. 2013.
[4] Y. Zhao, D. Li, A. Dong, D. Kang, Q. Lv, and L. Shang, "Fault prediction and diagnosis of wind turbine generators using SCADA data," *Energies*, vol. 10, no. 8, p. 1210, Aug. 2017.
[5] S. Simani and S. Farsoni, *Fault Diagnosis and Sustainable Control of Wind Turbines: Robust Data-Driven and Model-Based Strategies*. Woburn, MA, USA: Butterworth-Heinemann, 2018.
[6] N. Mehranbod, M. Soroush, and C. Panjapornpon, "A method of sensor fault detection and identification," *J. Process Control*, vol. 15, no. 3, pp. 321–339, Apr. 2005.
[7] A. K. Papatzimos, P. R. Thies, and T. Dawood, "Offshore wind turbine fault alarm prediction," *Wind Energy*, vol. 22, no. 12, pp. 1779–1788, Dec. 2019.
[8] R. Pandit, D. Infield, and T. Dodwell, "Operational variables for improving industrial wind turbine yaw misalignment early fault detection capabilities using data-driven techniques," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–8, 2021.
[9] C. Velandia-Cardenas, Y. Vidal, and F. Pozo, "Wind turbine fault detection using highly imbalanced real SCADA data," *Energies*, vol. 14, no. 6, p. 1728, Mar. 2021.
[10] Q. He, Y. Pang, G. Jiang, and P. Xie, "A spatio-temporal multiscale neural network approach for wind turbine fault diagnosis with imbalanced SCADA data," *IEEE Trans. Ind. Informat.*, vol. 17, no. 10, pp. 6875–6884, Oct. 2021.

[11] S. Zhang and Z.-Q. Lang, "SCADA-data-based wind turbine fault detection: A dynamic model sensor method," *Control Eng. Pract.*, vol. 102, Sep. 2020, Art. no. 104546.
[12] Y. Wang, X. Ma, and P. Qian, "Wind turbine fault detection and identification through PCA-based optimal variable selection," *IEEE Trans. Sustain. Energy*, vol. 9, no. 4, pp. 1627–1635, Oct. 2018.
[13] J. Liu, F. Qu, X. Hong, and H. Zhang, "A small-sample wind turbine fault detection method with synthetic fault data using generative adversarial nets," *IEEE Trans. Ind. Informat.*, vol. 15, no. 7, pp. 3877–3888, Jul. 2018.
[14] Y. Li, W. Jiang, G. Zhang, and L. Shu, "Wind turbine fault diagnosis based on transfer learning and convolutional autoencoder with small-scale data," *Renew. Energy*, vol. 171, pp. 103–115, Jun. 2021.
[15] M. S. Li, D. Yu, Z. Chen, K. Xiahou, T. Ji, and Q. H. Wu, "A data-driven residual-based method for fault diagnosis and isolation in wind turbines," *IEEE Trans. Sustain. Energy*, vol. 10, no. 2, pp. 895–904, Apr. 2018.
[16] J. Lei, C. Liu, and D. Jiang, "Fault diagnosis of wind turbine based on long short-term memory networks," *Renew. Energy*, vol. 133, pp. 422–432, Apr. 2018.
[17] L. Wei, Z. Qian, and H. Zareipour, "Wind turbine pitch system condition monitoring and fault detection based on optimized relevance vector machine regression," *IEEE Trans. Sustain. Energy*, vol. 11, no. 4, pp. 2326–2336, Oct. 2020.
[18] X. Jin, Z. Xu, and W. Qiao, "Condition monitoring of wind turbine generators using SCADA data analysis," *IEEE Trans. Sustain. Energy*, vol. 12, no. 1, pp. 202–210, Jan. 2021.
[19] Y. Liu, Z. Wu, and X. Wang, "Research on fault diagnosis of wind turbine based on SCADA data," *IEEE Access*, vol. 8, pp. 185557–185569, 2020.
[20] J. Zhang, H. Sun, Z. Sun, W. Dong, and Y. Dong, "Fault diagnosis of wind turbine power converter considering wavelet transform, feature analysis, judgment and bp neural network," *IEEE Access*, vol. 7, pp. 179799–179809, 2019.
[21] Y. Li, S. Liu, and L. Shu, "Wind turbine fault diagnosis based on Gaussian process classifiers applied to operational data," *Renew. Energy*, vol. 134, pp. 357–366, Apr. 2019.
[22] G. Q. Jiang, H. He, J. Yan, and P. Xie, "Multiscale convolutional neural networks for fault diagnosis of wind turbine gearbox," *IEEE Trans. Ind. Electron.*, vol. 66, no. 4, pp. 3196–3207, Apr. 2018.
[23] W. Tuerxun, X. Chang, G. Hongyu, J. Zhijie, and Z. Huajian, "Fault diagnosis of wind turbines based on a support vector machine optimized by the sparrow search algorithm," *IEEE Access*, vol. 9, pp. 69307–69315, 2021.
[24] X. Zhang, P. Han, L. Xu, F. Zhang, Y. Wang, and L. Gao, "Research on bearing fault diagnosis of wind turbine gearbox based on 1DCNN-PSO-SVM," *IEEE Access*, vol. 8, pp. 192248–192258, 2020.
[25] W. Chen, Y. Qiu, Y. Feng, Y. Li, and A. Kusiak, "Diagnosis of wind turbine faults with transfer learning algorithms," *Renew. Energy*, vol. 163, pp. 2053–2067, Jan. 2021.
[26] J.-Y. Hsu, Y.-F. Wang, K.-C. Lin, M.-Y. Chen, and J. H.-Y. Hsu, "Wind turbine fault diagnosis and predictive maintenance through statistical process control and machine learning," *IEEE Access*, vol. 8, pp. 23427–23439, 2020.
[27] D. Zhang, L. Qian, B. Mao, C. Huang, B. Huang, and Y. Si, "A data-driven design for fault detection of wind turbines using random forests and XGboost," *IEEE Access*, vol. 6, pp. 21020–21031, 2018.
[28] Y. Shi, Y. Liu, and X. Gao, "Study of wind turbine fault diagnosis and early warning based on SCADA data," *IEEE Access*, vol. 9, pp. 124600–124615, 2021.
[29] J.-Y. Hsu, Y.-F. Wang, K.-C. Lin, M.-Y. Chen, and J. H.-Y. Hsu, "Wind turbine fault diagnosis and predictive maintenance through statistical process control and machine learning," *IEEE Access*, vol. 8, pp. 23427–23439, 2020.
[30] J. Xie, Z. Li, Z. Zhou, and S. Liu, "A novel bearing fault classification method based on XGBoost: The fusion of deep learning-based features and empirical features," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–9, 2021.
[31] M. Tang, Q. Zhao, S. X. Ding, H. Wu, L. Li, W. Long, and B. Huang, "An improved LightGBM algorithm for online fault detection of wind turbine gearboxes," *Energies*, vol. 13, no. 4, p. 807, Feb. 2020.

