International Journal for Housing Science and Its Applications
Publish August 10, 2025. Volume 46, Issue 4 Pages 5192-5205

https://doi.org/10.70517/ijhsa464360

# Data mining and fault diagnosis of multi-dimensional condition monitoring data for wind turbines based on machine learning

**Shuqiao Chen[1,*], Peng Zhang[1], Hui Ma[1] and Shuo Zhou[1]**
[1] Mengdong Concord Zalutqi Wind Power Co., Ltd., Tongliao, Inner Mongolia, 029100, China
Corresponding authors: (e-mail: lyj005566@126.com).

**Abstract** Wind power, as an important part of clean energy, plays a key role in the global energy transition. However, wind turbines operate in harsh environments for a long time, and equipment failures occur frequently, which seriously affects power generation efficiency and economic benefits. Aiming at the difficulty of fault identification under complex working conditions of wind turbines, this study proposes a multi-dimensional anthropomorphic condition monitoring method based on CEEMDAN-TCN. The method firstly adopts fully adaptive noise ensemble empirical modal decomposition to decompose the signals of the unit operation data to eliminate the modal aliasing phenomenon, and then utilizes time-domain convolutional network to predictively model the decomposed intrinsic modal components and combines with adaptive crag analysis to realize the fault feature extraction. The experimental results show that the proposed method triggers the alarm 2 h 17 min, 55 min, and 1 h 13 min ahead of time compared with CNN, LSTM, and GRU models, respectively, in gearbox fault warning, and the prediction accuracy is significantly improved. In the pitch system fault diagnosis, the pitch power ratio in the fault state crosses the range of 0.5-2.0, while the normal state is only 1.0-2.0. The method effectively solves the problem of misjudgment and omission of the traditional method through deep mining of spatio-temporal correlation information, and provides a reliable technical support for the intelligent operation and maintenance of wind turbines.

**Index Terms** Machine Learning, Multi-dimensional Mimicry, Condition Monitoring, Fault Identification, CEEMDAN-TCN, Adaptive Cliffiness

## I. Introduction

As a renewable energy source, the use of wind energy for power generation can not only reduce the consumption of resources and alleviate China's resource constraints, but also greatly reduce the pollution caused by the environment, and make a great contribution to the promotion of China's energy consumption structure [1]-[3]. Wind turbine is the core equipment for wind power generation, which mainly converts kinetic energy into mechanical energy, and then converts mechanical energy into electrical energy [4], [5]. However, due to the location of mostly some remote areas and high mountains, the harsh natural environment, variable wind speeds and the unstable long-term impact of external loads on the internal components of the wind turbine can easily cause failures, especially the main failures of the three parts of the gearbox, generator, and inverter [6]-[9]. Therefore, wind turbine condition monitoring and fault identification are of great significance for its safe and stable operation.

Detection as well as fault diagnosis of wind turbines is a crystallization of artificial intelligence through the integration of several systems such as computer systems, electrical systems, control systems, etc. Numerous wind farms in China can be integrated into a single monitoring system, and in a single monitoring system it can be detected whether or not the power plants across the country are operating normally [10]-[13]. We need to collect fault data, develop generator set components suitable for power generation under local environmental conditions according to different conditions in different regions, effectively solve the occurrence of faults fundamentally, improve the service life of parts, increase the cycle of power generation, and combine advanced technology to improve the accuracy of monitoring technology, so that the occurrence of faults can be dealt with in a timely manner, and China's monitoring and fault diagnosis technology for generator sets can be improved to a greater extent [14]-[17].

In this study, a multi-dimensional anthropomorphic condition monitoring model is constructed by integrating signal decomposition technology and deep learning method, which realizes real-time health assessment of key components of wind turbine. Firstly, the CEEMDAN algorithm is used to adaptively decompose the complex unit operation signal, effectively eliminate the modal aliasing problem and extract the intrinsic modal components of different frequency features, then use the parallel computing advantage of TCN network and the long-term dependence modeling ability to predict and analyze each component, and finally combine the adaptive kurtosis analysis method to realize the accurate extraction and classification and identification of fault features, so as to establish a complete technical system from data preprocessing to fault diagnosis.

# II. Construction of wind turbine condition monitoring model based on multi-dimensional mimicry

## *II. A. Wind turbine mechanism and SCADA system*

Wind turbines are devices that convert wind energy into electrical energy, and because of their more complex construction, they are highly susceptible to abnormal failures. In order to ensure their smooth operation at high altitude, the manufacturing process, material structure, and control strategy of wind turbines are extremely critical, so in order to better diagnose and monitor the failures of wind turbines, it is particularly important to have an understanding of the structure and mechanism. In addition, most of the wind farms have been managed by installing SCADA systems, which store a large amount of historical data information for fault diagnosis. This paper briefly describes the mechanism of wind turbines and SCADA systems as a basis for research on fault diagnosis.

### II. A. 1) Introduction to wind turbines

(1) The main structure and principle of wind turbines

The internal structure of wind turbine mainly contains blades, hubs, gearboxes, generators, nacelles, towers and so on. It includes nine key components, including “wind wheel system, tower, nacelle, transmission system, generator, pitch system, yaw system, hydraulic system and control system”.

The working principle of wind turbine is the process of energy conversion, wind energy through the impeller into mechanical energy to complete the first conversion, and then the energy through the transmission system in the main shaft, gear box and flexible coupling to the generator, the generator will eventually be converted into mechanical energy to complete the second conversion of energy, and finally through the transformer and other appropriate equipment to the power grid feed.

(2) Classification of wind turbines

According to the classification of grid-connected: divided into off-grid and grid-connected two types, the difference between the two is whether or not they are directly connected to the grid, off-grid type independent operation and do not access the grid, applicable to the place where the power consumption is relatively small, the scope of application is small.

### II. A. 2) Common forms of failure of wind turbines

(1) Common Failure Forms

Wind turbine structure is complex, and most of the wind farms are built in remote areas, the natural conditions are relatively harsh, by the rain, snow, wind and sand and other extreme weather, coupled with the design height of the tower is rising, it is very susceptible to gusts of wind brought about by the impact of the load, as well as a variety of loads generated during operation. Wind turbine failure types are more, roughly divided into two categories, one is mechanical failure, the other is electrical failure.

(2) Wind turbine blade icing failure

Blade icing includes two kinds of icing in the cloud and precipitation icing, and icing in the cloud can be divided into two kinds of freezing rain and freezing fog, while precipitation icing is mainly divided into freezing rain and snow and frost. This paper mainly through the deep learning method, fully utilizes the wind turbine data collected in the SCADA system to analyze and model, and finally completes the diagnosis of icing faults.

