Condition monitoring of wind turbines based on spatial-temporal feature aggregation networks

Jun Zhan*
College of Computer Science, National University of Defense Technology, Changsha 410073, China
E-mail: zhanjun20@nudt.edu.cn

Chengkun Wu
State Key Laboratory of High Performance Computing, College of Computer Science, National University of Defense Technology, Changsha 410073, China
E-mail: chengkun_wu@nudt.edu.cn

Canqun Yang
National SuperComputing Center in Tianjin, China
E-mail: canqun@nudt.edu.cn

Qiucheng Miao
College of Computer Science, National University of Defense Technology, Changsha 410073, China
E-mail: miaoqiucheng19@nudt.edu.cn

Shilin Wang
Beijing Goldwind Smart Energy Technology CO., Ltd, Beijing, China
E-mail: wangshilin@goldwind.com.cn

Xiandong Ma*
Engineering Department, Lancaster University, Lancaster, LA1 4YW, UK
E-mail: xiandong.ma@lancaster.ac.uk

Abstract: The existing supervisory control and data acquisition (SCADA) system continuously collects data from wind turbines (WTs), which provides a basis for condition monitoring (CM) of WTs. However, due to the complexity and high dimension and nonlinearity of data, effective modeling of spatial-temporal correlations among the data still becomes a primary challenge. In this paper, we propose a novel CM approach based on the **m**ultidirectional **s**patial-**t**emporal **f**eature **a**ggregation **n**etworks (MSTFAN) to accurately evaluate the performance and hence diagnose the faults of the turbines. Firstly, the data collected from various sensors are formulated into graph-structured data at each timestamp. Spatial-temporal network constructed by combing a graph attention network (GAT) and a temporal convolutional network (TCN) is used to extract spatial-

* Corresponding author
Email address: *zhanjun20@nudt.edu.cn* (Jun Zhan), *xiandong.ma@lancaster.ac.uk* (Xiandong Ma)

temporal features of the data. Then, a bi-directional long short-term memory (BiLSTM) neural network is adopted to further study long-term spatial-temporal dependency of the extracted features. Finally, the condition score is obtained and the streaming peaks over threshold (SPOT) is applied to determine the abnormal threshold for early warning of the fault occurrence. Experiments on datasets from real-world wind farms demonstrate that the proposed approach can detect the early abnormal situation of the WTs, and outperform other established methods.

Keywords: Wind turbine, condition monitoring (CM), graph attention network (GAT), temporal convolutional network (TCN), spatial-temporal correlation, streaming peaks over threshold (SPOT)

1. Introduction

Wind energy is a clean source and has become renewable energy for the most promising commercial development [1]. According to the latest global wind power report from GWEC (Global Wind Energy Council), the global capacity of installed wind power has reached 837GW to date [2]. However, the rapid expansion of wind farms has been affected by operations and maintenance (O&M) issues and high O&M costs. The designed service-life of a WT is usually 20 years while the total O&M costs constitute up to 30% of the total income of the turbines over their operating lifetime [3]. In particular, on-site maintenance for offshore wind turbines becomes more expensive since it requires complicated offshore operations [4]. Hence, it is imperative to design an efficient CM approach based on artificial intelligence techniques to improve the O&M strategies from post maintenance and planned maintenance to condition-based maintenance and predictive maintenance， which will help to reduce O&M costs and ensure the long-term healthy and stable development of the wind power industry [5].

Currently, condition monitoring of wind turbines has been performed by the supervisory control and data acquisition (SCADA) system and the specifically designed condition monitoring system (CMS) [6-8]. CMS system adds high-precision sensors in the corresponding positions of key components of WTs to collect the parameters such as vibration [9] and temperature [10]. The collection frequency is generally higher than 50Hz [11]. Then the methods such as spectral analysis [12], envelope analysis [13] and machine learning [14] are applied to analyze the data to achieve the CM. This way can accurately identify the specific types, positions and damage degree of equipment faults. However, the cost of a CMS is relatively high, which can be more than 11,000 Euros per turbine [15]. The SCADA system has been developed for real-time monitoring and control of WT operations by collecting data from a large number of parameters covering all key components in WTs [8, 16]. Moreover, the sampling frequency is generally lower than 1Hz. The SCADA system can provide valuable online information with depth and breadth regarding the performance and operational history of the WTs. Therefore, SCADA data have been widely used for fault detection and CM purpose [17, 18]. However, the CM based on SCADA data has faced a number of challenges [19] as follows: 1) Poor data quality, the operation of WTs is affected by variable conditions, disturbances and manual debugging, resulting in the collected data being contaminated with a large amount of bad data that cannot accurately reflect real operation states. 2) Imbalanced data classification, WTs mostly operate in the normal condition, while abnormal data are usually scarce. Meanwhile, due to the high manual label cost, valuable labels are not yet added to the collected data. 3) Complex feature correlation, because of the inter-coupling among different components or subsystems of the WTs, SCADA data are naturally high-dimensional and have complex cross-correlation and self-correlation.

In order to tackle these challenges, a number of CM methodologies have been developed, which can be generally categorized as physical model-based and data-driven approaches. When a large amount of data cannot be obtained, it is a natural choice to use the physical model-based methods. The model-based methods simulate the dynamic process of the system by establishing an accurate physical model for the checked objects [20]. Then, the residual between estimated and actual values of the parameters can be calculated for CM [21]. For instance, Feng et al. [22] created a wind power transmission model for gearbox condition monitoring by considering the heat transfer mechanism of the gearbox lubrication system, thus providing a sound theoretical basis for CM of the WTs. Dong et al. [23] used the gaussian mixture model (GMM) to build a multi-regime model of selected parameters that are greatly affected by working conditions.

On the contrary, data-driven methods do not require excessive prior knowledge, thus making this approach advantageous when performing CM tasks with complex-coupling effects and highly nonlinear dynamic performances [24]. In recent years, data-driven methods have been gained more attention, including shallow learning approach and deep learning approach [25]. Shallow learning algorithms usually construct a data-based probability-statistical model to achieve forecasting and analysis of data, mainly including regression [26], clustering [27], classification [28] and boosting algorithm [29]. For instance, Meik et al. [30] utilized a linear regression model to solve the correlation among the variables of WTs. Tang et al. [31] and Liu et al. [32] proposed fault diagnosis approaches based on Shannon wavelet support vector machine and clustering binary tree support vector machine, respectively. Furthermore, the clustering algorithm has been extensively applied especially for dealing with the common abnormal alignments [33]. Kouadri et al. [34] proposed the hidden Markov models (HMM) by incorporating machine learning-based HMM and principle component analysis to improve the availability and reliability of the fault diagnosis model under different operating conditions. Trizoglo et al. [29] developed an ensemble model of the extreme gradient boosting (XGBoost) framework to achieve a higher accurate detection at low computational costs. Tang et al. [35] developed an improved LightGBM fault diagnosis method for WT gearboxes by embedding the confusion matrix as a performance indicator. However, when dealing with a large amount of heterogeneous data, most shallow learning algorithms have some shortcomings. For instance, logistic regression is easy to underfit and decision trees are prone to overfitting [35]. Moreover, these methods usually construct features by combining expert knowledge and may exist the problems such as slow convergence speed and low prediction accuracy when a large amount of data is processed [36].

Compared with the shallow learning algorithms, the deep learning method has a better adaptability and mapping capability [37], which has shown evident advantages when dealing with highly nonlinear SCADA data [38, 39]. Applying a deep belief network to the abnormal detection of vibration signals in WTs can learn more extensive feature representations and improve recognition accuracy [40]. SCADA data, being regarded as time series data and capturing the long-term dependency relationship among features, would be vital for fault classification. Long short-term memory (LSTM) can utilize its specific gates mechanism to satisfy this requirement [29, 41]. These methods are implemented mostly based on auto-encoder (AE) structure, and discriminate anomalies through reconstruction errors. For instance, Chen et al. developed the AE-LSTM to assess sequential CM data. Wu et al. [42] combined the LSTM with statistical Kullback-Leibler divergence (KLD), where the LSTM network was used to capture long-term dependency among the monitoring data while the KLD value was applied for making decisions. Compared with LSTM, convolutional neural network (CNN) has a strong learning ability for the spatial features of data. In order to deal with the multiscale characteristics inherent in the vibration signals of a gearbox, Jiang et al. [43] proposed a multiscale CNN architecture for simultaneous multiscale feature extraction and classification. In addition, the CNN-LSTM or CNN-GRU models have also achieved good results in fault diagnosis because of their abilities for spatial-temporal information extraction [44-46]. However, CNN performs remarkably in the field of image processing and is based on the assumption that the data exist in Euclidean space, which implies that the correlations among data can be measured by the Euclidean distance. This is clearly not enough for multivariable SCADA data because there are different dimensions and physical significances among various parameters, such as power and temperature.

How to better identify the complex relations among different variables of SCADA data thus becomes a problem worthy of thinking. Recently, graph neural network (GNN) [47] has been proved to possess a stronger ability for dealing with relationship dependence, including graph convolutional network (GCN) [48], graph attention network (GAT) [49] and graph spatial-temporal networks [50]. These methods have received extensive applications in the fields of traffic forecasting and molecular property forecasting. For instance, Diao et al. [51] proposed a dynamic spatial-temporal GCN for accurate traffic forecasting, which tracks the spatial dependencies among traffic data. Achievements have also been made in processing time series data. To solve the anomaly detection problem of large IT systems, Scheinert et al. [52] proposed a network with the GCN architecture to extract spatial and temporal features. Deng et al. [53] proposed a GCN-based multivariate time-series anomaly detection method, which combines the relationships among sensor variables and sensor embeddings. Besides, in the latest research, Su et al. [54] proposed a method to extract the spatial features by using an attention module instead of CNN or GCN and showed promising results for the gearbox operating status detection of offshore wind turbines.

In this paper, we propose a novel spatial-temporal aggregation network for condition monitoring of WTs, which makes full use of multiple monitoring variables related to WTs specific faults by allowing information to propagate through directed graphs and temporal subsequences. Specifically, firstly the multiple monitoring variables are preprocessed from the feature and temporal dimensions, respectively. Then a flexible multidirectional spatial-temporal feature aggregation network (MSTFAN) is constructed to capture the inherent relations among them. From the feature extraction perspective, the complex cross-correlation among variables is learned through the edges of the graph nodes so that different attribute sensor data can be distinguished. The models based on graph attention network can allow the correlations among sensors to be represented in a non-Euclidean space, which is more suitable for the actual complex data structures. From the temporal perspective, the correlation among variables at different timestamps is extracted by dilated casual convolution to obtain larger receptive fields while keeping the network stability. BiLSTM is further used to reconstruct spatiotemporal information in order to capture long-term temporal dependencies. For fault detection, we introduce a SPOT-based approach to identify the condition states, which makes no assumptions about the data distribution and has stronger adaptability to real failures. The main contributions of this paper are summarized as follows:

(1) *A new CM framework for WTs.* Specifically, this approach is proposed to automatically learn complex spatial-temporal features of SCADA data for constructing the normal behavior model of operating WTs, by which the abnormal behaviors of WTs deviating from this normal model are recognized and diagnosed. For the first time, spatial-temporal features of multivariate

SCADA data are investigated for CM of the WTs.