LISHU WANG was born in Yushu, Jilin, China, in 1973. He received the B.S., M.S., and Ph.D. degrees, in 1996, 2003, and 2006, respectively. In 2006, he studied as a Senior Visiting Scholar at Tottori University, Japan. He is currently a Professor and the Doctoral Supervisor with the Institute of Electrical and Information, Northeast Agricultural University. He is also the Executive Director and the Deputy Secretary General of the Heilongjiang Agricultural Engineering Society. His research interests include agricultural electrification, dynamic analysis and control of power systems, and fault diagnosis and condition monitoring of renewable energy power equipment.

SHUHAN JIA was born in Shenyang, Liaoning, China, in 1998. She is currently pursuing the master's degree with the Institute of Electrical and Information, Northeast Agricultural University. Her research interest includes fault diagnosis of wind turbine.

XIHUI YAN received the B.S. and M.S. degrees from North China Electric Power University, Baoding, China, in 2017 and 2020, respectively. He is currently working at State Grid Shijiazhuang Electric Power Supply Company. His research interests include fault diagnosis and condition monitoring of power equipment, demand response, and non-invasive load monitoring.

LIBO MA was born in Shijiazhuang, Hebei, China, in 1994. He is currently pursuing the Ph.D. degree with the School of Electrical and Electronic Engineering, North China Electric Power University. His main research interests include fault diagnosis of power equipment and hydrogen storage applications in power systems.

JUNLONG FANG was born in Harbin, Heilongjiang, China, in 1971. He received the B.S., M.S., and Ph.D. degrees, in 1993, 2000, and 2006, respectively. He began to work at Northeast Agricultural University, in July 1993. From November 2006 to October 2007, he studied at Tottori University, Japan. He is currenty a Professor and a Ph.D. Supervisor at the Institute of Electrical and Information, Northeast Agricultural University. He is also the Executive Director of the Heilongjiang Electrical Engineering Society, the Heilongjiang Agricultural Engineering Society, and the Heilongjiang Automation Society; and a member of the Teaching Committee of Applied Undergraduate Electrical Engineering and Automation Subject of China Machinery Industry Education Association.