### II. A. 3) Introduction to SCADA systems

At present, most of the wind farms have installed SCADA system [18] to maintain and manage the data of the wind turbines in the wind farm, especially for large wind farms with hundreds of units, the application of this system greatly improves the management efficiency. The SCADA system mainly consists of the wind turbines in the wind farm, the upper computer, the lower computer, the communication line, the data acquisition and the monitoring equipment, etc. The data acquisition equipment in the wind farm will obtain the operating information of the units in real time. The data acquisition equipment of the wind farm will make timely and accurate acquisition of the operation information of the wind turbine in real time, and the collected data include wind speed, power, blade angle, blade speed and acceleration, temperature, etc., which can be generally categorized into two major categories, namely, discrete quantity and continuous quantity.

Discrete quantity refers to the two different states represented by 0 and 1, and is mainly collected from the generator, yaw, lubrication and hydraulic system; while continuous quantity refers to the numerical values in a continuous period of time to show the trend of the unit's performance, including the temperature, speed and pressure of the unit's key position and other parameters.

The most important role of the monitoring center is to monitor the operating status of the unit and to provide alarms for faults, and to provide timely alarms to the manager of the electric field when faults occur.

In this paper, based on SCADA data and under the premise of analyzing the mechanism of wind turbine, the data set is preprocessed and features are extracted, and a model is built using the deep learning method to realize timely and accurate early warning of faults.

## *II. B.Acquisition and Processing of Key Data for Wind Power Systems*

### II. B. 1) Data collection

In this section, the raw operational data from February to April 2024 for a particular wind turbine in a wind farm in Sichuan is analyzed. In the process of data collection, the data reaches tens of thousands or even hundreds of thousands. If every data is used as an analysis sample, it will take a long time, so it is very necessary to sample the data. In this paper, the systematic sampling method is adopted to carry out. There are dozens of wind turbine operation data such as time, average active power, wind speed, turbulence intensity, etc. Since this paper studies the mining of temperature data, the data related to temperature is selected for analyzing and mining. The extracted data are time, wind speed, average active power, air temperature, gearbox speed, gear oil temperature, gearbox inlet temperature, gearbox bearing temperature, and nacelle temperature.

### II. B. 2) Data quality analysis

(1) wind power introduction

The output power of the wind turbine is related to the size of the wind speed at the hub height, air density, the diameter of the wind wheel, the wind energy utilization factor, transmission efficiency and mechanical efficiency, and the relationship between power and each variable is:

$$P=0.5\rho SV^3C_p\eta_t\eta_g \tag{1}$$

where $P$ denotes the output power of the wind turbine, kW; $\rho$ denotes the air density, $kg/m^3$; $S$ denotes the wind turbine swept area, $m^2$; $C_p$ denotes the wind energy utilization coefficient, generally between 0.2 and 0.6, with a maximum of 0.55; $\eta_t$ denotes the mechanical efficiency of the wind turbine drive unit; $\eta_g$ denotes the mechanical efficiency of the generator. $V$ denotes the hub height wind speed, $m/s$; Eq. (2) denotes the wind turbine swept area, and Eq. (3) is the final wind power calculation method:

$$S=\pi\left(\frac{D}{2}\right)^2 \tag{2}$$

$$P=\frac{1}{8}\pi\rho D^2V^3C_p\eta_t\eta_g \tag{3}$$

The proportionality between the energy obtained from the wind and the energy contained in the wind is known as the Bates power coefficient.

(2) Analysis of outliers

Wind turbines are greatly affected by climate, and wind speed and air temperature will affect their power generation capacity.

Icing: the wind turbine studied in this paper is constructed in a wide area, the air humidity difference is large, every winter, the temperature is lower than 1 ℃, the blade will freeze, the unit relative to the sunny state power generation capacity is greatly reduced, at this time, wind speed, temperature, power and other data can be regarded as anomalies, can be eliminated.

Wind speed: wind turbine generator in the range of 2.5m/s-28m/s can be normal generator, lower than 2.5m/s when the unit is in standby mode, 15m/s to meet the conditions of full generation, the generator output power of 1550kW, the maximum instantaneous power can be up to 1590kW, the wind speed of more than 15m/s and less than 28m/s generator continues to maintain the output power in the 1550kW, when the wind speed is greater than or equal to 28m/s, the generator cuts out and the unit stops running. Therefore, in the theoretical power calculation, the power can not be increased with the increase of wind speed has been increased.

Power: When drawing the wind speed power curve, when the wind speed is more than 5m/s, the corresponding power value differs greatly from the theoretical value.

### II. B. 3) Data pre-processing

(1) Data screening

When the temperature of the wind farm studied in this paper is lower than 1°C, the blades are basically in the icing state, so the data with the temperature lower than 1°C can be excluded. After screening, the data were transformed from a table of 7925×10 to a table of 6473×1 with 10 variables. Variables in the data that were below 1°C were eliminated.

(2) Outlier processing

Theoretical power can be calculated based on the wind speed in the raw data, and the wind speed and power in the raw data can be fitted to a curve, and since it is a smoothing process for the cluttered data, the smoothing spline interpolation method is used. The fitted curve is evaluated by the following three metrics to evaluate it.

Sum Square Error SSE: This metric calculates the sum of the squares of the errors between the fitted data and the points corresponding to the original data:

$$SSE = \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2 \tag{4}$$

where $w_i$ is the weight, $y_i$ is the actual power, and $\hat{y}_i$ is the estimated power.

Root Mean Square RMSE: This metric calculates the fitted standard deviation of the regression system:

$$RMSE = \sqrt{\frac{SSE}{n}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2} \tag{5}$$

where $w_i$ is the weight, $y_i$ is the actual power, and $\hat{y}_i$ is the estimated power.

Coefficient of Determination:The goodness of a fit is indicated by the variation of the data, which is determined by two other metrics, SSR and SST.SSR:The sum of squares of the difference between the mean of the predicted data and the mean of the original data. I.e:

$$SSR = \sum_{i=1}^{n} w_i (\hat{y}_i - \overline{y}_i)^2 \tag{6}$$

where $w_i$ is the weight, $\hat{y}_i$ is the estimated power, and $\overline{y}_i$ is the mean power.

SST: the sum of squares of the difference between the raw data and the mean:

$$SST = \sum_{i=1}^{n} w_i (y_i - \overline{y}_i)^2 \tag{7}$$

where $w_i$ is the weight, $y_i$ is the actual power, and $\overline{y}_i$ is the average power.