(2) *A novel spatial-temporal network.* A flexible MSTFAN is designed for improving the performance of signal reconstruction by modeling and capturing both short- and long-term spatial and temporal correlations. To the best of our knowledge, this graph neural network is the first time applied to CM of the WTs.

(3) *A new abnormality warning strategy.* An extreme value theory-based SPOT approach is proposed to calculate the threshold to distinguish the normal and abnormal behaviors, by which the abnormality detectability is improved. Furthermore, we propose a novel "**d**elay **p**erception" (DPs) pre-warning strategy, which can reduce false warnings.

The rest of this paper is structured as follows. Section 2 describes the proposed WT fault detection framework. Section 3 presents the structure and working principle of MSTFAN in detail. Two case studies, namely, the drivetrain bearing fault of doubly-fed induction (DFIG)-based WTs and the pitch system fault of direct-driven WTs, are presented in Section 4, which are used to validate the effectiveness of the proposed method. In Section 5, the performance of the proposed CM method is assessed and compared with other existing mainstream methods. Finally, the conclusions and future improvements are given in Section 6.

2. System framework overview

2.1. CM flowchart

An unsupervised anomaly detection often refers to the task of identifying data patterns from a test dataset that appears the most divergent from the prevalent patterns of previously observed data [55]. Therefore, for the abnormal detection of WTs, we need to first create the normal behavior model and calculate the condition index of the WT. When the index exceeds a certain threshold during the diagnosis process with the test dataset, the WT will be regarded as abnormal. Generally, the newly-operated or major-repaired WTs after operating stably for a period of time are considered to be normal [56], and data from these periods are used to train the normal behavior models and calculate the alarm thresholds.

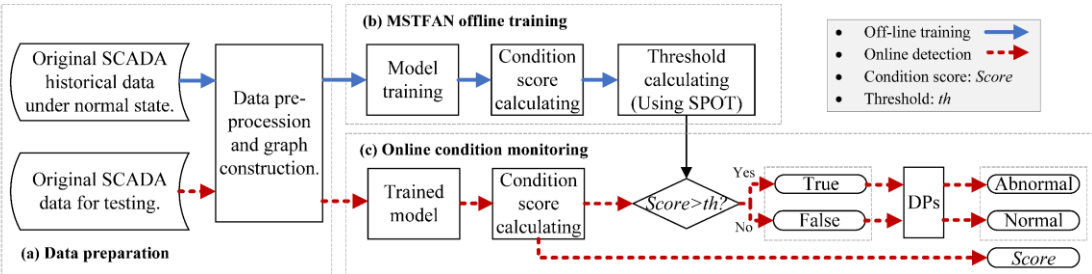


Figure 1 CM flowchart of WTs

The designed CM flowchart of WTs is shown in **Error! Reference source not found.**, where the process can be classified into three stages: data preparation and graph construction, MSTFAN offline training and online condition monitoring. In the stage of data preparation, the original SCADA data are cleaned first. This is because the SCADA system may cause data missing or mutation when communication failures or maintenance activities happen which do not indicate the actual abnormal state of the wind turbine. We adopt a quarterback method for mutation detection and revise the outlier and missing data as the average of the fore and aft values. In the stage of MSTFAN offline training, the normal operation pattern of WT is learned from a large number of historical SCADA data and the monitoring parameters available for the targeted WT under normal operating conditions are thus selected in this stage. The condition scores of normal data are then calculated, by which the threshold th of the scores is obtained by SPOT to quantify the abnormal level of the testing data. The model training and testing will be presented in detail in Section 3. In the stage of online condition monitoring, the testing data are inputted into the trained MSTFAN model to obtain their reconstructed values. The deviation between the actual and the reconstructed values is then used to calculate the condition score. The data with a score larger than the threshold th are discriminated as abnormal. It is worth noting that the length of the input sequence and the number of layers of the network could affect the detection results, which will be discussed in detail in Section 4.3. Besides, when the WT assembly operates gradually from normal to abnormal conditions, there appear inevitably fluctuations in the prediction results of the model, resulting in difficult decision-making for proper CM. Hence, after the detection results are initially obtained, a DPs is further designed to adjust the results for the second time in an attempt to eliminate the frequent alarms or false alarms. Finally, the abnormal condition is represented by binary (true or false), and scores are outputted simultaneously.

2.2. Graph construction

The MSTFAN-based CM method for WT aims to learn complex spatial-temporal correlations of SCADA data. Therefore, we creatively create a graph using the data from each window after processing the multivariate time series data with a sliding window, which are described in detail as follows:

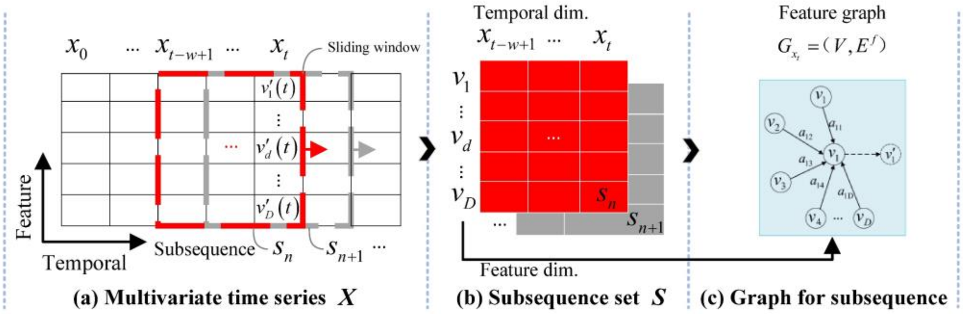


Figure 2 The construction process of the subsequences and the associated graph

Generally, for a given multivariate time series $X = \{x_1, x_2, \cdots, x_t, \cdots, x_T\}$, where T and $x_t \in R^D$ represent the length of data samples and data collected from D sensors at timestamp t, respectively. The subsequences are required to be processed with a uniform length. The construction process of the subsequences is shown in **Figure 2**. We define the feature vector at timestamp t as $x_t = \{v'_d(t) | d \in [1, D]\}$, and a sliding window containing certain features as $v_d = \{x_t | t \in [t - w + 1, t]\}$. As shown in **Figure 2**(a), a sliding window with length w is assigned to partition a long sequence X along the temporal dimension into N subsequences $s_1, s_2, \cdots, s_n, \cdots s_N$, which are collected as a subsequence set S. Then we build a feature graph for each subsequence s_n $(n = 1, \cdots, N)$, as can be seen in **Figure 2**(b) and (c). Each feature v_d is viewed as a node in the feature graph $\boldsymbol{G}_{x_t} = (\boldsymbol{V}, \boldsymbol{E}^f)$, where $V = \{\vec{v}_1, \vec{v}_2, \cdots, \vec{v}_D\}$ represents the node set, $\vec{v}_i \in \mathbb{R}^\omega$ represents the feature vector of each node, and E^f represents edge set (each edge denotes the connection between two corresponding features, and the mutual contribution between nodes i and j is expressed as a_{ij}) among feature nodes.

3. MSTFAN-based CM

The proposed MSTFAN-based CM framework is shown in **Figure 3**. The key motivation of MSTFAN is that using GAT and TCN simultaneously can capture implicit normal conditions from both feature and temporal dimensions of SCADA data. The framework consists of four parts, namely data representation, feature extraction, data reconstruction, and anomaly discrimination. We establish the MSTFAN to explore spatial dependence and temporal dependence among different sensors in the SCADA system. For each timestamp, spatial features and temporal features are aggregated together by splicing and then inputted into the reconstruction network. For the reconstruction network, a BiLSTM is adopted to capture long-term temporal correlations from spatial-temporal aggregation vectors, and then the output layer composed of a fully-connected network reconstructs the implicit vectors from BiLSTM to obtain the output vectors with the same length and dimension as the input subsequence. Here the reconstructed value and actual value are used to calculate the condition score, and the final abnormal condition is assessed based on the SPOT-based threshold.

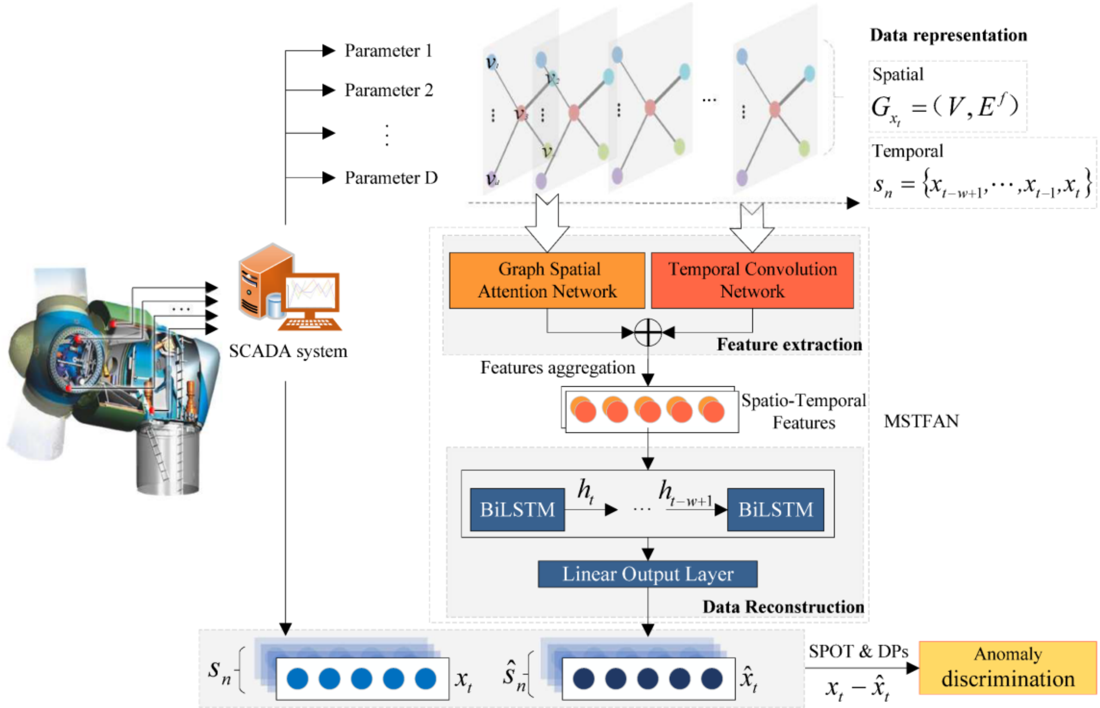


Figure 3 The proposed MSTFAN-based framework for WT CM

3.1. Spatial feature learning with GAT

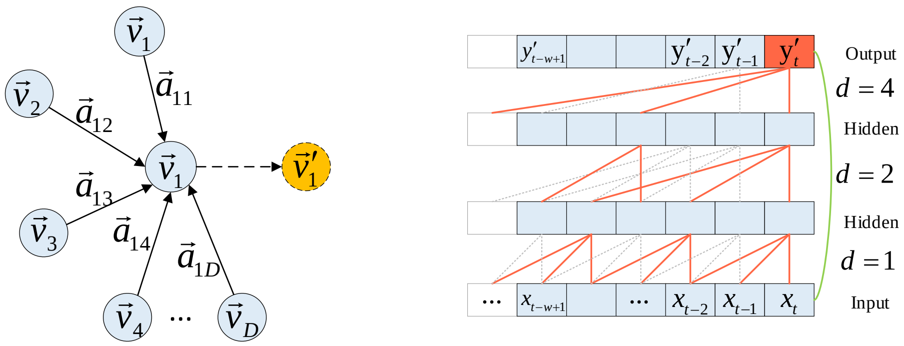
Graph attention network (GAT) was firstly proposed by Velikovi et al. [49], which utilizes attention mechanism to make a weighted sum of neighboring nodes, thus aggregating information and totally removing the constraints of the graph structure. To fully capture the spatial correlation among sensor parameters, we utilize the graph attention mechanism at each subsequence to process the signals. As shown in **Figure 3**, subsequence s_n is processed into the graph-structured data in the feature dimension and inputted into the GAT. The spatial correlation between the $i\text{-}th$ and the $j\text{-}th$ nodes can be represented by their attention coefficient being computed as:

$$\alpha_{ij} = \frac{\exp(\delta(\vec{a}^T[\mathbf{W}\vec{v}_i \,\|\, \mathbf{W}\vec{v}_j]))}{\sum_{k\in N_i}\exp(\delta(\vec{a}^T[\mathbf{W}\vec{v}_i \,\|\, \mathbf{W}\vec{v}_k]))} \quad (1)$$

where δ is an activation function, and generally uses LeakyReLU which has a relatively small positive gradient for negative inputs. $\vec{a} \in \mathbb{R}^w$ is a learnable weight vector, $\mathbf{W}$ is the shared weight vector and the operator $||$ represents the information concatenation of two nodes, and N_i denotes the number of adjacent nodes for the $i\text{-}th$ node. Finally, the output of each node can be obtained by aggregating its adjacent nodes, as shown in **Figure 4**(a). The implicit vector of i_{th} sensor node through one layer of GAT can be denoted as $\vec{v}_i'$. In order to avoid focusing too much on the position of the node itself, we adopt the multi-head attention mechanism [57], which is determined as follows.

$$\vec{v}_i' = \sigma\left(\frac{1}{H}\sum_{h=1}^{H}\sum_{j\in N_i} a_{ij}^h W^h \vec{v}_j\right) \quad (2)$$

where h is the head number, α_{ij}^h is the attention coefficient of the h-th head, H is the total number of the attention heads, and σ is the activate function. For the final result, an average value is adopted for the output vector of each attention head.



(a) GAT aggregation representation of different sensors through its adjacent nodes. a_{ij} denotes attention coefficient between sensor nodes.

(b) 4-layer TCN structure with residual connection, and d is dilation factor.

Figure 4 Implicit layer representation

3.2 Temporal feature learning with TCN

We adopt TCN to extract temporal features. The basic idea of TCN [58] is the combination of 1D-fully convolutional network (1D-FCN) and causal convolutions. Meanwhile, in order to obtain a larger receptive field and keep the network stability, TCN uses the extended causal convolution and residual modes to replace the general causal convolution network and the general convolutional layer, respectively, enabling the receptive field to expand exponentially. For 1-D input sequence $s_n = \{x_{t-w+1}, \cdots, x_{t-1}, x_t\}, x \in \mathbb{R}^D$, the convolution kernel is $f: \{0, \ldots, k-1\} \to \mathbb{R}$, and the convolution at timestamp t is defined as:

$$F(t) = (x *_d f)(t) = \sum_{i=0}^{k-1} f(i) \cdot x_{t-d\cdot i} \quad (3)$$

where d is dilation factor, k is the size of the convolution kernel, and $t - d \cdot i$ indicates the direction of the past. An example structure of 4-layer TCN network is shown in **Figure 4** (b).

After a series of convolution operations, the input sequence is mapped into the implicit vector y_t' containing temporal information:

$$y_t' = \mathcal{F}(x_t, \{W_t\}) + Conv_{1*1}(x_t) \quad (4)$$

where $\mathcal{F}$ represents the convolution operation module composed of a nonlinear causal expansion convolution, a nonlinear activation function ($ReLU$), a weight normalization and a dropout regularization. $Conv_{1*1}$ is used to adjust the dimension of input vector for realizing vector addition operation connected by residuals, and W_t is the learnable weight vector.

3.3 Feature aggregation and reconstruction

After obtaining the vectors of spatial and temporal features, we adopt a splicing approach to fuse them and send them to the BiLSTM-based reconstruction network. Different from the general LSTM network, BiLSTM deals with the data from two different directions, which can better capture bi-directional information dependence. **Figure 5** shows the basic structure of the signal reconstruction network.

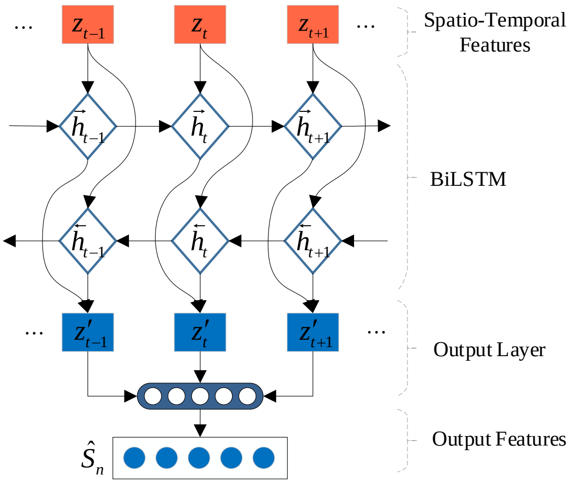


Figure 5 BiLSTM-based signal reconstruction network

In the figure, z_t is the spatial-temporal feature vector outputted from the MSTFAN at timestamp t. In the forward process, the implicit vector is updated by $\vec{h}_t$, while in the reverse process, the implicit vector is updated from the reverse direction and is denoted by $\overleftarrow{h}_t$. The combined implicit vector is represented as z'_t, and the relevant updating formulas are given as below:

$$\vec{h}_t = ReLU\left(W_{\vec{h}_t} z_t + U_{\vec{h}_t} \vec{h}_{t-1} + b_{\vec{h}_t}\right) \tag{5}$$

$$\overleftarrow{h}_t = ReLU\left(W_{\overleftarrow{h}_t} z_t + U_{\overleftarrow{h}_t} \overleftarrow{h}_{t+1} + b_{\overleftarrow{h}_t}\right) \tag{6}$$

$$\hat{s}_n = ReLU\left(W_{\vec{h}o} \vec{h}_t + W_{\overleftarrow{h}o} \overleftarrow{h}_t + b_o\right) \tag{7}$$

where $W_{\vec{h}_t}$ and $W_{\overleftarrow{h}_t}$ denote the learnable weight vectors from different directions for the spatial-temporal feature vector z_t, $U_{\vec{h}_t}$ and $U_{\overleftarrow{h}_t}$ denote the learnable weight vectors from different directions for hidden condition h_t, $W_{\vec{h}o}$ and $W_{\overleftarrow{h}o}$ denote learnable weight vectors from different directions for the output layer, $b_{\vec{h}_t}$, $b_{\overleftarrow{h}_t}$ and b_o denote the learnable bias, and $ReLU$ is the activation function.

3.4 Fault detection

To accurately reflect WT operation condition, we calculate the condition score at each timestamp. For the input subsequence s_n, the corresponding sequence $\hat{s}_n$ of the same size as s_n can be reconstructed, as described above. A residual signal is taken from the difference between the actual value and reconstructed value of all the subsequences for discriminating the data deviation and then calculating the condition score at each timestamp. To eliminate the effect of different variable dimensions, the scores are standardized [59], and calculated as follows:

$$score = \frac{1}{D}\left\| x_t - \hat{x}_t \right\|_2 \tag{8}$$

When the dataset from a WT under normal operation is applied to train the model to calculate the condition scores, the majority of scores are within a normal range. Hence, we introduce SPOT to define the threshold, since neither manually pre-set threshold nor distributional assumption is required in this method. We denote the obtained condition scores in the subsequence as $C = \{score_0, score_1, \cdots, score_t, \cdots score_T\}$, where T is the length of data and $score_t$ represents condition score at timestamp t. The SPOT is used to calculate the threshold th, ensuring the probability that $score_t > th$ is smaller than the given probability value q we have set, that is $P(score_t > th) < q$. In this paper, we set $q = 0.001$, an empirical value based on investigation into the nature of the datasets. With the testing dataset, when $score_t > th$, the data are regarded as abnormal, otherwise seen as normal. The predicted label describing the data condition can be defined as $Y' = \begin{cases} 1, (score \geq th) \\ 0, (score < th) \end{cases}$.

Then we use DPs for the second discrimination of Y' in order to obtain the final binary discrimination result of Y, thus improving the detection reliability. The detailed implementation process of DPs is shown in **Figure 6**. We observe the predicted results through two designated windows, namely, a voting area and an observation area, where the observation area contains several voting areas. During the CM process, the window representing the observation area continuously moves forward as time goes by. We denote the ratio of abnormal samples in each voting area as p ($i\text{-}th$ voting area is denoted as p_i). As shown in voting area ①, when p continuously increases to above 90%, the samples in this voting area are adjusted to be abnormal. On the contrary, as shown in voting area ②, when p continuously decreases to below 5%, the samples in this voting area are adjusted to be normal, while the area without continuous change of p in the observation area is not adjusted. This adjustment can avoid frequently repeated alarms in the voting area. In addition, after the WT operation returns to normal, false alarms can be avoided, thus improving the reliability of CM results for practical O&M arrangements.

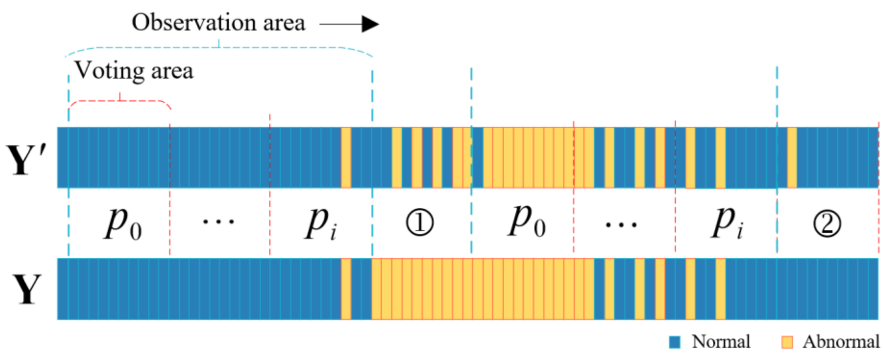


Figure 6 Use of delayed perception to reduce frequent alarms or false alarms

4. Experiment results and analysis

4.1 Datasets

Different types of WTs (direct-driven or doubly-fed) or WTs installed in different locations (mountainous or coastal) have great differences in data characteristics and high-risk components. For example, the hub of the direct-driven WT is directly connected to the rotor of the generator through rigid bearings; therefore, the hub speed is equal to the generator speed, which makes the variables of pitch system closely related to the variables of the generator. For DFIG, the hub is indirectly connected to the generator through the gearbox. Hence, the variables of the gearbox are closely related to the variables of the generator and it is the gearbox that is a high-risk component of failures in DFIGs.

