IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 26, 2019, accepted April 14, 2019, date of publication April 22, 2019, date of current version May 13, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2912621

Condition Monitoring of Wind Turbine Gearbox Bearing Based on Deep Learning Model

JIAN FU[1], JINGCHUN CHU[2], PENG GUO[1], AND ZHENYU CHEN[1]
[1]School of Control and Computer Engineering, North China Electric Power University, Beijing 102206, China
[2]Guodian United Power Technology Co., Ltd., Beijing 100039, China

Corresponding author: Jian Fu (huadianfj@ncepu.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 51677067.

ABSTRACT Wind turbines condition monitoring and fault warning have important practical value for wind farms to reduce maintenance costs and improve operation levels. Due to the increase in the number of wind farms and turbines, the amount of data of wind turbines have increased dramatically. This problem has caused a need for efficiency and accuracy in monitoring the operating condition of the turbine. In this paper, the idea of deep learning is introduced into wind turbine condition monitoring. After selecting the variables by the method of the adaptive elastic network, the convolutional neural network (CNN) and the long and short term memory network (LSTM) are combined to establish the logical relationship between observed variables. Based on training data and hardware facilities, the method is used to process the temperature data of gearbox bearing. The purpose of artificial intelligence monitoring and over-temperature fault warning of the high-speed side of bearing is realized efficiently and conveniently. The example analysis experiments verify the high practicability and generalization of the proposed method.

INDEX TERMS Adaptive elastic network, condition monitoring, deep learning, wind turbines.

I. INTRODUCTION

As the key component in mechanical equipment for connecting and transmitting power, gearbox plays an extremely important role in wind turbines. However, subcomponents in the gearbox are prone to various types of failures due to adverse working conditions and long term continuous operation. This affects the safety and reliability of the overall operation of the mechanical system. Not only will it lead to a decline in the quality of products or services, but also cause huge economic losses and casualties. Therefore, research on gearbox condition monitoring technology can identify fault modes in a timely and accurate manner and provide guidance for subsequent maintenance and repair. It is very important to ensure the safe and reliable operation of mechanical systems and to avoid the occurrence of major accidents [1] .

In the traditional wind turbine gearbox condition monitoring field, reference [2] monitors oil temperature by using data collected by the SCADA (Supervisory Control And Data Acquisition) system for gearbox fault detection. Reference [3] uses the motor current characteristics during shifting to monitor the health of the gearbox. Reference [4] uses an adaptive signal resampling algorithm with fault feature extraction and fault detector for time-varying frequency analysis. Reference [5] uses online sensors to study the remaining life of the gearbox and uses particle filter technology to explain it. The results summarize a good monitoring of lubricant degradation. Reference [6] proposes a method based Autocorrelation Time Synchronization Average (ATSA) to deal with the physical interaction between the ring and the gear in the gearbox. In reference [7] , a fiber refractometer is fabricated and the degradation of the gearbox synthetic lubricant is characterized. Physical models have an advantage in predicting long-term trends in components. However, some work have low precision and high time consumption.

In terms of statistical models, large amounts of data are often used to train fit models between inputs and outputs. Reference [8] analyzes the faults of different parts of the wind turbine through three different artificial neural network models. Reference [9] detects and classifies wind turbine faults by image texture analysis. Reference [1] introduces a variety of gearbox fault diagnosis models based on SCADA data. Reference [10] uses Deep Random Forest

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

Fusion (DRFF),sensors and accelerometers to extract monitoring data. The wavelet packet transform is used to extract the statistical parameters. Reference [11] uses an Artificial Neural Network (ANN) to combine SCADA data analysis systems based on temperature data and gearbox bearing fault detection. Reference [12] designs a solution based on linear Support Vector Machine (SVM) for wind turbine fault detection. Reference [13] uses guided waves and supervised learning classifiers to detect and diagnose wind turbine blades. Reference [14] and [15] used the Principal Component Analysis (PCA) and Multi-way Principal Component Analysis (MPCA) models to model normal units to evaluate faulty units. However, the traditional model algorithm is difficult to solve the huge data volume caused by the increase in the number of wind turbines and the observed variables. This leads to a series of problems such as over-fitting and low precision.

Different from traditional artificial intelligence algorithms, Deep Neural Networks (DNN) adopt parallel distributed structure which are adaptive, scalable and self-organizing. Based on huge training data and hardware facilities, DNN can accurately establish the logical relationship between multivariables. Reference [16] uses the Optimized Deep Belief Network to identify faults in the gearbox. Reference [17] uses wavelet neural network to diagnose wind turbine gearbox. Reference [18] applies multi-scale convolutional neural networks to the fault diagnosis of wind turbine gearboxes. References [19]–[21] have achieved better results in wind turbines fault identification and classification by using deep learning ideas. Therefore, great potential of deep learning idea in wind turbine monitoring is obvious.

This paper introduces idea of deep learning into wind turbine condition monitoring. DNN is used to solve the surge in data volume due to the increase in the number of wind farm units and observed variables. Efficiency and accuracy of online monitoring are improved too. For the gearbox bearing temperature, the Convolutional Neural Network (CNN) can quickly complete the characteristics of the whole process of feature extraction, dimension reduction and classification. CNN is combined with Recursive Neural Network (RNN) to establish model and use its huge neural network and multi-hidden layer to train data quickly and efficiently. At the same time, RNN makes good use of historical data information and accurately predicts future information. The model can eliminate the problems of gradient explosion, over-fitting and generalization of traditional neural networks. This method can not only check the working condition of the unit during most normal working hours, but also avoid the early warning caused by the accidental parameter change and play the role of condition monitoring.

The rest of this paper will be organized as follows. In section 2, the data is preprocessed. Then the paper performs modeling analysis in the 3-5 sections. Section 6 analyses the superiority of the model through data calculation. Finally, the high-speed and rapid warning and judgment of the over-temperature fault of the main bearing high-speed end proves that the monitoring method is of great significance for large-scale wind farms.