The coefficient of determination R-square is the ratio of SSR to SST:

$$R - square = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST} \tag{8}$$

The fitted curve neutralizes the variance $SSE = 1.046e + 08$, and although it does not converge to 0, the curve is closer to the actual result.

Residual: uses the difference between the actual power and the estimated value as the observed value of the error:

$$\overline{\sigma}_i = y_i - \hat{y}_i \tag{9}$$

where $\overline{\sigma}_i$ is the residual, $y_i$ is the actual power, and $\hat{y}_i$ is the estimated power.

The nonparametric residuals, let the estimated value equal the average of the bootstrap values. To wit:

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{10}$$

where $x_i$ is the new data generated, $\overline{X}$ is the average of this row of data, and $n=1000$.

At this point the nonparametric residuals can be expressed as:

$$\theta_i = y_i - \overline{X}_i \tag{11}$$

where $\theta_i$ is the new residual, $y_i$ is the actual power, and $\overline{X}_i$ is the average of the $i$ th row of the newly generated data set.

To detect outliers, use the $normplot(\theta_i)$ statement to see if the new residuals exist. The first step in processing these outliers is to first find the outliers, which can be accomplished by absolute value processing, using the abs() statement to make the residuals all positive:

$$a_i = \left| y_i - \overline{X}_i \right| \tag{12}$$

where $a_i$ is the absolute value of the new residuals, $y_i$ is the actual power, and $\overline{X}_i$ is the average of the $i$ th row of the newly generated data set.

The threshold value can determine the outliers in the original data:

$$th = 1.1S = 1.1\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{X}_i)^2\right)^{\frac{1}{2}} \tag{13}$$

where $th$ is the threshold, $y_i$ is the actual power, and $\overline{X}_i$ is the average of the $i$ th row of the newly generated data set.

## *II. C.CEEMDAN-TCN based condition monitoring model for generator sets*

### II. C. 1) Fully adaptive noise ensemble empirical modal decomposition

Empirical Modal Decomposition (EMD) is a new method of time-frequency analysis, the method is based on the variation of the data itself, which is only related to the sampling frequency, and it is very effective in dealing with non-smooth and non-linear signals. However, the method has a more important drawback, which is the modal aliasing problem. In order to suppress the phenomenon of modal aliasing, ensemble empirical modal decomposition (EEMD), complementary ensemble empirical modal decomposition (CEEMD) [19], and CEEMDAN have been successively proposed. Compared with other methods, CEEMDAN adds finite adaptive white noise at each decomposition, which solves the problem of difficult component alignment and minimizes the noise residual in the final reconstructed signal. Therefore, in this paper, CEEMDAN is used to decompose the complex PDI curves of the unit to obtain simple intrinsic modal components (IMFs) with different frequency characteristics, which reduces the difficulty of prediction. The steps of the CEEMDAN algorithm implementation are as follows:

(1) Gaussian white noise $n_i(t)$ is added to the original signal $x(t)$ to get the noise-containing signal $x(t)+\gamma_0 n_i(t)$, $\gamma_0$ is the noise coefficient, the noise-containing signal is decomposed into $n$ components by EMD, and the first $\overline{IMF}_1(t)$ is obtained by taking the mean value of the decomposed modal components, the The expression is:

$$\overline{IMF}_1(t) = \frac{1}{n}\sum_{i=1}^{n} IMF_1^i(t) \tag{14}$$

where $IMF_1^i(t)$ is the first-order modal component obtained from the $i$ th decomposition. The residual signal is $r_1(t) = x(t) - \overline{IMF}_1(t)$.

(2) Repeat the first step with $r_1(t)$ as the original signal to obtain the second eigenmode component $\overline{IMF}_2(t)$ with the expression:

$$\overline{IMF}_2(t) = \frac{1}{n}\sum_{i=1}^{n} E_1\{r_1(t) + \gamma_1 E_1[n_i(t)]\} \tag{15}$$

where $E_1$ is the first order IMF operator obtained by decomposition. At this time, the residual signal is expressed as $r_2(t) = r_1(t) - \overline{IMF}_2(t)$.

(3) Repeat the above steps until the EMD stopping condition is satisfied, at which point the original signal $x(t)$ is:

$$x(t)=\sum_{i=1}^{k}\overline{IMF}_i(t)+r_i(t) \tag{16}$$

where $k$ is the highest order obtained from the decomposition.

### II. C. 2) Time-Domain Convolutional Networks

TCN is a network structure for sequence data, and its main feature is to utilize the idea of convolutional neural network to process time series data. Compared with the traditional recurrent neural network, the convolutional layer of TCN can be computed in parallel, which improves the speed of the model to process long time series, and at the same time, the model can capture the long term dependencies that exist in the sequence, which alleviates the gradient vanishing and gradient explosion problems faced by recurrent neural networks to a certain extent. Due to its outstanding performance, TCN is widely used in wind power prediction, electric load prediction, rolling bearing remaining life prediction, and lithium-ion battery remaining life prediction. In this paper, TCN is used to predict the components obtained from decomposition.

### II. C. 3) CEEMDAN-TCN based prediction models

The model utilizes CEEMDAN to decompose the PDI curves of pumped storage units to obtain multiple IMFs, and then a TCN network [20] is used to predict each IMF separately, and finally, the predicted values of the IMFs are reconstructed to obtain the prediction results of the deterioration trend. The TCN model used is set up as a three-layer model, with each layer containing a two-layer inflated convolution, weight normalization, ReLU activation function, and Dropout block. The expansion coefficient of the inflated convolutional network in each convolutional block is $2n-1$ and $n$ is the $n$ th layer of the TCN model.

In order to verify the accuracy of the proposed method for the prediction of unit PDI curves, the mean square error (MSE), root mean square error (RMSE), and mean absolute error (MAE) are used to quantitatively evaluate the prediction results, assuming that $x$ and $x'$ are the true and predicted values, respectively, and the formulas for the three evaluation indexes are as follows:

$$MSE=\frac{1}{N}\sum_{i=1}^{N}\left(x_i-x_i'\right)^2 \tag{17}$$

$$RMSE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i-x_i'\right)^2} \tag{18}$$

$$MAE=\frac{1}{N}\sum_{i=1}^{N}\left|x_i-x_i'\right|^2 \tag{19}$$

## *II. D.Based on wind turbine condition and fault diagnosis*

Adaptive Kurtosis method for different signal conditions, for different fault history data and normal data indicators for comparison, so as to derive a specific characterization, in order to further fault nodes of the fault when the domain of time to analyze, which will produce a different analysis of the indicators. Through experiments in the short-time Fourier transform (STFT) through the application of the spectral crag method concluded that the spectral crag method can effectively determine the noise unstable signal. This method can analyze the frequency bands after time-frequency domain processing in detail and determine the location of the largest frequency band.