Therefore, to evaluate the performance of the proposed method, the SCADA data from two different wind farms with representative types of faults, i.e., pitch system fault for direct-driven WTs and drivetrain bearing fault for DFIG WTs, are selected. These two wind farms are denoted as WF_1 and WF_2, respectively. The wind farm WF_1 is located in the southern coast of China, which consists of 25 WTs with nominal power 2MW and DFIG. Having checked the onsite O&M records, we select a WT with good operation condition and a WT with drivetrain bearing fault to build two different datasets represented by WF_1-WT_1 (without abnormality) and WF_1-WT_2 (with abnormality), respectively. The wind farm WF_2 is located in the south-central hilly area of China, which consists of 25 WTs with nominal power 2MW and direct-driven WT generators. We select a WT with good operation condition and a WT with pitch system fault to build two different datasets represented by WF_2-WT_1 (without abnormality) and WF_2-WT_2 (with abnormality), respectively. The datasets in detail are given in **Table 1**. More physical descriptions of these faults are described in subsequent case studies.

Table 1 Datasets in detail

Datasets	Samples	Anomaly ratio	Sampling intervals	Dimension (Monitoring parameters)	Location/Type	Description
WF_1-WT_1	52790	-	10 minutes	9	Coast/DFIG	Without abnormality
WF_1-WT_2	86829	7.3%	10 minutes	9	Coast/DFIG	Drivetrain bearing fault
WF_2-WT_1	915000	-	1 second	32	Hill/Direct-driven	Without abnormality
WF_2-WT_2	995800	7.5%	1 second	32	Hill/Direct-driven	Pitch system fault

4.2 Case studies

4.2.1 Case 1: main bearing fault

The main bearing is one of the most important components for DFIG-based WTs. Once the abnormality of bearing occurs, the operating safety of WT will be seriously threatened. As shown in **Figure 7**, the main bearing abnormalities mainly include bearing cage wear and deformation of the ball. The continuous rotation of the main bearing causes the variety of bearing temperatures. Under the normal condition, the temperatures at each location of bearing change with the rotating speed; however, their changes are still maintained within a certain range. However, the situation under abnormal operating conditions is different. Thus, as shown in **Table 2**, nine parameters were selected among hundreds of SCADA parameters [60, 61], primarily because the rotating hub and the rotor of the generator are connected through the gearbox, implying that the changes in the temperature of main bearings are closely correlated to the speeds, ambient temperature, and active powers. For instance, when the bearing cracks, the temperature difference between the front-end and back-end of the gearbox may change greatly.

Figure 7 Physical photos of the drivetrain bearing fault

Table 2 SCADA parameters of bearing fault diagnosis in WF_1

Location	**Parameters**	**Unit**
Weather station	Ambient temperature	℃
	Wind speed	m/s
Hub	Rotation speed	rpm
Generator	Active power at generator side	kW
	Active power at grid side	kW
Drivetrain	Bearing temperature 1 (middle in front-end)	℃
	Bearing temperature 2 (front-end)	℃
	Bearing temperature 3 (back-end)	℃
	Bearing temperature 4 (middle in back-end)	℃

Figure 8 and **Figure 9** present CM results using the datasets $WF_1 - WT_1$ and $WF_1 - WT_2$, respectively. **Figure 8(a)** and **(b)** show, under the normal conditions, that the actual value and reconstructed value almost coincide, and **Figure 8(c)** shows that the condition score maintains within the range of 0.01. This is because the model only uses the data under the normal conditions for training and learns the normal behavior of the WT, demonstrating the accuracy of the predicted model.

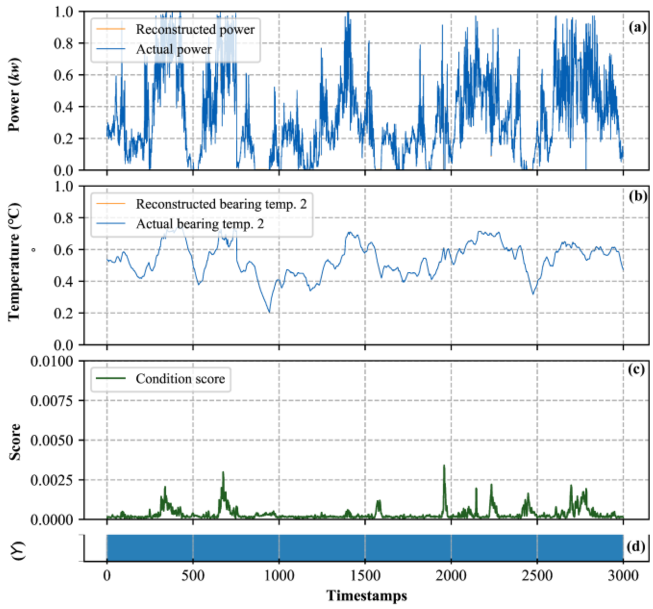


Figure 8 Abnormality detection results of the $WF_1 - WT_1$

According to the maintenance records during the routine inspection, the O&M personnel detected a grease abnormality of the main bearing and smelled the stink produced due to bearing deformation at the timestamp 1900. However, there was no fault alarm because the value of temperature did not exceed the alarm threshold set by the existing SCADA system. **Figure 9(a)** and **(b)** show the actual value and reconstructed value for the active power and main bearing temperature 2, respectively. The condition scores in **Figure 9**(c) show that the proposed algorithm is able to detect the abnormality at the timestamp 1500, at which the WT was still generating power, showing that the fault had not caused the WT to shut down. Then the condition score gradually increases until it fully deviates from the original normal condition. Based on the initial value of the SPOT threshold calculated using the dataset only containing health data, the threshold calculated using the testing dataset becomes higher up to 0.0116, enabling to differentiate the condition scores from the two datasets properly. It is noticed that in the early stage of fault occurrence, the incipient abnormality results in the condition scores with certain fluctuations. As shown in **Figure 9**(d), after the DPs adjustment, the condition scores are adjusted to be anomalies near the timestamp 1500, which is beneficial for the O&M personnel to be able to make decisions nearly 66 hours in advance about maintenance plans for WTs.

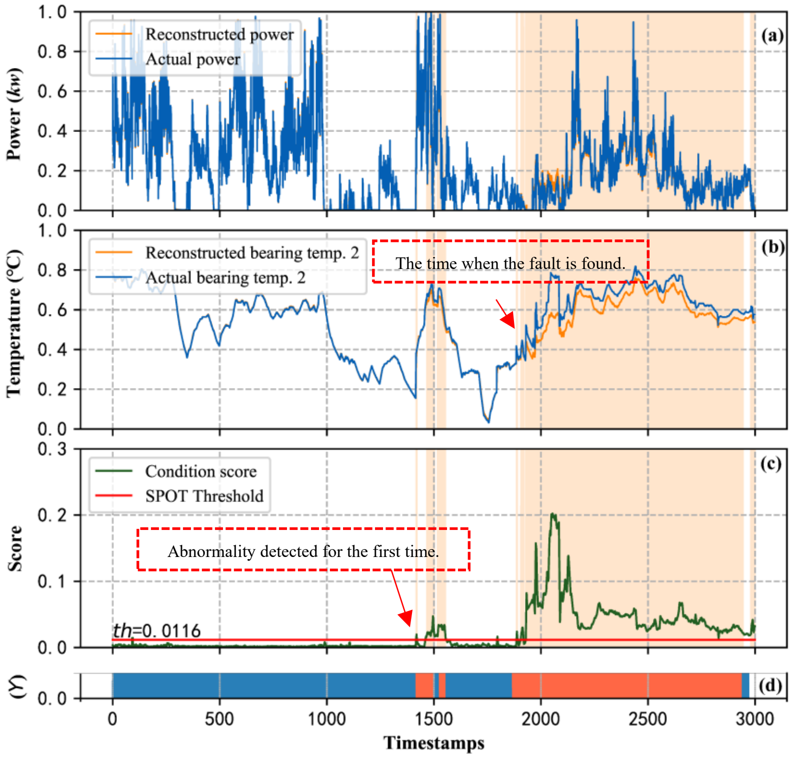
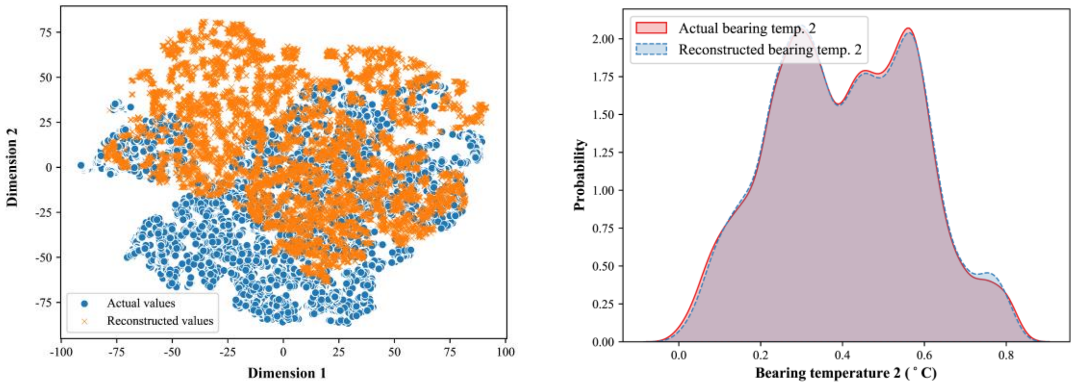


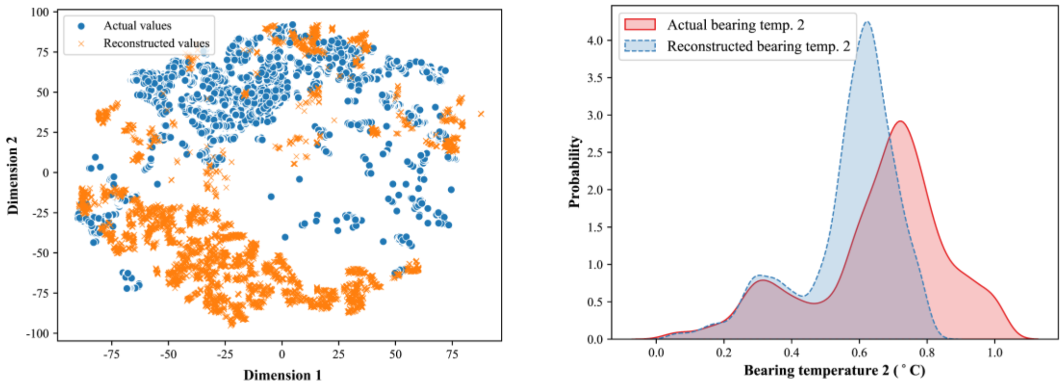
Figure 9 Fault detection results of the $WF_1 - WT_2$

Figure 10 and **Figure 11**, respectively, show the probability density distribution of actual value and reconstructed value of WF_1-WT_2 within the normal and abnormal periods, where sub-figure (a) shows the distribution of all parameters after dimension reduction by t-distributed stochastic neighbor embedding (t-SNE) algorithm and sub-figure (b) shows the probability density distribution of the main bearing temperature 2. The comparison of **Figure 10**(a) and **Figure 11** (a) clearly shows that our method fits the data well in normal conditions, while the distribution of the reconstructed values deviates from the original distribution in abnormal conditions. As shown in **Figure 10**(b), under normal conditions, the main bearing temperature 2 is distributed in the range of 0 to 0.84℃. However, within the abnormal conditions, as seen in **Figure 11**(b), there are significant differences in the temperature distributions, indicating the existence of abnormality. This is consistent with the main bearing cracking discovered during the maintenance activities.



