SGG-DGCN: Wind Turbine Anomaly Identification by Using Deep Graph Convolutional Networks with Similarity Graph Generation Strategy

Xiaomin Wang[1], Di Zhou[1]*, Xiao Zhuang[1], Jian Ge[1] and Jiawei Xiang[1]

[1] College of Mechanical and Electrical Engineering, Wenzhou University, Wenzhou 325035

*Corresponding author: zhoudi@wzu.edu.cn

Received Month X, XXXX | Accepted Month X, XXXX | Posted Online Month X, XXXX

In order to minimize wind turbine failures, fault diagnosis of wind turbines is becoming increasingly important, deep learning methods excel at multivariate monitoring and data modeling, but they are often limited to Euclidean space and struggle to capture the complex coupling between wind turbine sensors. To address this problem, we convert SCADA data into graph data, where sensors act as nodes and their topological connections act as edges, to represent these complex relationships more efficiently. Specifically, a wind turbine anomaly identification method based on deep graph convolutional neural network using similarity graph generation strategy (SGG-DGCN) is proposed. Firstly, a plurality of similarity graphs containing similarity information between nodes are generated by different distance metrics. Then, the generated similarity graphs are fused using the proposed similarity graph generation strategy. Finally, the fused similarity graphs are fed into the DGCN model for anomaly identification. To verify the effectiveness of the proposed SGG-DGCN model, we conducted a large number of experiments. The experimental results show that the proposed SGG-DGCN model has the highest accuracy compared with other models. In addition, the results of ablation experiment also demonstrate that the proposed SGG strategy can effectively improve the accuracy of WT anomaly identification.

Keywords: Wind turbine, Deep graph convolutional networks, Similarity graph generation, Anomaly identification

Introduction

In recent years, the global ecological environment has been deteriorating and the supply of fossil fuels is also gradually insufficient, so renewable energy has become a key research area for domestic and foreign researchers [1],[2]. Currently, wind energy as a renewable energy has the advantages of zero pollution and zero emission. However, in the actual utilization of wind energy, due to the remote location of wind farm installation, harsh working environment and changing working conditions and other factors, wind farm operation and maintenance (O&M) costs are often high.

High O&M costs are increasingly becoming a major obstacle to the sustainable development of the wind power industry [3]. Therefore, how to develop effective fault diagnosis techniques for wind turbines, reduce the maintenance cost of wind turbines and improve the reliability of wind turbines has become a hot and difficult issue for research in the field of wind turbines in recent years. Currently, the main methods commonly used in the field of wind turbine fault diagnosis include vibration analysis and oil fluid analysis, as well as intelligent diagnostic methods based on supervisory control and data acquisition (SCADA) data [4]. However, the practical application of the first two methods requires the installation of sensors, which adds additional costs to the implementation of the methods. In contrast, SCADA data contains rich information about the operating status and is widely used in the field of wind turbine fault diagnosis.

A novel wind turbine condition monitoring method based on temporal and spatial feature fusion of SCADA data with convolutional neural network (CNN) and DGCN recurrent unit (GRU) was proposed by Kong Z et al. [5]. Chen H et al. [6] proposed a method based on Long Short-Term Memory (LSTM) and Autoencoder (AE) neural networks for evaluating continuous condition monitoring data from wind turbines using SCADA parameters. Zhang Chen et al. [7] proposed a wind turbine anomaly detection and diagnosis method using Long Short-Term Memory-based Stacked Denoising Self-Encoder (LSTM-SDAE) and Extreme Gradient Boost (XGBoost) to realize the wind turbine anomaly detection. Wen and Xu [8] proposed a hybrid fault diagnosis method based on Relief, principal component analysis and deep neural networks, by using part of the SCADA system data as input parameters. It is evident from the above literature that the use of SCADA data is often accompanied by the design and use of algorithms for deep learning methods. Deep learning algorithms in the process of data analysis and mining mainly focus on the input of vector type and can effectively deal with numerical features. However, deep learning-based models have limitations for handling complex data in non-Euclidean space with multiple subsystems. In wind turbine operation, faults may cause multiple sensor parameters to change simultaneously, such as temperature, oil temperature, inlet pressure, etc., when a gearbox bearing fails. These complex interactions involve multiple types of information relationships. Therefore, it is necessary to investigate more effective methods for modeling SCADA data in non-Euclidean space to explore the complex relationships of the data. Graph neural networks (GNNs) are widely used in various fields due to the unique advantages of graphs in terms of data structure and relational representation.