Combined with the autoregressive model through the autoregressive signal prediction not only can effectively determine the faulty signal path and crag value faults in the time domain of the correlation data, for the residual signal can be a timely response.

(1) Principle of autoregressive model

The application of autoregressive model features determine the sum of input and output relationship, thus realizing linear analysis, due to the autoregressive model application can be predicted and evaluated on the complex channel, so as to guarantee the fault response rate. Since the autoregressive model is due to time change has a close feature coverage, by analyzing the mathematical model, the intrinsic structure of the data in the sequence will be better understood, which will allow the minimum variance to be predicted effectively. We assume that $x$ to obtain a stable signal sequence is to be achieved by the zero mean, if its length is taken as: $N$ , and the autoregressive model order is expressed as $p$, therefore, we can express the autoregressive model of $y$ as:

$$y_k = -\sum_{i=1}^{p} a_i x_{i+k} + \eta_k \tag{20}$$

For the analysis of the non-stationary signal link and the stable signal link, the residual value difference between the two is more obvious, most of the more obvious residual values in the non-normal or faulty data acquisition in the noise signal can be obviously collected.

(2) Optimal order determination

Cliff as a relative statistical index, mainly based on the unit vibration signal distribution characteristics for real-time monitoring, usually due to the vibration signal distribution characteristics of the absence of quantitative outline, belongs to the time domain indicators. For this reason, for the given discrete vibration signals, the coefficient of crag can be defined in terms of $K$ as follows, which can be obtained as Eq:

$$K = \frac{1}{N}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_2}\right)^4 \tag{21}$$

$x_i$ primarily represents the discrete vibration signal, $\bar{x}$ represents the signal at the sampling average, $N$ represents the sampling duration, and $\sigma_2$ represents the signal standard deviation. The magnitude of faults occurring in the system can be represented by utilizing the magnitude of the cliff value.

For the characteristic vectors of the electronic control system, which can be constructed from statistical characteristic parameters, the identification of the spectral magnitude values can be realized by real-time monitoring of these signal links relative to the time-domain signals. Combined with the vibration signals, multiple frequency band index parameters in the time domain can be assembled to realize the analysis of the characteristics of the electronic control system.

Based on the analysis of different frequency band signal sequences in the time domain, the maximum peak, mean and average amplitude levels can be calculated by using the following statistical indicators with scales:

a) Maximum Peak

The maximum peak is analyzed for the maximum value expressed in the time domain of the signal. The maximum peak reflects the relationship between the strength and size of the signal in the time domain, which is described and expressed as follows:

$$\hat{x} = \max(|x(n)|) \tag{22}$$

b) Mean Value and Average Magnitude

The mean value as a component value in the signal analysis process is mainly derived from the time domain signal estimation. It can be expressed as follows:

$$\bar{X} = \frac{1}{N}\sum_{n=1}^{N} x(n) \tag{23}$$

In the above equation, $N$ denotes the number of sequence points of the discrete signal.

The expression for the average amplitude is:

$$\overline{X_p} = \frac{1}{N}\sum_{n=1}^{N} |x(n)| \tag{24}$$

## III. Multi-dimensional mimetic wind turbine condition monitoring and fault identification analysis

### *III. A. Wind turbine gearbox condition monitoring model validation and analysis*

In order to verify the effectiveness of the proposed method, SCADA monitoring data collected from an actual wind farm is used in this paper. The dataset used is from the SCADA monitoring system of a wind turbine at a wind farm in Shaanxi, China, with a time range of 2 MW rated power of the wind turbine under test.The SCADA system that comes with the wind turbine at the wind farm was alarmed on February 4, 2024 at 10:24:00, and after on-site maintenance by the maintenance personnel, it was found that the wind turbine's gearbox had a fault of a damaged temperature-control valve, and the system under which it belonged was the gearbox The system was lubricated and cooled, and the system was subsequently maintained for 45 days and 14.5 hours for this purpose. In this section, the first 100,000 minutes of the alarm time node of this turbine are extracted for instance validation, where the first 60,000 minutes of data are the training set and the last 40,000 minutes are the test set. Before utilizing this data for instance validation, it is necessary to select the monitoring variables that can represent the wind turbine operation status through SCADA data variable selection. Since most of the monitoring quantities cannot be directly used in the wind turbine gearbox, this paper selects the monitoring quantities that are closely related to the status of the gearbox, and discards the rest of the monitoring quantities. The further selected monitoring variables of the SCADA system for wind turbine gearboxes are shown in Table 1.