II. DATA PREPROCESSING

SCADA system records 47 parameters related to the operation of the wind turbine and there is correlation and inertia between the various variables. In general, the performance of a classifier increases with the number of feature variables. When a certain value is exceeded, the performance does not rise but fall. This depends on the number of training samples, the complexity of the decision boundaries, and the type of classifier. In fact, because the number of training samples is limited, it is inevitable to avoid řdimensional disasters ś. Therefore, it is necessary to select the appropriate variables related to the transmission bearing temperature as the research object before analyzing the data collected by SCADA. Variables with different correlations have different effects on gearbox bearing temperatures. For example, if the transmission oil temperature is too low, the high speed shaft bearing temperature is too high. The viscosity of the lubricating oil is high at low temperatures, and the oil passing through the oil inlet hole becomes very small, and the viscosity of the high viscosity oil is poor. The thermal conductivity is also much worse, resulting in higher and higher bearing temperatures and entering a vicious circle. The above situation is mainly reflected in the winter gearbox and water-cooled lubrication system. Therefore, the oil temperature of the gearbox occupies a considerable weight, and the wind speed has less influence on the bearing temperature due to its randomness and volatility. Therefore, these analyzed variables related to the observed parameters need to be selected for the study.

In this part, this paper uses the adaptive elastic network based on statistical theory to implement variable selection. The new statistical method not only inherits the sparsity characteristics of the elastic network, but also has a group selection ability with more reasonable related features. The adaptive elastic network is assigned to different weights according to the size of the variable coefficients in the framework of the elastic network regularization method. Variables with larger coefficients are assigned to smaller weights, while variables with smaller coefficients are assigned larger weights, so that important variables are retained and unimportant variables are removed. More preferably, the adaptive elastic network estimates have oracle properties, namely sparsity and asymptotic normality, to ensure optimality of the model.

The adaptive elastic network is weighted on the basis of the regularization combination, and the following regularization is obtained:

$$\mathrm{P}(\beta)=\lambda_1\sum_{j=1}^{s} w_j\left|\beta_j\right| + \lambda_2\sum_{j=1}^{s} w_j\beta_j^2 \tag{1}$$

In the formula, w_j is the penalty weight. $\lambda_1 \sum_{j=1}^{s} w_j \left|\beta_j\right|$ is the improved penalty function, and w_j is the key to the variable selection in the model.

The estimated value of the elastic network is:

$$l(\beta) = \sum_{i=1}^{n} [(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})] \quad (2)$$

$$\overset{\wedge}{\beta} = \arg\min \left\{ l(\beta) + \lambda_1 \sum_{j=1}^{p} w_j \left|\beta_j\right| + \lambda_2 \sum_{j=1}^{p} w_j \beta_j^2 \right\} \quad (3)$$

In the formula, β_0 is the intercept term and $\beta = (\beta_1, \beta_2 .... \beta_p)$ is the regression coefficient of the model. $(x_{i1}, x_{i2} ... x_{ip}; y_i)$ is n sets of observed variables. x is an n×p matrix$(i = 1, 2, ..n; j = 1, 2...p)$. In order to avoid the denominator being zero, the value of the weight becomes meaningless, and the weight coefficient is set to:

$$w_j = (\left|\overset{\wedge}{\beta_j}(PLS)\right|)^{-\gamma} \quad (4)$$

γ is a normal number, $\overset{\wedge}{\beta_j}(PLS)$ is the coefficient solution of the partial least squares method of the linear regression model. λ_1 and λ_2 are used for sparse estimation and group effect respectively. The modified PLS regression coefficient is chosen as the weighting factor to improve the drawbacks of the traditional adaptive elastic network that cannot filter out the important variables in the group.

Since the elastic net penalty used in the method is a convex combination of lasso penalty and ridge penalty, it not only achieves variable compression, but also avoids the problem of excessive compression of lasso penalty. At the same time, elastic net can not only solve the problem of strong correlation of data, but also solve the difficulty that Lasso penalty method can not select group variables. Through the selection of variables, the input scale of the model can be effectively reduced, and the generalization ability of the model can be improved. The relevant variables selected in this paper are shown in Table 1.

TABLE 1. The modeling parameters selected in this article.

Indicator	Wind turbine status parameter	Units
T1	gearbox high speed shaft side temperature	$^\circ C$
T2	gearbox low speed shaft side temperature	$^\circ C$
T3	gearbox oil temperature	$^\circ C$
T4	motor bearing temperature	$^\circ C$
T5	cabin temperature	$^\circ C$
T6	gearbox cooling water temperature	$^\circ C$
T7	hydraulic oil temperature	$^\circ C$
T8	control cabinet temperature	$^\circ C$
V	wind speed	m/s
P	active power	kw

When multiple variables are used for wind power prediction, the dimensions are different between different variables, and the numerical differences are also large. To avoid neuron saturation, the input and output range of the nonlinear activation function are considered in the model. The variables need to be normalized. The selected modeling variables are normalized by the annual statistical limit values, and their values are reduced to the interval [−1, 1].

$$x' = \frac{x - (x_{\max} + x_{\min})/2}{(x_{\max} - x_{\min})/2} \quad (5)$$

$x_{\max}$, $x_{\min}$ are the maximum and minimum values of the variable respectively. The wind power prediction data obtained by the prediction model is subjected to inverse normalization to make it have physical meaning. The inverse normalization calculation formula is:

$$x' = 0.5 + \left[x'(x_{\max} - x_{\min}) + (x_{\max} + x_{\min})\right] \quad (6)$$

After the data processing procedure presented in this section, the variables related to bearing temperature are effectively screened out. Subsequent modeling analysis will be presented in the next section.

III. CONVOLUTIONAL NEURAL NETWORK FEATURE EXTRACTION AND DIMENSIONALITY REDUCTION

This chapter introduces the idea of deep learning into the neural network through modeling the single unit and using historical data as a training sample. First, the collected turbine normal operation data is substituted into the convolution layer. Subsequently, the improved long short term memory network is used to train the reduced-dimensional data.

The essence of CNN is to construct multiple filters that can extract data features, and extract the hidden topological features between the data by layer-by-layer convolution and pooling operations on the input data. As the number of layers increases, the extracted features become more and more abstract. Finally, these abstract features are merged through the fully connected layer, and the classification problem and regression problem are solved by the ReLU (Rectified Linear Unit)activation function. Figure 1 shows the dimension reduction process of convolutional neural network. M represents the number of convolutional layers. K represents the number of fully connected layers. Generally it does not exceed 3.

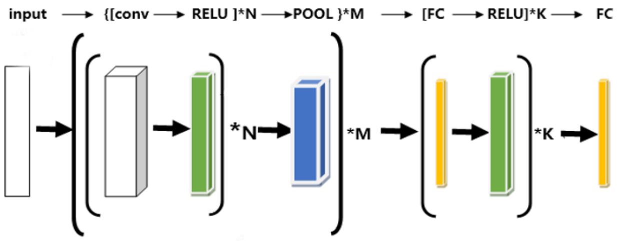

FIGURE 1. Schematic diagram of convolutional neural network dimension reduction.