(a) t-SNE distribution of all parameters **(b)** Probability density distribution of main bearing temperatures

Figure 10 The distribution of actual value and reconstructed value from $WF_1 - WT_2$ under normal conditions



(a) t-SNE distribution of all parameters (b) Probability density distribution of main bearing temperatures

Figure 11 The distribution of actual value and reconstructed value from $WF_1 - \mathrm{WT}_2$ under abnormal conditions

4.2.2 Case 2: pitch system fault

In this section, a CM case of direct-drive WT pitch system is presented. Because of the time-varying nature of winds, the direct-drive WT needs to adjust the blade angle through the pitch system to achieve a stable energy output. Once a pitch system fault occurs, as shown in **Figure 12**, the common abnormalities include pitch-bearing cracks and bolt fractures. As shown in **Table 3**, thirty-two parameters were selected among hundreds of SCADA parameters [62, 63], primarily because the bearing cracking may lead to an increase in running resistance, which results in the deviation of related pitch system parameters away from the original normal condition, such as the wide fluctuations of pitch-motor currents and changes in pitch-motor temperatures and temperatures of those components housed in the hub. At the same time, the failure of any blade bearing will cause damage to the WT due to imbalance. Therefore, those parameters related to blades, such as blade angles and nacelle vibrations are also important, along with the wind conditions and generator operation parameters.

Figure 12 Physical photos of the pitch system fault

Table 3 SCADA parameters of the pitch system fault in WF_2

Location	Parameters	Unit	Parameters	Unit
Weather station	Ambient temperature	℃	Wind direction	°
	Wind speed	m/s		
Hub	Rotation speed	rpm	Hub temperature	℃
	Pitch motor current 1	A	Angle of blade 1	°
	Pitch motor current 2	A	Angle of blade 2	°
	Pitch motor current 3	A	Angle of blade 3	°
	Pitch motor temperature of blade 1	℃	Pitch motor power 1	kW
	Pitch motor temperature of blade 2	℃	Pitch motor power 2	kW
	Pitch motor temperature of blade 3	℃	Pitch motor power 3	kW
	Battery temperature of blade 1	℃	Inverter temperature of blade 1	℃
	Battery temperature of blade 2	℃	Inverter temperature of blade 2	℃
	Battery temperature of blade 3	℃	Inverter temperature of blade 3	℃
Nacelle	Vibration x	m/s2	Hydraulic brake pressure	bar
	Vibration y	m/s2		
Generator	Active power at generator side	kW	Turbine state (e.g., startup, generation, and stop)	
	Active power at grid side	kW	Generator current	A
	Generator frequency	Hz	Generator torque	kNm

The CM results on the datasets $WF_2 - WT_1$ and $WF_2 - WT_2$ are respectively shown in **Figure 13** and **Figure 14**. It can be seen from **Figure 13** that under the normal condition, the actual value almost coincides with the reconstructed value, and the condition score maintains in the range of 0.04. Although there is a jump at the end of the condition score in **Figure 13**(c), it is still normal and within the allowable range. We can see clearly from **Figure 13**(a) that the actual power changes to zero at this time instant when the turbine shuts down, indicating that our method is still robust in dealing with such sudden changes in operations.

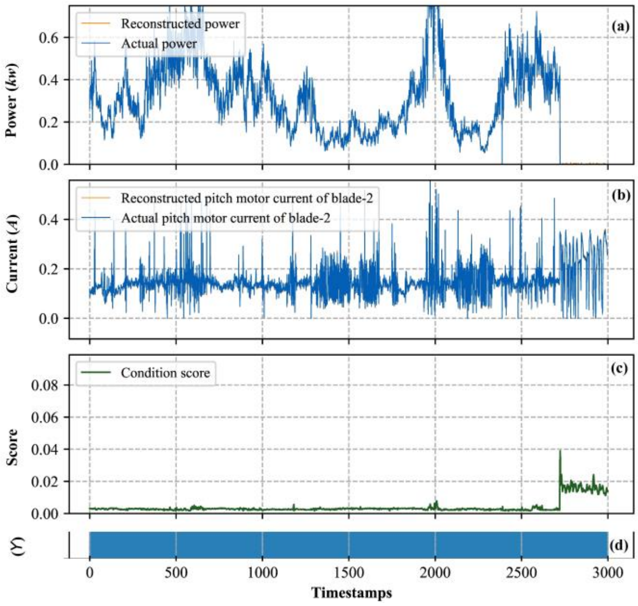


Figure 13 Abnormality detection result of the $WF_2 - WT_1$

According to the maintenance records during the routine inspection, the maintenance personnel detected a bolt fracture at the outer pitch bearing of blade-2 and a fracture at the location of the bearing ball-plugging hole at timestamp 46800. By analyzing the data, it is found that the pitch motor current of blade-2 is clearly higher than that of blade-1 and blade-3, and furthermore the motor current presents large fluctuations at high frequencies, as can be seen in **Figure 14**(b). However, the SCADA system does not detect the abnormality. Instead, it only gives an alarm. The condition score curve in **Figure 14**(c) shows that at timestamp 33800, the proposed algorithm has detected the abnormalities above the threshold 0.1337 for several times, indicating that the pitch system fault occurs at that moment. As shown in **Figure 14**(d), after the DPs adjustment, the proposed CM method identifies the fault occurrence 3.6 hours earlier. It is noteworthy that the threshold value of 0.1337 in this case is much higher than the threshold value of 0.0116 in case 1, indicating a decline in the model's capacity in terms of fitting the normal data. This is primarily because the two cases work at different sampling rates and with different number of parameters (9 for case 1 while 32 for case 2). The maximum scores under abnormal conditions are also significantly different with 0.2 and 0.8 for the two cases respectively. This shows that although the detection difficulty increases, the proposed model still demonstrates good CM ability in dealing with the high-dimensional and complex parameters.

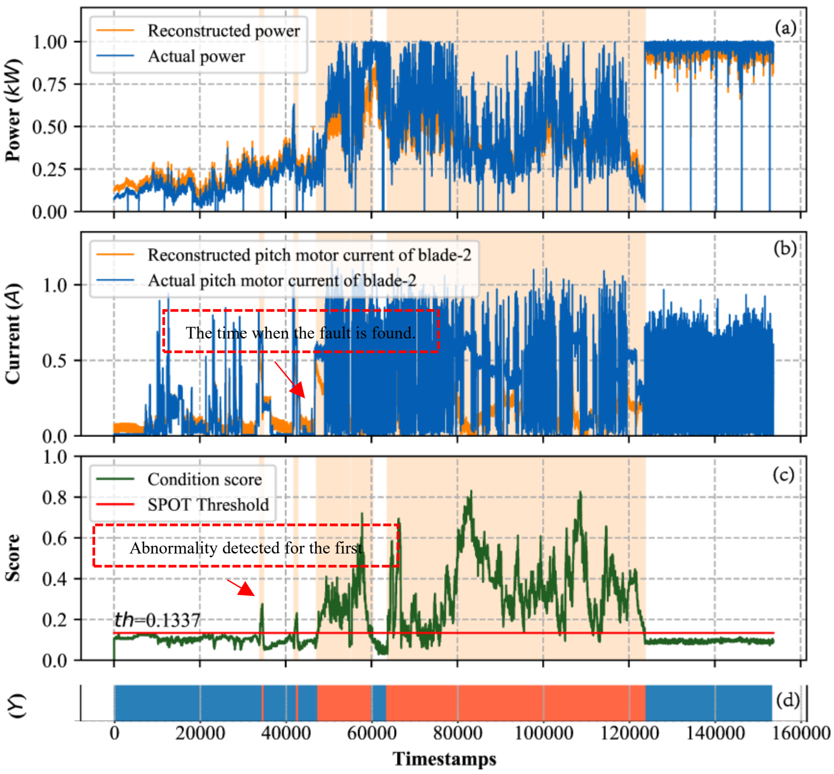
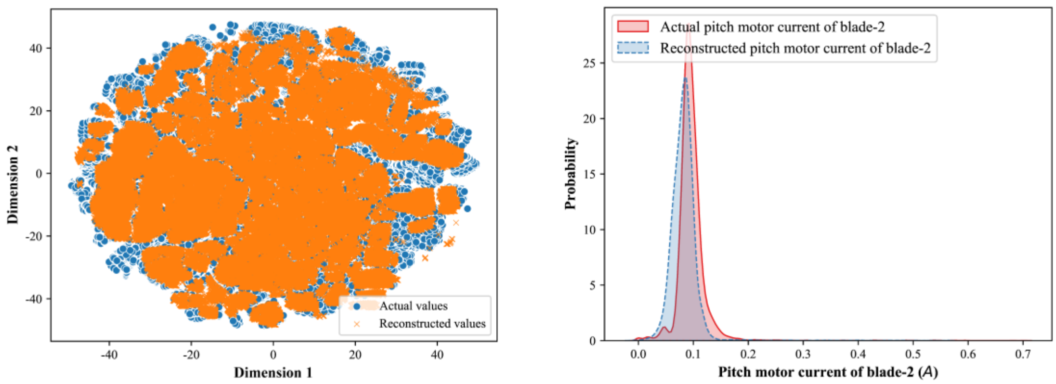


Figure 14 Fault detection result of the $WF_2 - WT_2$

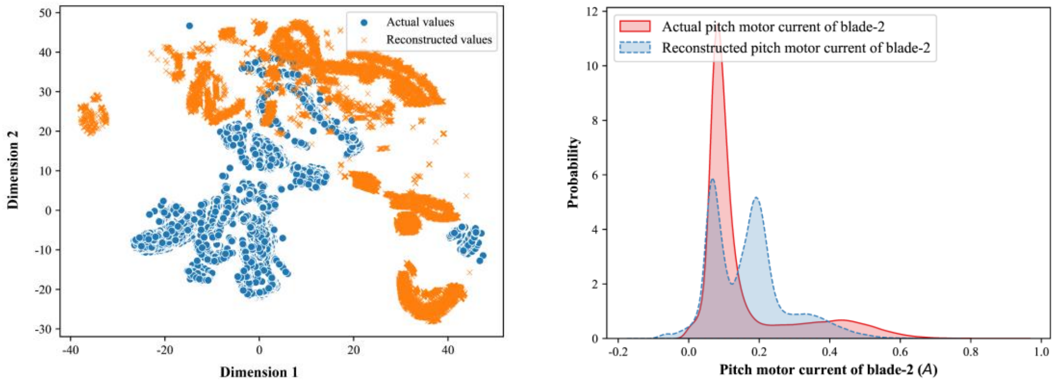
The probability density distribution of actual values and reconstructed values of $WF_2 - WT_2$ within the normal and abnormal periods are given in **Figure 15** to **Figure 16**. As shown in **Figure 15**(a), under the normal condition, the distributions of actual values and reconstructed values of all selected 32 parameters are consistent after t-SNE dimension reduction. Taking the pitch motor current of blade-2 as an example, it can be seen from **Figure 15**(b) that the current values of blade-2 are mainly distributed in the range of 0 to 0.2A, indicating that the proposed model has an accurate expression capability under the normal operation. However, within the abnormal period, **Figure 16** shows the clear differences between the distributions of actual values and reconstructed values after t-SNE dimension reduction. The actual values of the pitch motor current of blade-2 are shifted to the range of 0 to 0.65A under the abnormal operation.