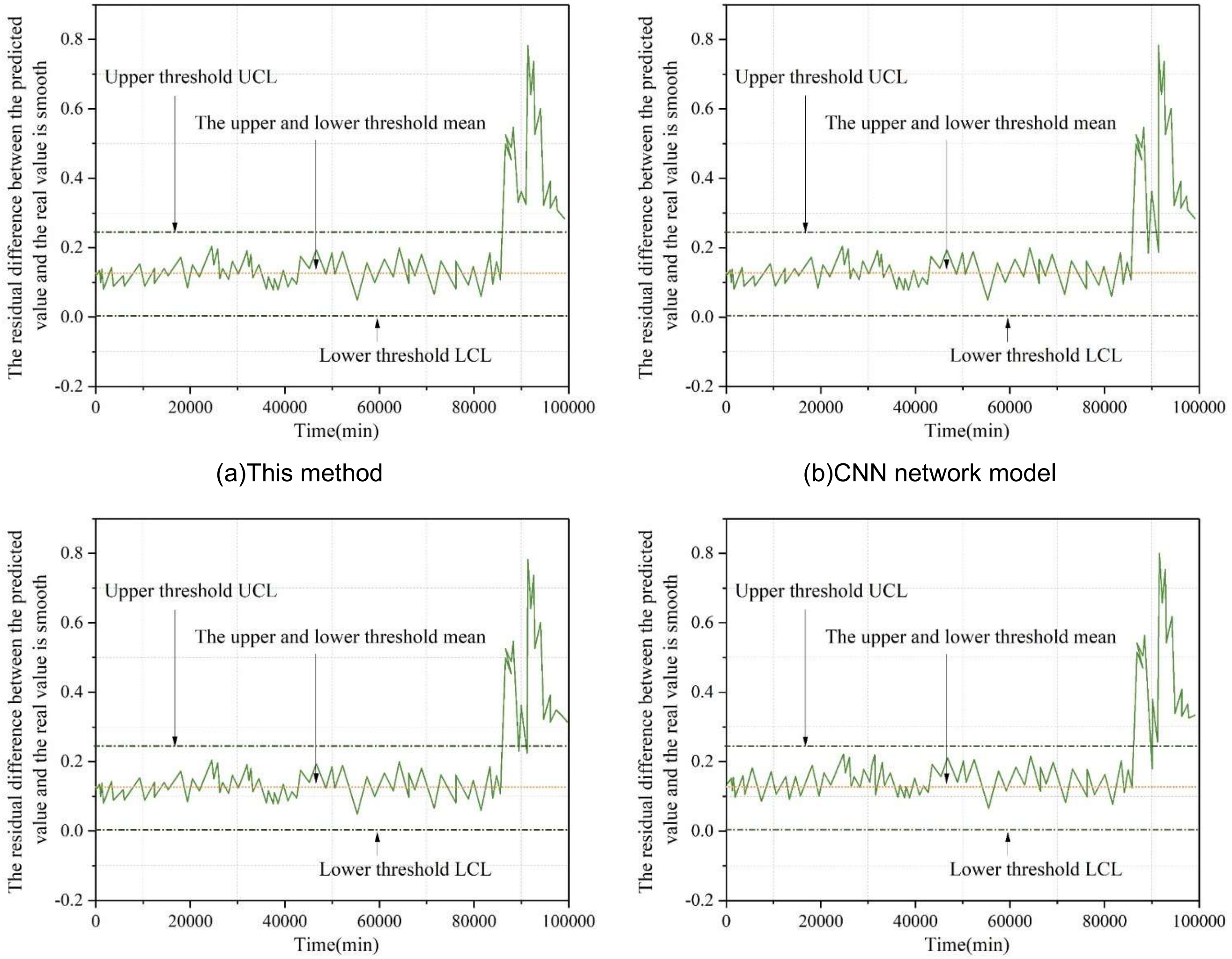
Table 1: Further selected fan gear box SCADA system monitoring variable

| Serial number | Variable name |
|---|---|
| 1 | Generator torque |
| 2 | Gear case cooling water temperature |
| 3 | Reactive power |
| 4 | Work ratio |
| 5 | Generator speed |
| 6 | No power |
| 7 | Dynamo power |
| 8 | Instantaneous wind speed |
| 9 | Generator winding maximum temperature |
| 10 | Engine direction |
| 11 | The speed axis of the gearbox is high |
| 12 | Instantaneous wind |
| 13 | The rear end temperature of the gearbox |
| 14 | 1 # blade Angle |
| 15 | Gear box oil pool temperature |
| 16 | Wind speed |
| 17 | Gear box inlet oil temperature |
| 18 | Ambient temperature |
| 19 | Gear box inlet pressure |
| 20 | Engine room temperature |
| 21 | Gear housing pump outlet pressure |
| 22 | Engine vibration effective value |

Data cleaning and restoration work was performed on the 22 parameter data screened. The data cleaning work cleans the missing, duplicated, and abnormal data, and the data repair work is to repair the cleaned data with single values. Then after that, the repaired data are divided into data samples. The self-attention-improved CEEMDAN-TCN constructed in this paper constructs the training samples by sliding-window sampling with 26-minute timeseries lengths and one-minute sliding steps for a training set with a total length of 60,000 minutes. In each constructed sample with a chronological length of 25 minutes, the 22 parameter data of the first 24 minutes are used as inputs, and the 22 parameter data of the last one minute are used as labels. The samples are divided for the test set in the same way as the training set. Finally, the training learning rate is set to 0.002, the number of training rounds is 110, and the training batch is 450, and the condition monitoring model is trained based on the divided training samples, and the model parameters are updated with the reverse gradient with the model prediction and the labeled mean squared construction loss.

In order to verify the superiority of the wind turbine gearbox condition detection method proposed in this chapter, this paper chooses “convolutional neural network CNN, long and short-term memory neural network LSTM and gated neural network GRU” as the control group, and compares the monitoring results with the method in this paper.

The monitoring results of the four methods are shown in Fig. 1, where (a) to (d) represent the method of this paper, the CNN network model, the LSTM network model and the GRU network model, respectively. From the figure, it can be seen that the monitoring results of the proposed method in this paper are bounded by a certain time point, before which the value is always between the upper threshold UCL and the lower threshold LCL, which means that the wind turbine gearbox is in good operating condition during this period and will not trigger an alarm; after this time point, the value rises rapidly and crosses the upper threshold UCL in a short time, after which the value is always greater than the upper threshold UCL and will not fall back to the lower threshold UCL. UCL and will not fall back below the lower threshold UCL, and the alarm is continuously triggered. However, the CNN and LSTM and GRU exceed the upper threshold UCL for the first time at a certain point in time, triggering an alarm, and after a period of time the alarm continues to fall back to below the upper threshold UCL and the alarm is canceled; after that, the value will exceed the upper threshold UCL for the second time at a certain point in time, triggering an alarm, which will continue to be triggered until the alarm time that comes with the SCADA system. The reason for the above phenomenon is that the model proposed in this paper fully exploits the spatio-temporal correlation information, and compared with the traditional CNN neural network and the time series models such as LSTM and GRU, this model has a better ability to discriminate between the abnormal and normal data, so there will not be any misjudgement and omission of judgment.



(a)This method

(b)CNN network model

(c)LSTM network model

(d)GUR network model

Figure 1: Monitoring results of four methods

The first time when different models exceeded the upper threshold and alarm occurred is shown in Table 2, the first time when the threshold was exceeded by the self-attention improvement CEEMDAN-TCN proposed in this paper is March 25, 2024 at 10:22 p.m. As seen in the table, the proposed method of this model is earlier than the GRU time-dependence mining model by 1 hour and 13 minutes, earlier than the LSTM time-dependence mining model by 55 minutes, and 2 hours and 17 minutes earlier than the CNN time-dependent mining model. This indicates that this model is more capable than the other three models in fitting the normal operation state and discovering abnormal data, which verifies the superiority of the method proposed in this paper. From the above comparison experiments, it can be seen that the SCADA monitoring model proposed in this chapter can successfully warn the abnormal faults of the wind turbine gearbox before the alarm of the maintenance system, and the alarm time of this model is earlier than that of the traditional deep learning model, which suggests that the method proposed in this paper can evaluate the health of the wind turbine gearbox more effectively by mining the spatial and temporal correlation information in the monitoring data of the wind turbine gearbox. The proposed method in this paper can more effectively assess the health of wind turbine gearboxes by mining the temporal and spatial correlation information in the monitoring data.