The main feature of CNN is that it can extract the local features of the input data and combine the abstraction to generate advanced features layer by layer. As a multi-level neural network, its filtering level is used to extract features of the input signal, and the classification level classifies the learned features. At the same time, because the convolutional layer and the pooled layer of CNN can well reduce the dimensionality of multi-dimensional observation variables, it can quickly and accurately extract different features in a considerable amount of data. This is also the reason why this method has achieved remarkable results in the fields of image processing and speech recognition. Similarly, this method is also significant for wind turbine data processing.

The modeling method here draws on the idea that the CNN model is divided into multi-dimensional arrays and random sample selection for image processing. The ten variables selected by the above adaptive elastic network method are shown in Table1. At the same time, the gearbox temperature at the previous moment has a direct impact on the current temperature, because the parameter changes have a large inertia. Thus, ten adjacent historical moments are selected as a set of samples. The training sample input contains a total of 13,000 samples (1 minute -level data), and each set of samples is a 10×10 vector. D is the matrix setting for a single sample ($n = 1, 2, 3 \ldots 10$; $m = 1, 2, 3 \ldots .10$). Among them, abnormal data points such as the shutdown point and the test dead points have been removed. This vector contains 10 sets of data of 10 variables adjacent to each other at the historical moment. The output is the high speed shaft end temperature of the gearbox bearing. Figure2 shows the convolution model built in this paper.

$$D = [X(1)\,X(2)\ldots\ldots X(m)] = \begin{bmatrix} x_1(1)\,x_1(2)\ldots\ldots\,x_1(m) \\ x_2(1)\,x_2(2)\ldots\ldots x_2(m) \\ \vdots\ \vdots\ \vdots \\ x_n(1)\,x_n(2)\ldots\ldots\,x_n(m) \end{bmatrix}_{n\times m} \quad (7)$$

The single sample described above is a 10×10 matrix in the text. When substituting into the convolutional neural network, each batch randomly selects (10×10) samples for batch training. It has been proved by experiments that the selection of 100 samples in a single batch for training has the shortest experimental time and the highest accuracy. By calculating the gradient of the random sample, it is updated to the model weight once after the summation. Furthermore, the process of the convolution layer and the pooling layer reduces the dimension of the input variable. The specific input changes are shown in the table below. Table 2 shows the parameters configuration in the convolutional neural network. The boundary parts of the convolution operation are not filled.

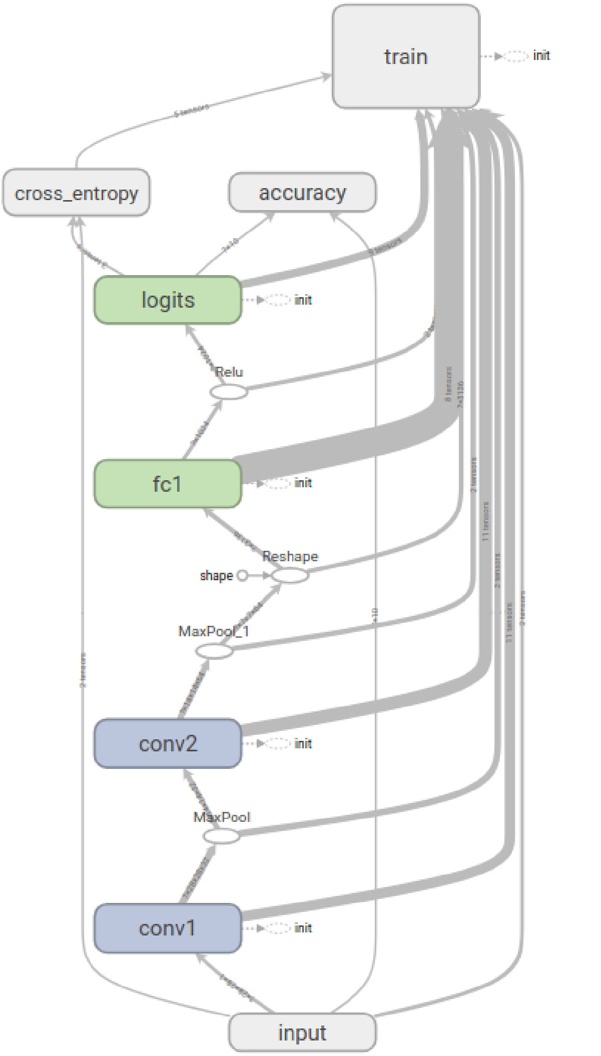

FIGURE 2. Schematic diagram of convolutional neural network structure in this paper.

TABLE 2. Convolutional neural network parameters configuration.

No.	Network layer	Kernel size	stride	Kernels	Output size
1	Convolution 1	11×11	1×1	32	88×88×32
2	Pooling 1	4×4	1×1	1	22×22
3	Convolution 2	5×5	1×1	16	18×18×16
4	Pooling 2	3×3	1×1	1	6×6
5	Fully connected	100	1	1	100×1
6	Softmax	10×1	1	1	10

IV. LONG SHORT TERM MEMORY NETWORK REGRESSION PREDICTION

Current wind farm sensors generally have high precision and resolution. This article takes 1-min level data as an example, while the data volume brought by the 1-second sampling interval will be larger. Therefore, proper dimensionality reduction is necessary. In this paper, after the training sample is substituted into the above process, the output of the fully connected layer is used as the input of LSTM. The output of the wind turbine gearbox bearing temperature is predicted as the output of LSTM. Through the dimensionality reduction and feature classification of the CNN, the dimension of the input substituted into the LSTM is greatly reduced. Moreover, under the premise of ensuring a sufficient number of samples, the weight of each variable affects the output of the output is trained. This is critical for establishing a logical relationship between multiple input and output variables as well as for further predicting output values.

The LSTM is an improved model of the recurrent neural network (RNN). The neurons in the hidden layer of the cyclic neural network have a feedback mechanism to realize the transmission of context information, so that the model has the ability to process sequence data. The cyclic neural network uses a certain length of sequence data as a training set, and predicts the output at the next moment after completing the training. The combination with historical data is especially important for predicting results. The BPTT (back propagation through time) algorithm is commonly used for training RNN. The central idea of BPTT is to continuously search for better points along the negative gradient direction of the parameters that need to be optimized until convergence. In essence, it is also based on the BP algorithm of the gradient descent method. Finding the gradient of each parameter is the core of this algorithm. As the time series continues to deepen, the multiplication of the decimals will cause the gradient to become smaller and smaller until it is close to zero, which in turn affects the accuracy of the model. That is the gradient disappearance problem. The improved LSTM based on RNN network can solve this problem well.