(a) t-SNE distribution of all parameters

(b) Probability density distribution of pitch motor current of blade-2

Figure 15 The distribution of actual value and reconstructed value from $WF_2 - WT_2$ under normal conditions



(a) t-SNE distribution of all parameters **(b)** Probability density distribution of pitch motor current of blade-2

Figure 16 The distribution of actual value and reconstructed value from $WF_2 - WT_2$ under abnormal conditions

4.2.3 Discussions of cases

From the results of the above two representative cases, we can see that, the model proposed in this paper can accurately detect the potential anomalies, although there exist large differences in data in the wind farms due to the differences in turbine type, installation area, working environment, fault type, and data collection frequency. For these cases, our methods can be adapted to the input with variation in data dimension, which is significant for real-world WT condition monitoring, because it is always hard to obtain the variables consistent with the faults. However, it is worth noting that many variables may affect the accuracy of the prediction model due to data redundancy. Moreover, different data collection frequency indicates the large differences in the contained information. Use of the lower sampling frequency leads to domination of the trend-varying information of the data, while use of the higher sampling frequency produces both low and high frequency components of the data. Hence, when the data collection frequency is different, we need to set different model hyper-parameters, mainly including sliding window length and number of network layers.

4.3 Hyperparameter effect on the model performance

We have studied the effects of different parameters on model performance. Here, we take the experimental results of the datasets WF_1-WT_2 and WF_2-WT_2 as examples to undertake the performance analysis. It is hoped that the model can capture spatial-temporal features under normal conditions as much as possible, thus providing an effective normal behavior model of the WT. However, for abnormal data, the opposite properties are required, that is to say, it is better if the reconstructed abnormal values deviate from the measured data. Therefore, the selection of hyperparameters would be crucial for appropriate modelling. We mainly focus on the effect from the number of layers of spatial-temporal feature extraction network composed of GAT and TCN and the sliding window width on the model performance, because the number of BiLSTM layers is found to have relatively little impact. Meanwhile, $F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$ is used to evaluate the performance of the model, where $Precision = \frac{TP}{TP+FP}$, $Recall = \frac{TP}{TP+FN}$. TP, FP and FN refer to true positives, false positives, and false negatives, respectively.

Figure 17 and **Figure 18** show the effects of different sliding window widths and the number of spatial-temporal network layers on $F1$ score. For the WF_1-WT_2, when the window width is 120, F1 reaches 0.9306; however, when the window width is 10, F1 reduces to 0.8866. This is because a longer time window contains more temporal features. It can be seen from **Figure 18** that F1 reaches the optimal value 0.9257 when the number of network layer is 2. For the WF_2-WT_2, when the window width is 60 and the number of network layer is 1, F1 reaches the optimal 0.9361 and 0.937, respectively. F1 does not increase further with the increase of network layer number, indicating that the number of network layer has little effect on the model performance. Note that the gradual increase in network layer number clearly results in the increase in computation load and possibility of the overfitting. Hence, in the practical applications, it is appropriate to select 1 to 2 layers of the model. The parameter configuration of MSTFAN proposed in this paper is given in Table 4.

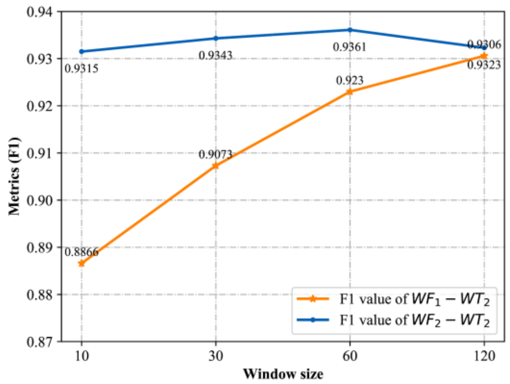


Figure 17 Detection accuracy with different widths of the sliding window

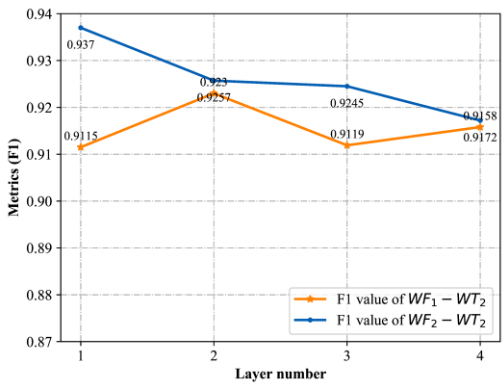


Figure 18 Detection accuracy with different layers of the proposed model

5. Performance comparison

To further verify the accuracy and effectiveness, the proposed MSTFAN method is compared with TCN [58], LSTM, LSTM-VAE [64], CNN-LSTM and CNN-GRU [65]. All parameters of these deep learning methods are kept to be consistent with the proposed MSTFAN method in this paper. The TCN model is a 1-D dilated causal convolution-based model, which enables a better extraction of spatial and temporal features from multivariate SCADA data. For the LSTM model, it is a 4-layer LSTM network, a typical model for temporal feature extraction, which can capture potential temporal dependency information from SCADA data. The LSTM-VAE-based model is used for multimodal fault detection. The potential distribution of multivariate spatial-temporal signals is modelled and then the information is reconstructed using the expected distribution. The CNN-LSTM method performs similarly to the spatial-temporal feature extraction method proposed in this paper, by utilizing the CNN network and LSTM network to extract spatial features and temporal features, respectively. The CNN-GRU method performs similarly to CNN-LSTM since GRU is a variant model of LSTM. Omitting the output gate in the model structure can make the number of GRU parameters fewer, thus making the training easier.

Table 4 Parameter configuration of MSTFAN

layer		Filter (Dropout)	Channels	Heads	Kernel (Strides)	Activation	Padding (Dilation)
Spatial network	GATConv		10/15	4	/	/	/
	Linear	64			/	/	/
	GATConv		10/15	4	/	/	/

	Linear	64	/	/	/	/	/
Temporal network	Conv1d	128（20）	/	/	7（1）	ReLU	6（1）
	Conv1d	128（20）	/	/	7（1）	ReLU	6（2）
Reconstruction network	BiLSTM	150（20）	/	/	/	/	/
	Linear	10/15	/	/	/	/	/

In addition, for all methods, the batch size is set to 128; the initial learning rate is 0.001; the maximum number of iterations is 100; and the early stopping mechanism is used to prevent the model from overfitting. To obtain a better convergence performance of the model, ReduceLROnPlateau, a kind of learning rate adjustment method based on epoch training times, is adopted for dynamic adjustment of learning rate. Meanwhile, $Precision$ ($Pre.$), $Recall$ ($Rec.$), $F1$ and area under ROC curve (AUC) are used as evaluation indexes. Here the receiver operating characteristic (ROC) curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.

Table 5 Performance comparison of all methods for WF_1 and WF_2

	WF_1-WT_2				WF_2-WT_2				Average			
Methods	**F1**	**Pre.**	**Rec.**	**AUC**	**F1**	**Pre.**	**Rec.**	**AUC**	**F1**	**Pre.**	**Rec.**	**AUC**
TCN	0.879	0.862	0.898	0.949	0.922	0.952	0.894	0.983	0.901	0.907	0.896	0.966
LSTM	0.820	0.970	0.713	0.925	0.903	0.958	0.854	0.982	0.861	0.964	0.784	0.953
LSTM-AE	0.719	0.837	0.642	0.930	0.781	**0.968**	0.661	0.966	0.750	0.903	0.652	0.948
CNN-LSTM	0.893	0.971	0.827	0.970	0.868	0.964	0.791	0.982	0.880	**0.968**	0.809	0.976
CNN-GRU	0.848	**0.972**	0.756	0.964	0.892	0.953	0.842	0.978	0.870	0.962	0.799	0.971
MSTFAN (proposed)	**0.931**	0.963	**0.900**	**0.980**	**0.932**	0.953	**0.913**	**0.983**	**0.931**	0.958	**0.907**	**0.981**

Table 5 shows the performance of all methods against these two datasets $WF_1 - WT_2$ and $WF_2 - WT_2$. Figure 19 graphically illustrates the performance comparison of the results from all the models investigated, where the optimal performance of the metric is shown in bold. It can be seen that the overall performances in terms of F1 and AUC are 0.931 and 0.981, respectively, which are produced from MSTFAN. The proposed MSTFAN is more optimal than those other methods, indicating that MSTFAN can effectively capture the complex spatial-temporal features of multivariate SCADA data and has a better generalization performance. Notably, the recall of MSTFAN is 0.907, which is also more optimal than other models. High recall indicates the model has a higher true negatives (TN) value, meaning that the model has a higher detection accuracy for abnormal data. This would be essential for fault detection. Furthermore, CNN-LSTM and CNN-GRU also present good performances whose AUC values are over 0.970; however, their average F1 (0.880 and 0.870, respectively) is still lower than that of the MSTFAN, and even worse than that of the TCN prediction network alone. It indicates that although CNN has a good local feature extraction capability, its ability to extract temporal correlations is weak when compared with the TCN. Meanwhile, CNN deals with problems under the Euclidean space, which may not be enough for complex multivariate time series data. We also notice that the average F1 of LSTM-AE is only 0.750, which is the worst performance in all models. This may be caused by abnormal patterns due to the trend anomaly of the variables and correlation anomaly among the variables. The AE-based methods use the reconstruction error to achieve the condition monitoring; however, sometimes the abnormal patterns may be allowed to be reconstructed. This fully illustrates the importance that the MSTFAN adopts the graph attention network to model specifically the spatial correlation of multivariate data, thus improving the detection accuracy.