Table 2: For the first time, it is more than the upper threshold

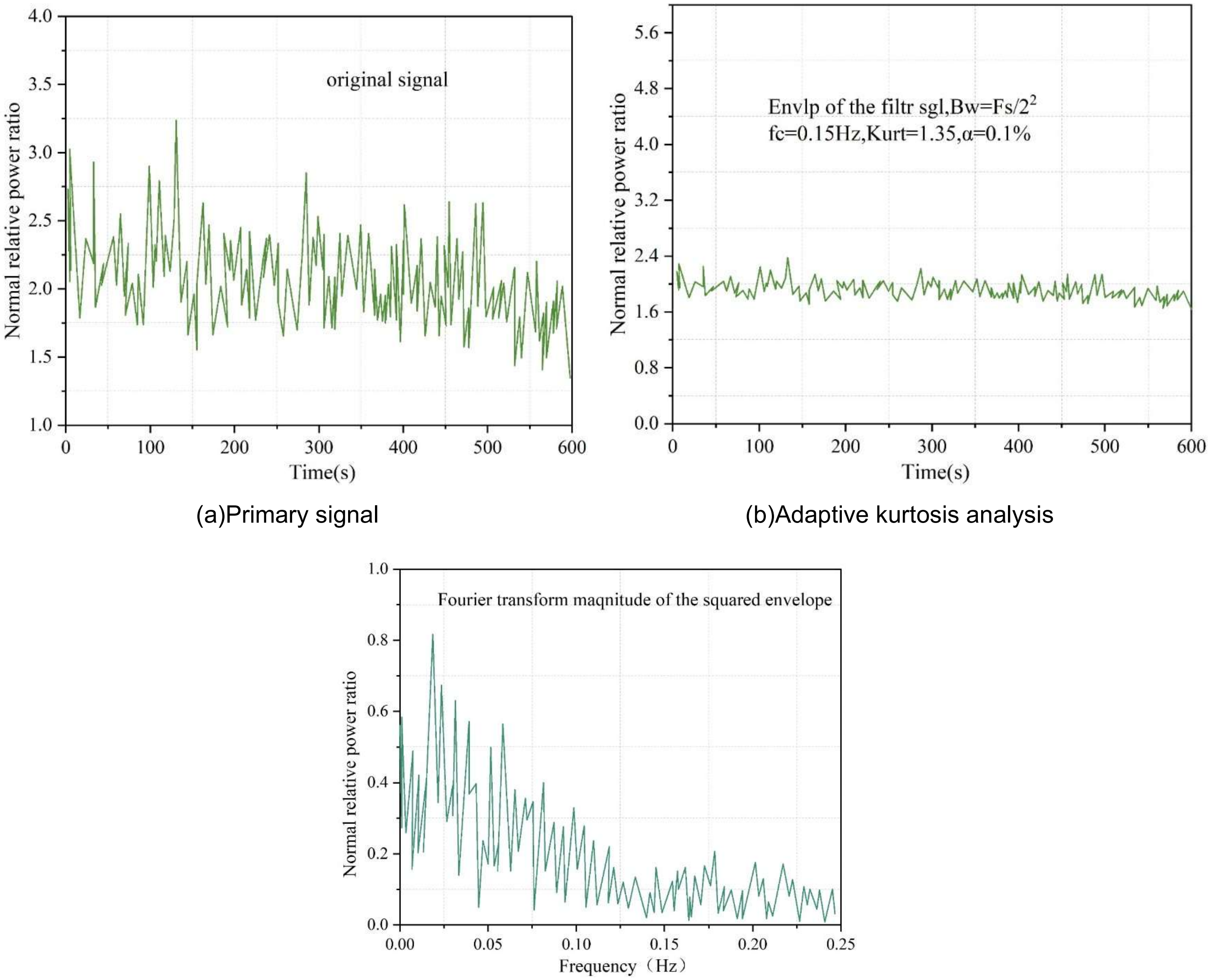
| Experimental method | Alarm time |
|---|---|
| CNN | On March 25, 2024, 12:39 |
| LSTM | On March 25, 2024, 11:17 |
| GRU | On March 25, 2024, 11:35 |
| This method | On March 25, 2024, 10:22 |

### *III. B. Adaptive Kurtosis Analysis*

#### III. B. 1) Pitch system failure analysis

(1) Graphical analysis

The abnormal and normal relative power ratios are discussed. The results of the analysis of the normal relative power ratio are shown in Fig. 2, where (a) to (c) represent the original signal, adaptive Kurtosis analysis and Fourier transform, respectively. The results show that although there is no obvious pattern in the numerical changes of the normal values, there are no cases that suddenly show great differences, and the trend of changes is more stable.