The training process of the LSTM includes a forward propagation process and a back propagation process. The forward propagation process of the LSTM means that the sequence X of length T is recursively calculated according to the time step from $t = 0$ to $t = T$. The back propagation process refers to the time from $t = T$ to $t = 0$. At the moment, the BPTT algorithm is used for reverse error propagation. The RNN initialization sets the state of all cells to 0, and the weight correction term δ is set to 0 at $t = T$.

Before analyzing the network training process, the following parameters are used: w_{ij} is the connection weight of node i to node j. a_j^t is the input of unit j at time t. b_j^t is the calculation value of the activation function of unit j at time t. s_j^t is the state values of unit j at time t. l,φ,w,c respectively represent the input port, the output port, the forgetting port and the storage unit. The connection weights of the storage unit to the input port, the output port and the forgetting port are w_{cl}, $w_{c\varphi}$, w_{cw} respectively. I, K and H represent the number of input nodes, output nodes, and LSTM units respectively. h is the number of other LSTM units connected to the LSTM unit. Different from the standard RNN, the number of inputs of the LSTM unit is more than the number of outputs. The input port, the output port and the forgetting port are the inputs of the LSTM unit, and only a single output is used by the network for other units. Therefore, the definition G represents the number of inputs of the hidden layer, including all inputs of the storage unit W and the LSTM port unit, and uses g to represent the input variable when it is not necessary to distinguish the input categories. $f(x)$, $g(x)$ represents the number of active faces, and l is the loss function at the time of training. Figure 3 is a comparison of RNN and LSTM.

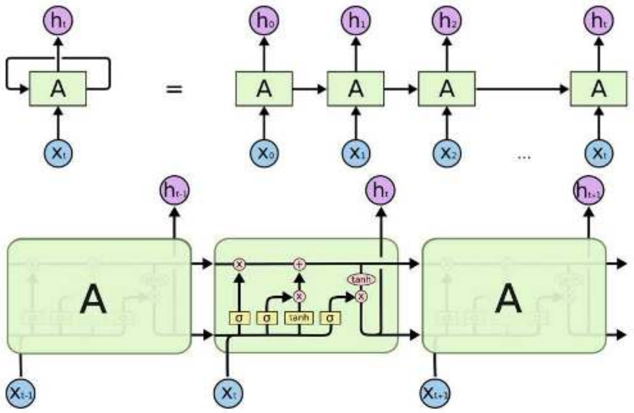

FIGURE 3. **RNN and LSTM comparison chart.**

A. FORWARD PROPAGATION PROCESS

The basic principle of the forward propagation process of the cyclic neural network is similar to that of the general neural network, except that the input of the hidden layer includes not only the input of the current moment but also the state of the network at the previous moment. For example, a sequence of length T is input to a cyclic neural network consisting of I input units, H hidden layer units, and K output units. Let x_i^t be the input i at time t, then the calculation of the hidden layer unit is as shown by the formula, where $\theta_{h(x)}$ represents the nonlinear activation function of the hidden layer.

$$a_h^t = \sum_{i=1}^{I} w_{ih} x_i^t + \sum_{h'}^{H} w_{h'h} b_{h'}^{t-1} \tag{8}$$

$$b_h^t = \theta_h(a_h^t) \tag{9}$$

The forward propagation process passes through the input port, the forgetting port, the storage unit, the output port, and the unit output in turn. Next, the formulas for each part are deduced separately.

The input a_l^t of the input gate is as shown in the formula, and the output is b_l^t.

$$a_l^t = \sum_{i=1}^{I} w_{il} x_i^t + \sum_{h=1}^{H} w_{hl} b_h^{t-1} + \sum_{c=1}^{C} w_{cl} s_c^{t-1} \tag{10}$$

$$b_l^t = f(a_l^t) \tag{11}$$

The input a_φ^t of the forgotten gate is shown in the formula, and the output is b_φ^t.

$$a_\varphi^t = \sum_{i=1}^{I} w_{i\varphi} x_i^t + \sum_{h=1}^{H} w_{h\varphi} b_h^{t-1} + \sum_{c=1}^{C} w_{c\varphi} s_c^{t-1} b_\varphi^t = f(a_\varphi^t) \tag{12}$$

The input a_c^t of the storage unit is shown in the formula, and the status value is s_c^t.

$$a_c^t = \sum_{i=1}^{l} w_{ic} x_i^t + \sum_{h=1}^{H} w_{hc} b_h^{t-1} s_c^t = b_\varphi^t s_c^{t-1} + b_l^t g(a_c^t) \quad (13)$$

B. BACKPROPAGATION PROCESS

After completing the forward propagation, the cyclic neural network uses the error back propagation method to update the weight. BPTT algorithms are simpler and more efficient in processing timing information. The loss function in the cyclic neural network depends not only on the influence of the hidden layer activation value on the output layer, but also on the influence of the hidden layer activation value on the hidden layer at the next moment. The weight correction term δ_h^t is as shown in the formula.

$$\delta_h^t = \theta'(a_h^t)(\sum_{k=1}^{K} \delta_k^t w_{hk} + \sum_{h'=1}^{H} \delta_{h'}^{t+1} w_{hh'}) \quad (14)$$

The complete sequence of the weight correction term is recursively calculated from the t=T time by the time step until the time t=0, and finally the weight correction term sequence is summed to obtain the final weight correction term. The backpropagation process is the inverse of the forward propagation process, so it goes through the unit output, output port, storage unit, forgetting port and input port in sequence. The calculation process is as follows:

ε_c^t and ε_s^t are the partial derivative of the loss function to the unit output and the state of the storage unit.

$$\varepsilon_c^t \overset{def}{=} \frac{\partial l}{\partial b_c^t} \varepsilon_s^t \overset{def}{=} \frac{\partial l}{\partial s_c^t} \quad (15)$$

The RNN backpropagation process first performs the calculation of the sum at the unit output.

$$\varepsilon_c^t = \sum_{k=1}^{K} w_{ck} \delta_k^t + \sum_{g=1}^{G} w_{cg} \delta_k^{t+1} \quad (16)$$

The RNN backpropagation process is followed by calculation of the weight correction term of the unit at the output port.

$$\delta_\gamma^t = f'(a_\gamma^t) \sum_{c=1}^{C} h(s_c^t) \varepsilon_c^t \quad (17)$$