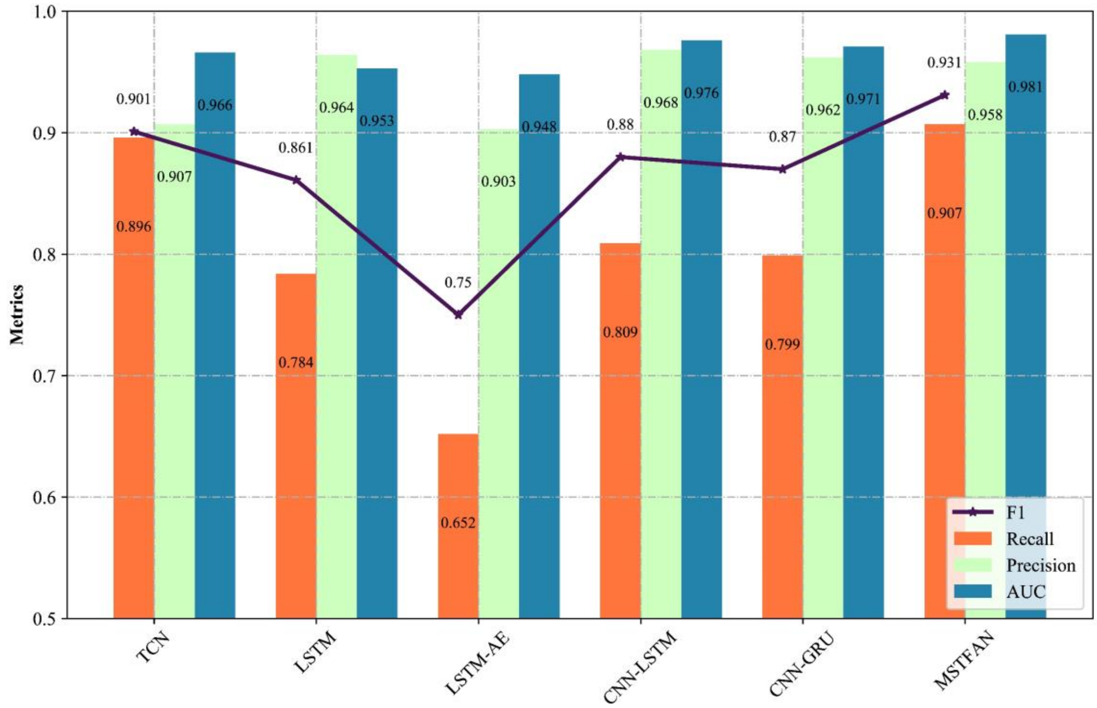


Figure 19 Performance comparison of the average results of different methods

Figure and **Figure** show the comparison of the actual and reconstructed values, taking main bearing temperature 2 from $WF_1 - WT_2$ and pitch motor current of blade-2 from $WF_2 - WT_2$ as examples. It can be seen that the proposed MSTFAN method has the best performances, although TCN and LSTM both present comparable fitting effect for the normal data. CNN-LSTM and CNN-GRU do not manifest perfect effects, possibly because these two methods only consider spatial correlation at the front-end of network while ignoring temporal correlation of data at the back-end of network. Moreover, although LSTM and TCN networks have good fitting effects on the normal state of WTs, it can be seen from **Figure** that the values of F1 and AUC are relatively low. This indicates that network over-fitting may happen, which is disadvantageous to fault detection. The above analyses verify the necessity of spatial-temporal feature extraction from SCADA data by combining graph network and temporal network. Meanwhile, the use of BiLSTM network has further improved the extraction ability of the model due to long-term dependency correlation of temporal data, making the proposed MSTFAN achieve an optimal performance in terms of accuracy and generalization.

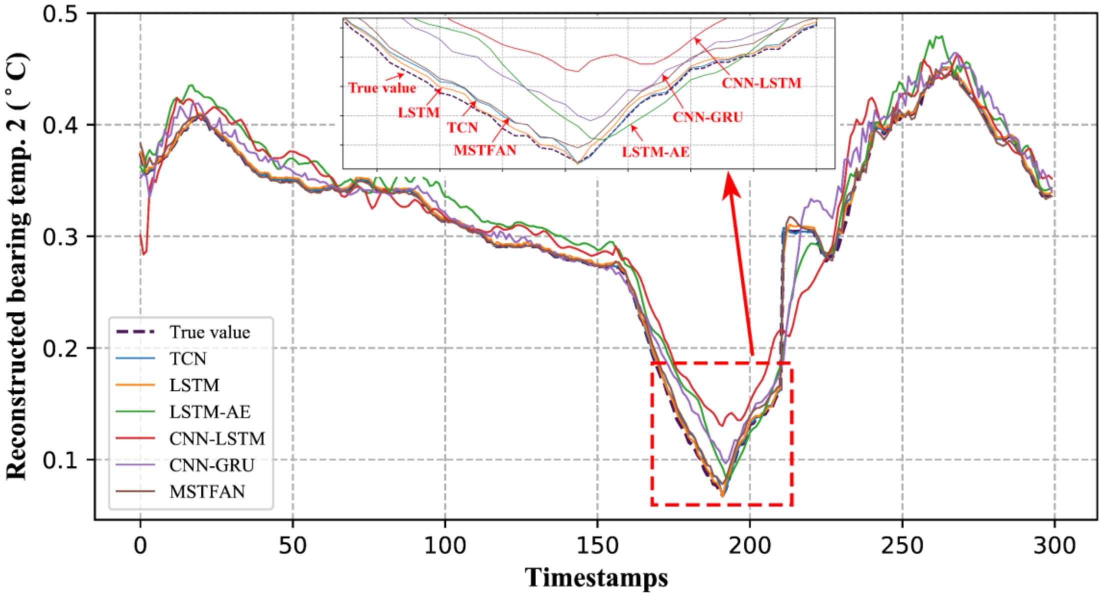


Figure 20 Comparison of the actual value and reconstructed value for main bearing temperature 2 of $WF_1 - \mathrm{WT}_2$

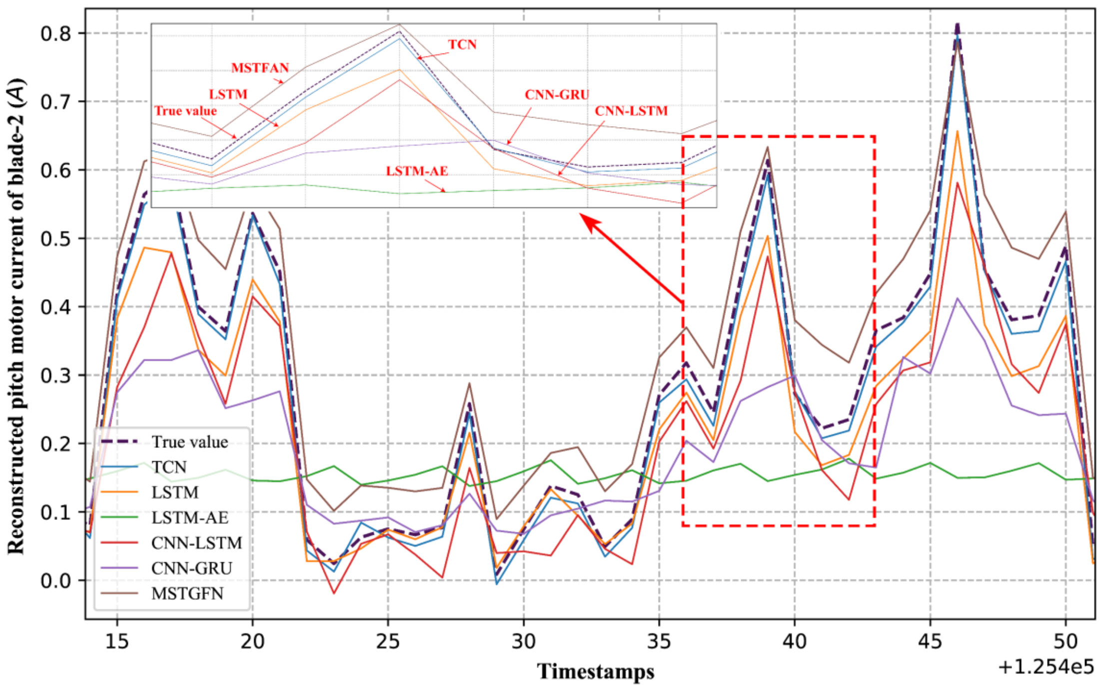


Figure 21 Comparison of the actual value and reconstructed value for pitch motor current of blade-2 of $WF_2 - \mathrm{WT}_2$

6. Conclusions

A novel approach is proposed for fault detection of WTs based on spatial-temporal information aggregation in this paper. It combines a graph neural network with a temporal convolution neural network to extract simultaneously spatial features and temporal features from multivariate SCADA data. A reconstruction network by means of BiLSTM network is utilized to capture the bi-directional information dependence to achieve an optimal prediction accuracy for the normal operation conditions of WTs. Furthermore, a condition scoring method is proposed based on reconstruction error, and DPs is further designed to improve the detection reliability for the early warning of the faults. The experimental results demonstrate that the proposed MSTFAN model can effectively detect early WT faults and be convenient for users to detect anomalies in the condition monitoring of real-world WTs. For future studies, the detection accuracy and generalization capability of the model need to be optimized further by the proper selection of monitoring parameters. Moreover, practical data with more extensive fault cases should be collected for training and verifying the model to improve the robustness of the approach.

Acknowledgment

The work is supported by National Natural Science Foundation of China (62006236), NUDT Research Project (ZK20-10), National Key Research and Development Program of China (2020YFA0709803), Hunan Provincial Natural Science Foundation (2020JJ5673), National Science Foundation of China (U1811462), National Key R&D project by Ministry of Science and Technology of China (2018YFB1003203), and Autonomous Project of HPCL (201901-11, 202101-15).

References

1. Zhao, J., Patwary, A. K., Qayyum, A., Alharthi, M., Bashir, F., Mohsin, M., Hanif, I. and Abbas, Q., The determinants of renewable energy sources for the fueling of green and sustainable economy. *Energy* **2022,** *238*, 122029.
2. GWEC *Global Wind 2021 Report*; https://gwec.net/global-wind-report-2021/. 2021.
3. Crabtree, C. J., Zappalá, D. and Hogg, S. I., Wind energy: UK experiences and offshore operational challenges. *Proceedings of the Institution of Mechanical Engineers, Part A: Journal of Power and Energy* **2015,** *229* (7), 727-746.
4. Ren, Z., Verma, A. S., Li, Y., Teuwen, J. J. and Jiang, Z., Offshore wind turbine operations and maintenance: A state-of-the-art review. *Renewable and Sustainable Energy Reviews* **2021,** *144*, 110886.
5. Zhou, P. and Yin, P., An opportunistic condition-based maintenance strategy for offshore wind farm based on predictive analytics. *Renewable and Sustainable Energy Reviews* **2019,** *109*, 1-9.
6. Yang, W., Tavner, P. J., Crabtree, C. J., Feng, Y. and Qiu, Y., Wind turbine condition monitoring: technical and commercial challenges. *Wind Energy* **2014,** *17* (5), 673-693.
7. Qiu, Y., Feng, Y. and Infield, D., Fault diagnosis of wind turbine with SCADA alarms based multidimensional information processing method. *Renewable energy* **2020,** *145* (Jan.), 1923-1931.
8. Astolfi, D., Pandit, R., Terzi, L. and Lombardi, A., Discussion of Wind Turbine Performance Based on SCADA Data and Multiple Test Case Analysis. *Energies* **2022,** *15* (15), 5343.
9. Romero, A., Soua, S., Gan, T. H. and Wang, B., Condition Monitoring of a wind turbine drive train based on its power dependant vibrations. *Renewable Energy* **2018,** *123*.
10. Guo, P., Infield, D. and Yang, X., Wind turbine generator condition-monitoring using temperature trend analysis. *IEEE Transactions on sustainable energy* **2011,** *3* (1), 124-133.
11. Swiszcz, G., Cruden, A., Booth, C. and Leithead, W. In *A data acquisition platform for the development of a wind turbine condition monitoring system*, 2008 international conference on condition monitoring and diagnosis, IEEE: 2008; pp 1358-1361.
12. Colherinhas, G. B., de Morais, M. V. G. and Machado, M. R., Spectral Model of Offshore Wind Turbines and Vibration Control by Pendulum Tuned Mass Dampers. *International Journal of Structural*

Stability and Dynamics **2022,** *22* (05), 2250053.