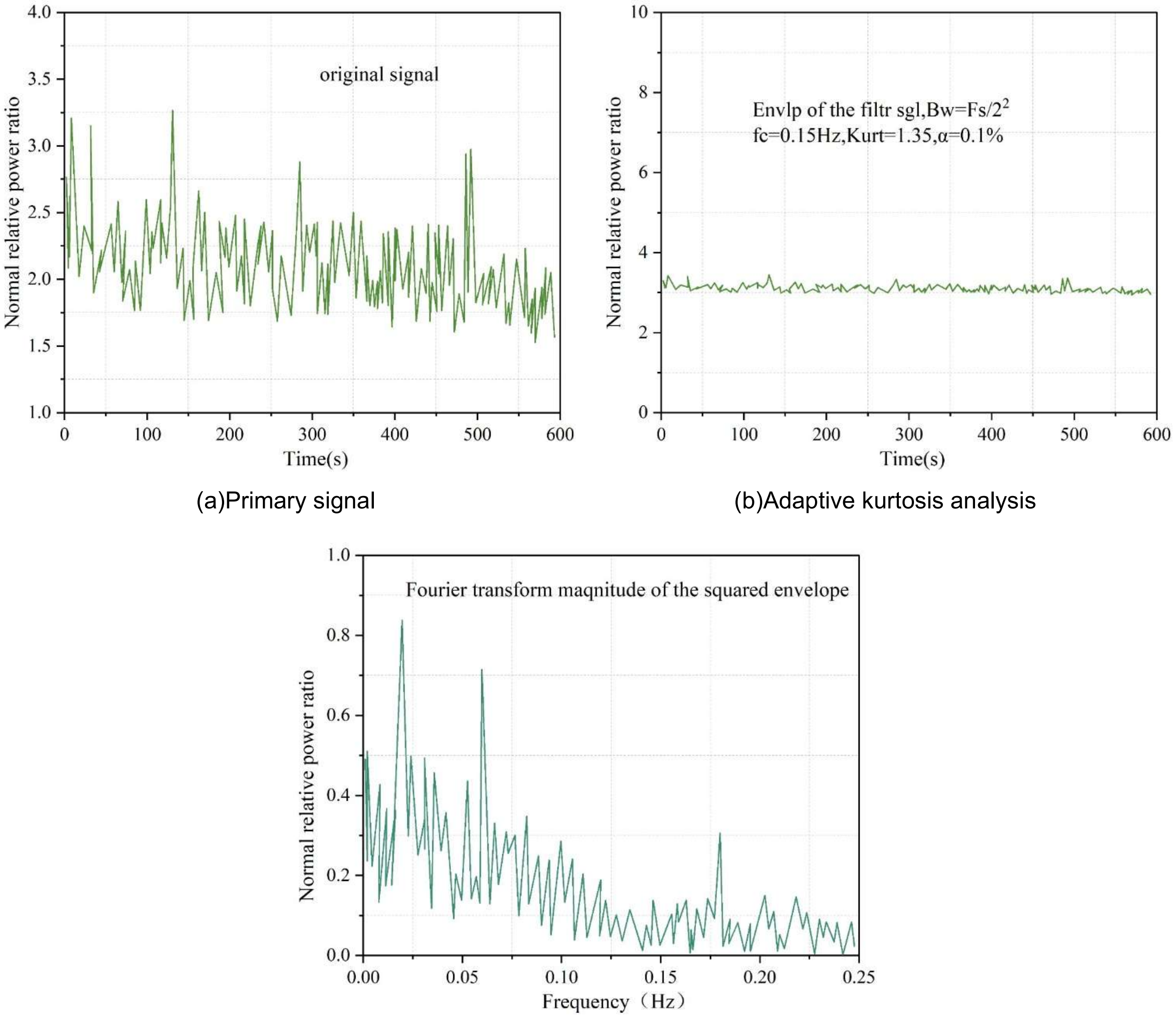
(a)Primary signal

(b)Adaptive kurtosis analysis

(c)Fourier transform

Figure 2: Normal relative power ratio analysis results

The anomalous relative power ratio analysis is shown in Fig. 3, where (a)~(c) represent the original signal, adaptive Kurtosis analysis and Fourier transform, respectively. As far as the relative power ratio of the anomalies is concerned, the original data obviously show a sharp increase in the four time periods, and the difference with the time periods before and after is large, which is obviously a deviation of its data from the normal data, so the data of the anomalous time periods can be analyzed as a basis for judging the data in comparison with the real values.



(a)Primary signal

(b)Adaptive kurtosis analysis

(c)Fourier transform

Figure 3: Abnormal relative power ratio analysis results

(2) Failure parameter analysis of electrical variable pitch system

The relative power ratio of the electrical pitch pitch control system is shown in Fig. 4. The results show that in the normal state, the pitch power ratio of the system is smaller than that in the fault state, and the majority of the pitch power ratio is mainly concentrated in the range of 1.0-2.0; whereas in the fault state, the majority of the pitch power ratio crosses a larger range, and it floats between 0.5-2.0. Thus, it can be seen that the pitch power ratio in the faulty state has a large deviation from that in the normal state.

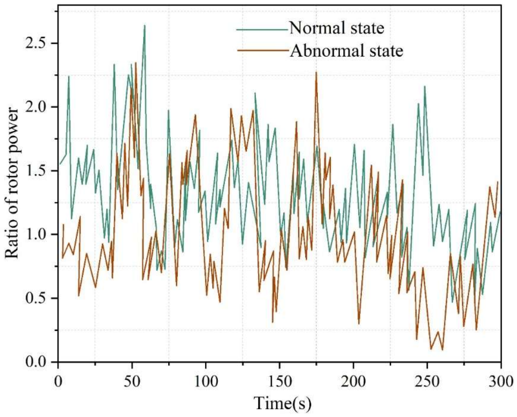


Figure 4: Relative power ratio of the control system

The parameter index of relative power ratio is shown in Table 3. It is not difficult to see that the fault data and normal data have significant differences in the mean, variance, crag, pulse and other indicators, according to which can be used as a basis for fault diagnosis.

Table 3: The parameters of the power ratio

| Index | Normal data | Failure data |
|---|---|---|
| Mean | 0.0486 | 1.9377 |
| Variance | 0.0011 | 0.0480 |
| Mean square amplitude | 0.0618 | 1.9511 |
| Peak | 0.2745 | 2.7235 |
| Mean amplitude | 0.0491 | 1.9402 |
| Mean square value | 0.0040 | 3.8103 |
| Sheer indicator | 1.8986 | -1.9463 |
| Peak index 4 | 4.4200 | 1.3951 |
| Waveform index | 1.2751 | 1.0082 |
| Pulse index | 5.6355 | 1.4038 |

### III. B. 2) Failure analysis of main control PLC system

(1) Graphical analysis

The abnormal and normal relative power ratio is discussed, and the relative power ratio under normal and fault condition is shown in Figure 5. Through the graphic obvious comparison, it can be clearly seen that under normal operation, the ratio of relative power are able to be in the presentation, while the fault data has exceeded the normal data change interval, can not be shown, and the relative power under the fault data is generally lower than the normal data.

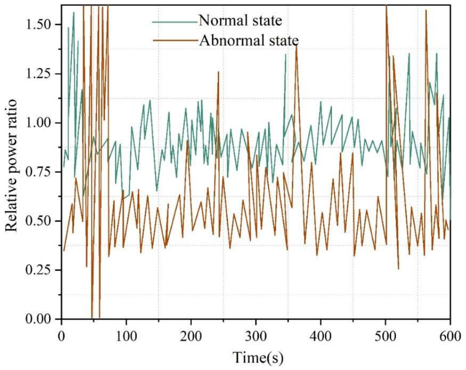


Figure 5: Relative power ratio analysis results

(2) Electrical main control PLC fault parameter analysis

Relative power ratio parameter indicators are shown in Table 4. It can be seen that the fault data and normal data in the average value, crag, peak and pulse and other indicators have significant differences, according to which can be used as a basis for fault diagnosis.