The RNN backpropagation process arrives at the storage unit to calculate the weight correction term of the unit.

$$\varepsilon_s^t = b_\gamma^t h'(s_c^t) \varepsilon_c^t + b_\varphi^{t+1} \varepsilon_s^{t+1} + w_{cl} \delta_l^{t+1} + w_{c\varphi} \delta_\varphi^{t+1} + w_{c\gamma} \delta_\gamma^t \delta_c^t \quad (18)$$

The RNN backpropagation process calculates the weight correction term for the unit at the forgetting port.

$$\delta_\varphi^t = f'(a_\varphi^t) \sum_{c=1}^{C} s_c^{t-1} \varepsilon_s^t \quad (19)$$

The RNN backpropagation process finally reaches the input port to calculate the weight correction term of the unit, such as the formula:

$$\delta_\varphi^t = f'(a_l^t) \sum_{c=1}^{C} g(_c^t) \varepsilon_s^t \quad (20)$$

C. PREDICTIVE EFFECT EVALUATION

In this paper, the root mean square error (RMSE) and the mean absolute percentage error (MAPE) are used to evaluate the prediction results. ε_{RMSE} embodies the performance of the model to control absolute error. The calculation formula is:

$$\varepsilon_{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left[\hat{x}(i) - x(i) \right]^2} \quad (21)$$

$\hat{x}(i)$ and $x(i)$ are the actual and predicted values of the bearing temperature; n is the number of predicted verification data; i is the predicted point sequence number.

ε_{MAPE} can avoid the problem that the errors cancel each other out, and accurately reflect the magnitude of the actual prediction error. The smaller the calculated values of the two indicators, the smaller the prediction error of the reflection model and the higher the accuracy.

$$\varepsilon_{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{\hat{x}(i) - x(i)}{x(i)} \right| \times 100\% \quad (22)$$

V. MODEL ACCURACY IMPROVEMENT METHOD

With the deepening of the number of layers, the traditional neural network has three typical problems: the first is the non-convex optimization problem. That is, the optimization function is more and more easy to fall into the local optimal solution; the second is the gradient disappearing problem; the third is the over-fitting problem. The gradient disappearance problem has been solved by the improved LSTM model. The other two issues have also been effectively addressed in this article.

A. LEARNING RATE AND MOMENTUM ITEMS

Learning rate controls the update speed of parameters. The excessive learning rate is that the model oscillates around the optimal solution, while too small learning rate makes the model converge too slowly and falls into the local optimal solution.

Generally, in the early stage of training, it is desirable to use a slightly larger learning rate to quickly obtain a better solution; and in the later stage of training, it is desirable to use a smaller learning rate to make the model training more stable. A small learning rate can make the training track smoother, but it will cause the network training speed to be too slow or even not converged. A large learning rate can improve the training speed of the network, but it is easy to cause network training to be unstable or to fall into the local maximum. Therefore, the global optimal solution cannot be obtained. In order to solve the contradiction between learning rate and training stability, this paper introduces the momentum term to adjust, so that the correction amount of the parameter changes. The expression is:

$$\Delta\theta_t = \beta \times \Delta\theta_{(t-1)} + \alpha \times \frac{\partial \ln L}{\partial \ln \theta} \tag{23}$$

$\Delta\theta_t$ and $\Delta\theta_{(t-1)}$ are the corrections of the parameters in the t and $t-1$ iterations respectively. β is the momentum term coefficient. α is the learning rate; $\frac{\partial \ln L}{\partial \ln \theta}$ is the gradient of the current sample log likelihood function. According to the formula, after adding the momentum term, the correction amount of the parameter is not only related to the current gradient, but is determined by the update amount of the parameter in the previous iteration and the gradient in this iteration, so that the training process is more stable and avoids excessive fluctuations and local optimal solutions. According to experience, the momentum term coefficient generally takes a number between 0.5 and 0.9.

B. OVER-FITTING PROBLEM

Model over-fitting is due to excessive consideration of data correlation and noise data. Excessive pursuit of řsmall deviation ś makes the model too complicated, resulting in a small deviation of the fitted data distribution from the real distribution but the variance is too large. Thus, although the model fits the training data very well, it performs poorly on the test data (unknown data).Generally, the dropout layer is added after the fully connected layer to prevent over-fitting and improve the generalization ability of the model. However, for multiple hidden layer neural networks, the effect is not ideal. Figure 4 is a schematic diagram of standard neural net and after applying dropout.

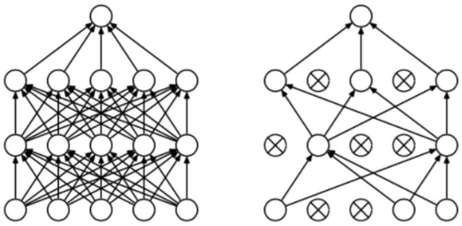

FIGURE 4. Standard neural net and after applying dropout.

Over-fitting problems often occur on complex models. By introducing a weight distribution that prevents over-fitting, the weight is adjusted as close as possible to loss function while adjusting the loss. One way to reduce the complexity of a network with a large number of weight parameters is to group the weights and then equalize the weights within the group. It integrates the network's translation invariance to the image into the network construction process. However, it only applies to the limited form in which the problem can be determined in advance. Therefore, the method of soft weight sharing is suitable for solving such problems. In this approach, hard limits with equal weights are replaced by a form of regularization, where the grouping of weights tends to take approximate values. In addition, the grouping of weights, the mean of each set of weights, and the range of values within the group are all determined as part of the learning process.

Gaussian mixture probability distributions are used to group weights.In the mixed distribution, the mean, variance, and mixing coefficient of each Gaussian component are determined as adjustable parameters during the learning process. Thus, there is a probability density of the form:

$$p(w) = \prod_i p(w_i) \tag{24}$$

$$p(w_i) = \sum_{j=1}^{M} \pi_j N(w_i \,|\mu_i\ , \sigma_j^2) \tag{25}$$

π_j is the mixing coefficient. By taking a negative logarithm of π_j, a regularization function can be get. The form isčž

$$\Omega(w) = -\sum_i \ln(\sum_{j=1}^{M} \pi_j N(w_i \,|\mu_i\ , \sigma_j^2)) \tag{26}$$

Thus, the total error function is:

$$\tilde{E} = E + \lambda\Omega \tag{27}$$

C. EFFECT DISPLAY

This part effectively shortens the training time and reduces the error. For the CNN-LSTM model used in this article, training time is significantly shortened under different training set lengths and test set lengths. At the same time, the error percentage is also effectively reduced. Table 3 intuitively reflects the superiority of the model improvement and fully embodies the necessity of model improvement. Model 1 represents the CNN-LSTM model before the accuracy improvement, and model 2 represents the model after the improvement.