13. Romero, A., Soua, S., Gan, T.-H. and Wang, B., Condition monitoring of a wind turbine drive train based on its power dependant vibrations. *Renewable energy* **2018,** *123*, 817-827.
14. Benbouzid, M., Berghout, T., Sarma, N., Djurović, S., Wu, Y. and Ma, X., Intelligent Condition Monitoring of Wind Power Systems: State of the Art Review. *Energies* **2021,** *14* (18), 5967.
15. Yang, W., Court, R. and Jiang, J., Wind turbine condition monitoring by the approach of SCADA data analysis. *Renewable Energy* **2013,** *53*, 365-376.
16. Qiu, Y., Feng, Y. and Infield, D., Fault diagnosis of wind turbine with SCADA alarms based multidimensional information processing method. *Renewable Energy* **2020,** *145*, 1923-1931.
17. Tautz-Weinert, J. and Watson, S. J., Using SCADA Data for Wind Turbine Condition Monitoring – a Review. *IET Renewable Power Generation* **2017,** *11* (4), 382-394.
18. Dao, P. B., Condition monitoring and fault diagnosis of wind turbines based on structural break detection in SCADA data. *Renewable Energy* **2022,** *185*, 641-654.
19. Chatterjee, J. and Dethlefs, N., Scientometric review of artificial intelligence for operations & maintenance of wind turbines: The past, present and future. *Renewable and Sustainable Energy Reviews* **2021,** *144*, 111051.
20. Peng, Q., Ma, X. and Cross, P., An integrated data-driven model-based approach to condition monitoring of the wind turbine gearbox. *IET Renewable Power Generation* **2017,** *11* (9), 1177-1185.
21. Zhan, J., Wu, C., Ma, X., Yang, C., Miao, Q. and Wang, S., Abnormal vibration detection of wind turbine based on temporal convolution network and multivariate coefficient of variation. *Mechanical Systems and Signal Processing* **2022,** *174*, 109082.
22. Feng, Y., Qiu, Y., Crabtree, C. J., Long, H. and Tavner, P. J., Use of SCADA and CMS signals for failure detection & diagnosis of a wind turbine gearbox. **2011**.
23. Dong, Y., Fang, F. and Gu, Y., Dynamic evaluation of wind turbine health condition based on Gaussian mixture model and evidential reasoning. *Journal of Renewable and Sustainable Energy* **2013,** *5* (3), 033117.
24. Chen, J., Li, J., Chen, W., Wang, Y. and Jiang, T., Anomaly detection for wind turbines based on the reconstruction of condition parameters using stacked denoising autoencoders. *Renewable Energy* **2020,** *147*, 1469-1480.
25. Stetco, A., Dinmohammadi, F., Zhao, X., Robu, V., Flynn, D., Barnes, M., Keane, J. and Nenadic, G., Machine learning methods for wind turbine condition monitoring: A review. *Renewable energy* **2019,** *133* (APR.), 620-635.
26. Schlechtingen, M. and Santos, I. F., Comparative analysis of neural network and regression based condition monitoring approaches for wind turbine fault detection. *Mechanical Systems and Signal Processing* **2011,** *25* (5), 1849-1875.
27. Zhao, H., Wu, Q., Guo, Q., Sun, H. and Xue, Y., Distributed Model Predictive Control of a Wind Farm for Optimal Active Power ControlPart I: Clustering-Based Wind Turbine Model Linearization. *IEEE Transactions on Sustainable Energy* **2017,** *6* (3), 831-839.
28. Kong, Y., Wang, T., Feng, Z. and Chu, F., Discriminative dictionary learning based sparse representation classification for intelligent fault identification of planet bearings in wind turbine. *Renewable energy* **2020,** *152* (Jun.), 754-769.
29. Trizoglou, P., Liu, X. and Lin, Z., Fault detection by an ensemble framework of Extreme Gradient Boosting (XGBoost) in the operation of offshore wind turbines. *Renewable Energy* **2021**.
30. Schlechtingen, M. and Ferreira Santos, I., Comparative analysis of neural network and regression

based condition monitoring approaches for wind turbine fault detection. *Mechanical Systems and Signal Processing* **2011,** *25* (5), 1849-1875.

31. Tang, B., Song, T., Li, F., Deng, L., Kalogirou, S. A. and Christodoulides, P., Fault diagnosis for a wind turbine transmission system based on manifold learning and Shannon wavelet support vector machine. **2014**.

32. Liu, WY, Wang, ZF, Han, JG and GF, Wind turbine fault diagnosis method based on diagonal spectrum and clustering binary tree SVM. *RENEW ENERG* **2013**.

33. Castellani, F., Astolfi, D., Sdringola, P., Proietti, S. and Terzi, L., Analyzing wind turbine directional behavior: SCADA data mining techniques for efficiency and power assessment. *Applied Energy* **2015,** *185* (pt.2), 1076-1086.

34. Aka, C., Mha, B., Mfha, D., Ka, E., Mm, A., Hn, A. and Mn, F., Hidden Markov model based principal component analysis for intelligent fault diagnosis of wind energy converter systems. *Renewable Energy* **2020,** *150*, 598-606.

35. Tang, M., Zhao, Q., Ding, S. X., Wu, H., Li, L., Long, W. and Huang, B., An Improved LightGBM Algorithm for Online Fault Detection of Wind Turbine Gearboxes. *Energies* **2020,** *13*.

36. Biswas, D. In *Adapting shallow and deep learning algorithms to examine production performance–Data analytics and forecasting*, SPE/IATMI Asia Pacific Oil & Gas Conference and Exhibition, OnePetro: 2020.

37. Zhong, G., Ling, X. and Wang, L. N., From shallow feature learning to deep learning: Benefits from the width and depth of deep architectures. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **2019,** *9* (1), e1255.

38. Yan, X., Liu, Y. and Jia, M., Multiscale cascading deep belief network for fault identification of rotating machinery under various working conditions. *Knowledge-Based Systems* **2020**, 105484.

39. Cao, L., Qian, Z., Zareipour, H., Huang, Z. and Zhang, F., Fault Diagnosis of Wind Turbine Gearbox Based on Deep Bi-Directional Long Short-Term Memory Under Time-Varying Non-Stationary Operating Conditions. *IEEE Access* **2019,** *7*, 1-1.

40. Xya, B., Ying, L. A. and Mj, B., Multiscale cascading deep belief network for fault identification of rotating machinery under various working conditions - ScienceDirect. *Knowledge-Based Systems 193*.

41. Lei, J., Liu, C. and Jiang, D., Fault diagnosis of wind turbine based on Long Short-term memory networks. *Renewable Energy* **2019,** *133* (APR.), 422-432.

42. Wu, Y. and Ma, X., A hybrid LSTM-KLD approach to condition monitoring of operational wind turbines. *Renewable Energy* **2022,** *181*.

43. Jiang, G., He, H., Yan, J. and Xie, P., Multiscale Convolutional Neural Networks for Fault Diagnosis of Wind Turbine Gearbox. *IEEE Transactions on Industrial Electronics* **2018,** *66* (4), 3196-3207.

44. Xiang, L., Wang, P., Yang, X., Hu, A. and Su, H., Fault detection of wind turbine based on SCADA data analysis using CNN and LSTM with attention mechanism. *Measurement* **2021,** *175* (8), 109094.

45. He, Q., Pang, Y., Jiang, G. and Xie, P., A spatio-temporal multiscale neural network approach for wind turbine fault diagnosis with imbalanced SCADA data. *IEEE Transactions on Industrial Informatics* **2020,** *PP* (99), 1-1.

46. Xiang, L., Yang, X., Hu, A., Su, H. and Wang, P., Condition monitoring and anomaly detection of wind turbine based on cascaded and bidirectional deep learning networks. *Applied Energy* **2022,** *305*.

47. Xu, K., Hu, W., Leskovec, J. and Jegelka, S., How Powerful are Graph Neural Networks? **2018**.

48. Kip F , T. N. and Welling, M., Semi-Supervised Classification with Graph Convolutional Networks. **2016**.

49. Velikovi, P., Cucurull, G., Casanova, A., Romero, A., Liò, P. and Bengio, Y., Graph Attention Networks. **2017**.

50. Guo, S., Lin, Y., Feng, N., Song, C. and Wan, H., Attention Based Spatial-Temporal Graph Convolutional Networks for Traffic Flow Forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence* **2019,** *33*, 922-929.

51. Diao, Z., Wang, X., Zhang, D., Liu, Y. and He, S., Dynamic Spatial-Temporal Graph Convolutional Neural Networks for Traffic Forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence* **2019,** *33*, 890-897.

52. Scheinert, D. and Acker, A., TELESTO: A Graph Neural Network Model for Anomaly Classification in Cloud Services. **2021**.

53. Deng;, A. and Hooi., B., Graph Neural Network-Based Anomaly Detection in Multivariate Time Series. *aaai2021* **2021**.

54. Su, X., Shan, Y., Li, C., Mi, Y., Fu, Y. and Dong, Z., Spatial-temporal attention and GRU based interpretable condition monitoring of offshore wind turbine gearboxes. *IET Renewable Power Generation* **2022,** *16* (2), 402-415.

55. Chandola, V., Banerjee, A. and Kumar, V., Anomaly Detection: A Survey. *ACM Computing Surveys* **2009,** *41* (3).

56. Jannis, Tautz-Weinert, Simon, J. and Watson, Using SCADA data for wind turbine condition monitoring – a review. *IET Renewable Power Generation* **2017,** *11* (4), 382-394.

57. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I., Attention Is All You Need. *arXiv* **2017**.

58. Bai, S., Kolter, J. Z. and Koltun, V., An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv preprint arXiv:1803.01271.* **2018**.

59. Zhao, H., Wang, Y., Duan, J., Huang, C. and Zhang, Q., Multivariate Time-series Anomaly Detection via Graph Attention Network. **2020**.

60. Chen, H., Liu, H., Chu, X., Liu, Q. and Xue, D., Anomaly detection and critical SCADA parameters identification for wind turbines based on LSTM-AE neural network. *Renewable Energy* **2021,** *172* (1).

61. Zhang, Z.-Y. and Wang, K.-S., Wind turbine fault detection based on SCADA data analysis using ANN. *Advances in Manufacturing* **2014,** *2* (1), 70-78.

62. Wei, L., Qian, Z. and Zareipour, H., Wind turbine pitch system condition monitoring and fault detection based on optimized relevance vector machine regression. *IEEE Transactions on Sustainable Energy* **2019,** *11* (4), 2326-2336.

63. Wei, L., Qian, Z., Zareipour, H. and Zhang, F., Comprehensive aging assessment of pitch systems combining SCADA and failure data. *IET Renewable Power Generation* **2022,** *16* (1), 198-210.

64. Park, D., Hoshi, Y. and Kemp, C. C., A Multimodal Anomaly Detector for Robot-Assisted Feeding Using an LSTM-based Variational Autoencoder. *IEEE Robotics and Automation Letters* **2017,** *PP* (99).

65. Kong, Z., Tang, B., Deng, L., Liu, W. and Han, Y., Condition monitoring of wind turbines based on spatio-temporal fusion of SCADA data by convolutional neural networks and gated recurrent units. *Renewable Energy* **2020,** *146*.