Table 4: Relative power ratio parameters

| Index | Normal data | Failure data |
|---|---|---|
| Mean | 0.0843 | 1.7088 |
| Variance | 0.0251 | 0.0731 |
| Mean square amplitude | 0.1872 | 1.7000 |
| Peak | 1.3401 | 3.0909 |
| Mean amplitude | 0.0985 | 1.7838 |
| Mean square value | 0.0371 | 3.2723 |
| Sheer indicator | 15.0034 | -1.8263 |
| Peak index 4 | 7.3356 | 1.1350 |
| Waveform index | 1.8940 | 0.7828 |
| Pulse index | 12.2402 | 1.2203 |

## IV. Conclusion

In this study, the multi-dimensional anthropomorphic condition monitoring method based on CEEMDAN-TCN shows excellent performance in wind turbine fault identification by analyzing and verifying the actual operation data from February to April 2024 of a wind farm in Sichuan. The method successfully realizes the early warning of the gearbox temperature control valve damage fault, with the alarm time of 10:22 on March 25, 2024, which has obvious time advantage over the traditional deep learning model. In the fault diagnosis of pitch system, the pitch power ratio in normal state is mainly concentrated in the interval of 1.0-2.0 with small change amplitude, while the ratio fluctuates greatly in the range of 0.5-2.0 in the fault state, and there are significant differences between the two states in the statistical indicators such as the mean value, variance, and craginess, which provide a reliable basis for the fault diagnosis. The proposed signal decomposition and deep learning fusion strategy effectively solves the problem of misjudgment and omission of traditional methods under complex working conditions, and improves the accuracy and robustness of abnormality detection by fully exploiting spatial and temporal correlation information. This technical solution provides new ideas for intelligent operation and maintenance of wind turbines, and has important engineering application value in reducing maintenance costs and improving power generation efficiency.

## References

[1] Vargas, S. A., Esteves, G. R. T., Maçaira, P. M., Bastos, B. Q., Oliveira, F. L. C., & Souza, R. C. (2019). Wind power generation: A review and a research agenda. Journal of Cleaner Production, 218, 850-870.

[2] Pryor, S. C., Barthelmie, R. J., Bukovsky, M. S., Leung, L. R., & Sakaguchi, K. (2020). Climate change impacts on wind power generation. Nature Reviews Earth & Environment, 1(12), 627-643.

[3] Zhang, Y., Wang, J., & Wang, X. (2014). Review on probabilistic forecasting of wind power generation. Renewable and Sustainable Energy Reviews, 32, 255-270.

[4] Hossain, M. M., & Ali, M. H. (2015). Future research directions for the wind turbine generator system. Renewable and Sustainable energy reviews, 49, 481-489.

[5] Allaei, D., Tarnowski, D., & Andreopoulos, Y. (2015). INVELOX with multiple wind turbine generator systems. Energy, 93, 1030-1040.

[6] Bensalah, A., Barakat, G., & Amara, Y. (2022). Electrical generators for large wind turbine: Trends and challenges. Energies, 15(18), 6700.

[7] Reder, M. D., Gonzalez, E., & Melero, J. J. (2016, September). Wind turbine failures-tackling current problems in failure data analysis. In Journal of Physics: Conference Series (Vol. 753, No. 7, p. 072027). IOP Publishing.

[8] Olabi, A. G., Wilberforce, T., Elsaid, K., Sayed, E. T., Salameh, T., Abdelkareem, M. A., & Baroutaji, A. (2021). A review on failure modes of wind turbine components. Energies, 14(17), 5241.

[9] Santelo, T. N., de Oliveira, C. M. R., Maciel, C. D., & de A. Monteiro, J. R. B. (2022). Wind turbine failures review and trends. Journal of Control, Automation and Electrical Systems, 1-17.

[10] Chen, J., Pan, J., Li, Z., Zi, Y., & Chen, X. (2016). Generator bearing fault diagnosis for wind turbine via empirical wavelet transform using measured vibration signals. Renewable Energy, 89, 80-92.

[11] Zhao, Y., Li, D., Dong, A., Kang, D., Lv, Q., & Shang, L. (2017). Fault prediction and diagnosis of wind turbine generators using SCADA data. Energies, 10(8), 1210.

[12] Liu, W. Y., Tang, B. P., Han, J. G., Lu, X. N., Hu, N. N., & He, Z. Z. (2015). The structure healthy condition monitoring and fault diagnosis methods in wind turbines: A review. Renewable and Sustainable Energy Reviews, 44, 466-472.

[13] Dao, P. B. (2022). Condition monitoring and fault diagnosis of wind turbines based on structural break detection in SCADA data. Renewable Energy, 185, 641-654.

[14] Dey, S., Pisu, P., & Ayalew, B. (2015). A comparative study of three fault diagnosis schemes for wind turbines. IEEE Transactions on Control Systems Technology, 23(5), 1853-1868.

[15] Attallah, O., Ibrahim, R. A., & Zakzouk, N. E. (2022). Fault diagnosis for induction generator-based wind turbine using ensemble deep learning techniques. Energy Reports, 8, 12787-12798.

[16] Liu, Y., Wu, Z., & Wang, X. (2020). Research on fault diagnosis of wind turbine based on SCADA data. IEEE Access, 8, 185557-185569.

[17] Merizalde, Y., Hernández-Callejo, L., Duque-Perez, O., & López-Meraz, R. A. (2020). Fault detection of wind turbine induction generators through current signals and various signal processing techniques. Applied Sciences, 10(21), 7389.

[18] Liulin Yang,Zhenning Huang,Xiujin Mo & Tianlu Luo. (2025). Enhanced GAIN-Based Missing Data Imputation for a Wind Energy Farm SCADA System. Electronics,14(8),1590-1590.

[19] JiaJing Gao,HongMei Xing,YongSheng Wang,GuangChen Liu,Bo Cheng & DeLong Zhang. (2025). Ultra-short-term wind power prediction based on hybrid denoising with improved CEEMD decomposition. Renewable Energy,251,123352-123352.

[20] Guoyuan Qin,Xiaosheng Peng & Zimin Yang. (2025). Regional Short-Term Wind Power Prediction Based on CEEMDAN-FTC Feature Mapping and EC-TCN-BiLSTM Deep Learning. Wind Energy,28(6),e70025-e70025.