TABLE 3. Comparison of the effects before and after the model improvement at forecast size of 1000.

Model	Training size	Training time(s)	Forecast time(s)	$\varepsilon_{MAPE}(\%)$
1.	2000	375.92	0.120	5.017
	5000	420.16	0.108	4.824
	10000	476.35	0.146	4.756
2.	2000	322.64	0.116	4.525
	5000	389.51	0.105	4.064
	10000	399.82	0.127	3.792

VI. CASE ANALYSIS

A. NORMAL DATA CASE ANALYSIS

In order to verify the effectiveness of the condition monitoring method, this paper takes the gearbox bearing temperature of a 1.5MW turbine A06 as the main research parameter. The 1-min level data recorded by the SCADA system from November 2012 to January 2013 are used for analysis, which include time, active power, wind speed, ambient and cabin temperature, gearbox bearing temperature and other parameters. Among them, abnormal data points such as the shutdown

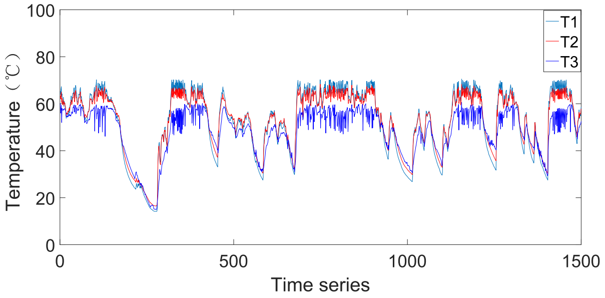

FIGURE 5. Schematic diagram of gearbox bearing temperature high speed shaft end (T1), low speed shaft end (T2), gearbox oil temperature (T3).

TABLE 4. The number of variables used by CNN-LSTM corresponds to the model of ε_{RMSE}(°C).

Prediction time series length	500	1000	3000
T1(1)	1.256	1.697	2.541
T1+T2+T3(3)	0.881	1.27	2.115
T1+T2...+T5(5)	0.825	1.265	2.101
T1+T2...T8+V+P(10)	0.781	1.17	2.023

point, the test dead points have been removed. At the same time, the SCADA system also saves the operating status information of the unit, such as unit start-up, shutdown, generator over-temperature, and pitch system failure. Figure 5 shows the sampled values for some main variables. Table 4 shows that the different numbers of parameters used for modeling affect the accuracy of the model. The numbers in parentheses represent the number of parameters involved in modeling. Conversely, the participation of parameters that are almost uncorrelated with gearbox bearing temperatures can increase the prediction error.

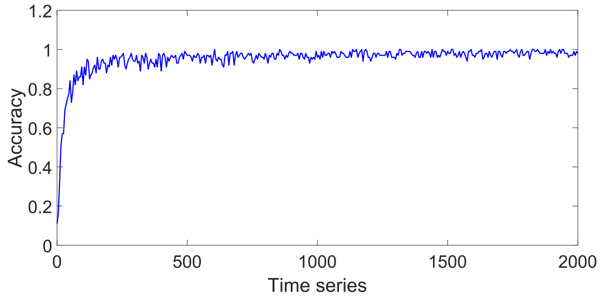

FIGURE 6. CNN accuracy value.

From the 1-min level data of 90 consecutive days collected by the SCADA system, matrixes consisting of ten variables and ten consecutive moments is taken as a sample input. A total of 13,000 sets of data are substituted into the CNN-LSTM model for training, and the gearbox bearing temperature is taken as the output. First, the model extracts and reduces the dimensions of the input data and then adaptively adjusts to form high quality parameters. Figure 6 and 7 show the value of the error function and accuracy the model. In this paper, the objective function threshold is 0.01 and the training accuracy is set to 99.5.

Subsequently, the dimensionally reduced data is substituted into a recurrent neural network for prediction. The chart below shows the prediction of the next 1000 data and the comparison with the actual data. It can be seen from the Figure 8 that the prediction results are almost similar to the actual results, which directly verifies the accuracy of the model. It can be seen from Table 5 that under the same evaluation criteria, CNN-LSTM has the smallest prediction error value and the best accuracy for time series prediction of different lengths. Obviously, the CNN-LSTM model used in this paper has the best accuracy for prediction after dimensionality reduction and feature extraction by convolutional neural network, which is compared with ARMA (Autoregressive moving average model) , DBN (Deep belief network) , BPNN (Back Propagation Neural Network) models. The four prediction methods selected as comparisons in this paper are both traditional and proven effective models. Among them, RNN and DBN are also one of the popular deep learning prediction models. By comparing with their predictions under the same trade-off indicator, the value and effectiveness of the work of this paper can be visually displayed. At the same time, this paper analyzes the different models of the same model in the same wind farm. Table 6 shows the ε_{MAPE} of A01, A02, A08, B02, B06, C01, C05, D02, D07, E05, E10, E16,

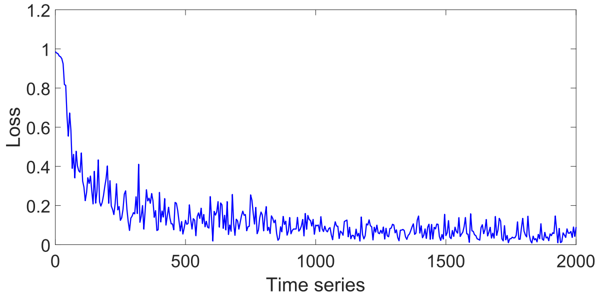

FIGURE 7. CNN objective function value.

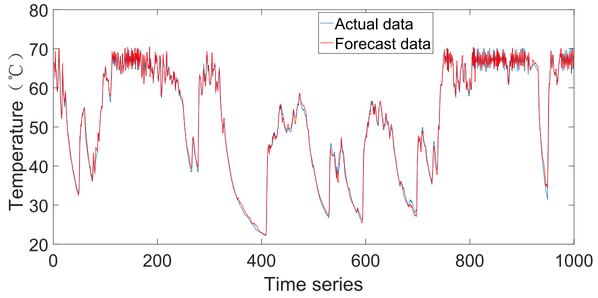

FIGURE 8. CNN-LSTM prediction results.

TABLE 5. ε_{RMSE}(°C)and ε_{MAPE}(%) corresponding to different models.

Prediction time series length	500	1000	3000
CNN-LSTM	0.781 3.675	1.17 4.064	2.023 4.623
RNN	1.325 4.618	1.845 5.329	3.325 6.683
DBN	1.431 4.814	1.931 5.462	3.628 7.513
ARMA	1.604 5.126	2.363 6.031	4.125 8.657
BPNN	1.856 5.398	2.728 6.653	4.687 9.926

TABLE 6. ε_{MAPE} **(%) corresponding to different turbines at prediction length 1000.**

A01	A02	A08	B02	B06	B10	C01	C02	D02
3.462	3.923	4.167	3.714	3.997	4.164	3.341	3.601	3.473
D07	E05	E08	E10	E16	F01	F06	G01	G04
3.227	4.009	3.491	3.348	3.401	3.429	3.308	3.728	4.335

F01, F06, G01 turbines. Obviously, the model still has good control error prediction ability for different turbines. It proves that the method has generalization and universal adaptability.

B. CASE ANALYSIS OF GEARBOX HIGH SPEED SHAFT OVER TEMPERATURE

According to the fault record, from January 2, 2013, the turbine A06's gearbox temperature was higher than P13.8 (70°C) for a period of time.P13.8 is the temperature point threshold of the actual field gearbox high-speed side under the cold season, which is usually set by the empirical method. Since the wind turbine bearing temperature is over temperature, the wind turbine will limit the power further before the rated power to lower the temperature and the power variation is relatively small. In this paper, combined with the above training model, the data of the 10 days before the failure occurred is used as input to predict the data for the next 5 days. The comparison between the predicted results and the actual data is reflected in Figure 9. The prediction results are almost the same as the actual results, which can reflect the accuracy of the model. At the same time, it can be seen from the figure that between the 100th and 400th points, the predicted gearbox bearing temperature is significantly higher than the normal unit data. This status reflects the abnormality of the gearbox bearing temperature.

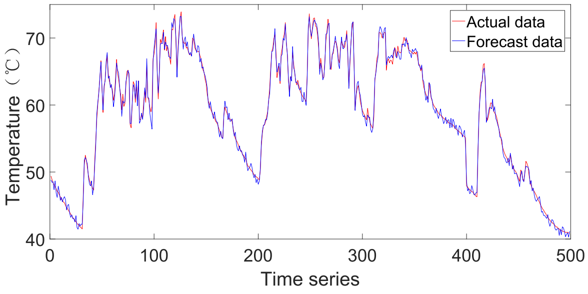

FIGURE 9. Comparison of forecast data and actual data(bearing temperature).

Combined with the fault log, the main cause of the fault is that the bearing is damaged and the roller is not running smoothly. In particular, when the high-speed shaft bearing has a high rotational speed, a large amount of heat is generated. Moreover, due to the interference friction of the parts and the excessive installation of the packing, a large amount of frictional heat is generated, which causes the bearing temperature to rise. This causes the device to restart. The brake level is 2, depending on the extent to which the fault occurred. Figure 10 reflects the bearing temperature comparison between the faulty turbine and the normal operating turbine.

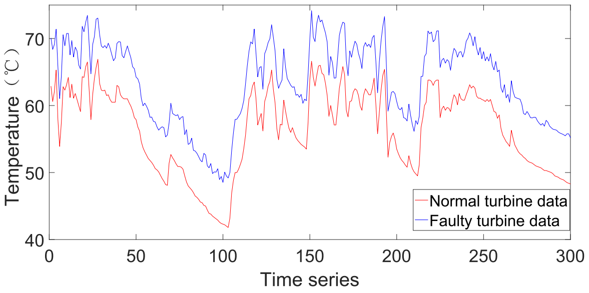

FIGURE 10. Comparison of faulty turbine forecast data with normal operating turbine data(bearing temperature).

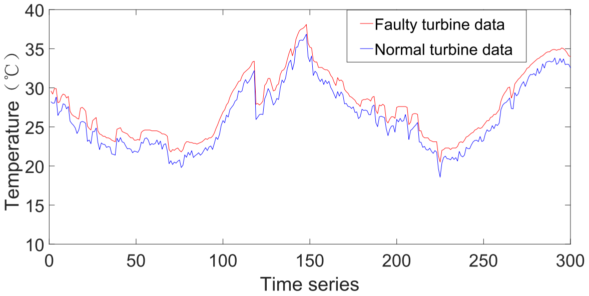

FIGURE 11. Comparison of faulty turbine data and normal operating turbine data(gearbox oil temperature).

At the same time, according to the Figure 11, it can be seen that the oil temperature of gearbox is lower compared with normal operating turbine, which is due to the lower temperature of the winter weather. Because the viscosity of the lubricating oil is very high at low temperatures, the oil passing through the oil inlet hole becomes very small. The fluidity of the high viscosity oil is poor, and the heat conduction ability is also much poor, resulting in higher and higher bearing temperatures, which fall into a vicious circle. This situation occurs mainly in the winter and on the gearbox of the water-cooled lubrication system.

On the whole, the cause of the failure is that the heat is too high due to wear of the bearing and the oil temperature of the gearbox is below the standard value. In this paper, the modeling of gearbox bearing temperature is a good predictor of unknown faults, which can reduce loss and repair equipment in advance.

VII. CONCLUSION

In this paper, this method is not blindly combining algorithms to analyze problems, but has practical results and significance. The following is a summary and outlook of the paper work:

1) Paper introduces the deep learning ideas into the modeling of wind turbine condition monitoring, and has achieved expected results in the actual wind farms.

2) The convolutional neural network used in this paper quickly complete the characteristics of the whole process of feature extraction, dimension reduction and classification. Its large neuron network and multiple hidden layers can quickly and efficiently train data. Combined with the recursive neural network in the good use of historical data information,

the CNN-LSTM model establishes the logical relationship between observed variables.

3) Many problems brought by traditional artificial intelligence algorithm are solved, such as gradient explosion, overfitting, which improves the accuracy of the model and reduce prediction error.

4) Good prediction result obtained in this paper can detect hidden dangers early and take preventive measures to improve the reliability of wind turbine operation and reduce maintenance costs.

The method used in paper solves the problem of surge in data volume due to the increase in the number of wind farm turbines and the observed variables, as well as the need for improved efficiency and accuracy of the turbine operating conditions monitoring. It has been effectively applied in actual wind farms and is of great significance for monitoring the turbines operation of large wind farms.

REFERENCES

[1] J. P. Salameh, S. Cauet, E. Etien, A. Sakout, and L. Rambault, ''Gearbox condition monitoring in wind turbines: A review,'' *Mech. Syst. Signal Process.*, vol. 111, pp. 251–264, Oct. 2018.

[2] Y. Feng, Y. Qiu, C. J. Crabtree, H. Long, and P. J. Tavner, ''Monitoring wind turbine gearboxes,'' *Wind Energy*, vol. 16, no. 5, pp. 728–740, 2013.

[3] I. Bravo-Imaz, H. D. Ardakani, Z. Liu, A. Garcia-Arribas, A. Arnaiz, and J. Lee, ''Motor current signature analysis for gearbox condition monitoring under transient speeds using wavelet analysis and dual-level time synchronous averaging,'' *Mech. Syst. Signal Process.*, vol. 94, pp. 73–84, Sep. 2017.

[4] D. Lu, W. Qiao, and X. Gong ''Current-based gear fault detection for wind turbine gearboxes,'' *IEEE Trans. Sustain. Energy*, vol. 8, no. 4, pp. 1453–1462, Oct. 2017.

[5] J. Zhu, J. M. Yoon, D. He, and E. Bechhoefer, ''Online particle-contaminated lubrication oil condition monitoring and remaining useful life prediction for wind turbines,'' *Wind Energy*, vol. 18, no. 6, pp. 1131–1149, 2015.

[6] J. M. Ha, B. D. Youn, H. Oh, B. Han, Y. Jung, and J. Park, ''Autocorrelation-based time synchronous averaging for condition monitoring of planetary gearboxes in wind turbines,'' *Mech. Syst. Signal Process.*, vols. 70–71, pp. 161–175, Mar. 2016.

[7] P. Sanchez *et al.*, ''Wind turbines lubricant gearbox degradation detection by means of a lossy mode resonance based optical fiber refractometer,'' *Microsyst. Technol.*, vol. 22, no. 7, pp. 1619–1625, 2016.

[8] M. Schlechtingen and I. F. Santos, ''Comparative analysis of neural network and regression based condition monitoring approaches for wind turbine fault detection,'' *Mech. Syst. Signal Process.*, vol. 25, no. 5, pp. 1849–1875, 2011.

[9] M. Ruiz *et al.*, ''Wind turbine fault detection and classification by means of image texture analysis,'' *Mech. Syst. Signal Process.*, vol. 107, pp. 149–167, Jul. 2018.

[10] C. Li, R.-V. Sanchez, G. Zurita, M. Cerrada, D. Cabrera, and R. E. Vásquez, ''Gearbox fault diagnosis based on deep random forest fusion of acoustic and vibratory signals,'' *Mech. Syst. Signal Process.*, vols. 76–77, pp. 283–293, Aug. 2016.

[11] Z.-Y. Zhang and K.-S. Wang, ''Wind turbine fault detection based on SCADA data analysis using ANN,'' *Adv. Manuf.*, vol. 2, no. 1, pp. 70–78, 2014.

[12] P. Santos, L. F. Villa, A. Reñones, A. Bustillo, and J. Maudes, ''An SVM-based solution for fault detection in wind turbines,'' *Sensors*, vol. 15, no. 3, pp. 5627–5648, 2015.

[13] A. A. Jiménez, C. Q. G. Muñoz, and F. P. G. Márquez, ''Dirt and mud detection and diagnosis on a wind turbine blade employing guided waves and supervised learning classifiers,'' *Rel. Eng. Syst. Saf.*, vol. 184, pp. 2–12, Apr. 2019.

[14] F. Pozo and Y. Vidal, ''Wind turbine fault detection through principal component analysis and statistical hypothesis testing,'' *Energies*, vol. 9, no. 1, p. 3, 2016.

[15] F. Pozo, Y. Vidal, and Ó. Salgado, ''Wind turbine condition monitoring strategy through multiway PCA and multivariate inference,'' *Energies*, vol. 11, no. 4, p. 749, 2018.

[16] Y. Qin, X. Wang, and J. Zou, ''The optimized deep belief networks with improved logistic sigmoid units and their application in fault diagnosis for planetary gearboxes of wind turbines,'' *IEEE Trans. Ind. Electron.*, vol. 66, no. 5, pp. 3814–3824, May 2019.

[17] L. Guo-Qi *et al.*, ''Fault diagnosis of wind turbines based on wavelet neural network,'' *J. Chin. Comput. Syst.*, vol. 36, no. 7, pp. 1504–1508, 2015.

[18] G. 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. 2019.

[19] L. Wang, Z. Zhang, H. Long, J. Xu, and R. Liu, ''Wind turbine gearbox failure identification with deep neural networks,'' *IEEE Trans. Ind. Informat.*, vol. 13, no. 3, pp. 1360–1368, Jun. 2017.

[20] H. Zhao, H. Liu, W. Hu, and X. Yan, ''Anomaly detection and fault analysis of wind turbine components based on deep learning network,'' *Renew. Energy*, vol. 127, pp. 825–834, Nov. 2018.

[21] F. Jia, Y. Lei, J. Lin, X. Zhou, and N. Lu, ''Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data,'' *Mech. Syst. Signal Process.*, vols. 72–73, pp. 303–315, May 2016.

JIAN FU was born in Nanyang, Henan, China, in 1996. He received the bachelor's degree from the School of Control and Computer Engineering, North China Electric Power University (NCEPU), Beijing, China, in 2017. He is currently pursuing the master's degree in control engineering with NCEPU. His research interests include condition monitoring and fault diagnosis for wind turbines, and data analysis based on deep learning.

JINGCHUN CHU was born in Inner Mongolia, China, in 1976. He received the Ph.D. degree from North China Electric Power University, Beijing, China, in 2004. He is currently with Guodian United Power Technology Co., Ltd. His research interests include wind power generation technology, the research and development of low wind speed wind turbine, offshore wind turbine, intelligent wind power technology, and tidal power generation technology and equipment.

PENG GUO was born in 1975. He received the master's and Ph.D. degrees from North China Electric Power University, in 2001 and 2004, respectively. His research interest includes new energy generation technology.

ZHENYU CHEN was born in Yongcheng, Henan, China, in 1995. He received the B.Eng. and M.Eng. degrees in control engineering from the School of Control and Computer Engineering, North China Electric Power University (NCEPU), Beijing, China, in 2016 and 2018, respectively. He is currently pursuing the Ph.D. degree in control theory and engineering with NCEPU. His research interests include wind turbine control and operation optimization, and wind farm power control and optimization.