Currently, graph neural networks (GNNs) pay more attention to the connection relationship of data [9]. The concept of GNNs was first proposed by Scarselli et al. Its goal is to establish a specific network connection for the data stored in graph domain. Some traditional neural networks such as CNNs do not have translation invariance in non-Euclidean structures (the same size convolution kernels can not be used for convolution). Accordingly, graph convolutional networks (GCNs) began to process such data, which makes it possible to perform convolution operations on irregular graph structures [10].

The similarity between the samples generates the adjacency matrix to be used by the GCN to achieve anomaly identification. Generally, the similarity between nodes is calculated as elements of the adjacency matrix by using a distance metric. Different distance

calculation methods note different information. Based on the distance between nodes the adjacency matrices can be constructed using metrics such as for example the k-nearest neighbor method [11], cosine similarity [12] or the Mahalanobis distance [13]. Different distance calculation methods note different information. KNN measures the numerical differences in the input data by using Euclidean distances. Cosine similarity is used to measure the directional differences in the input data. Mahalanobis distance is used to measure the relative differences in the input data. Manhattan distance is used to measure spatial distance differences between input data. Since different datasets may have different feature distributions and complexities, a single distance metric may not be able to fully capture all aspects of the data. Therefore, this paper proposes a similarity graph generation strategy, which generates the adjacency matrix by using four distance metrics and then fuses the matrices, integrating the advantages of each method.

The main insights and contributions of this paper are summarized as follows:

(1) A wind turbine anomaly recognition model based on DGCN is proposed. The model effectively captures the complex features of wind turbine data by dynamically adjusting the graph structure, which significantly improves the accuracy of anomaly detection.

(2) A similarity graph generation strategy for constructing sample similarity graphs is proposed. Similarity graphs are generated by fusing multiple adjacency matrices generated by four common distance metrics. The constructed similarity graph fully represents the similarity between samples from multiple perspectives.

(3) The proposed SGG-DGCN method can deeply explore the correlation between the data to realize the identification of anomalous wind turbines. The proposed SGG-DGCN can provide support for operators to detect potential faults and safety hazards of wind turbine in time so as to reduce the risk of accidents in wind farms.

This paper is organized as follows. Section 2 introduces the theoretical background. Section 3 introduces the proposed matrices fusion strategy. Section 4 proposes the framework of the proposed model. Section 5 gives the verification of the superiority of the proposed model. Section 6 shows the conclusion of this paper.

Graph convolutional networks (GCN)

In recent years, researchers have introduced graph convolutional neural networks to extend CNNs to non-Euclidean data, such as graph data. The graph convolution operation works by aggregating the node feature information of nodes and their neighboring nodes in the graph to obtain implicit representations or labels of the nodes. In the convolution operation, the relationships between the nodes are mined for feature representation, new node features are formed and nodes are updated. Eventually, the new node features are used to predict the labels of the nodes. When solving node-level tasks, the main focus is on learning better local representations for each node. In this case, pooling operations for graph-level tasks are not necessary, so we only need to concentrate on building the convolutional operations on the graph.

In this study, we use GCN to capture the spatial correlation of wind power data flow. The role of GCN is the same as that of CNN, but the object of GCN is the graph data

consisting of nodes and edges. The structure of GCNs based on frequency domain convolution is shown in Fig. 1.

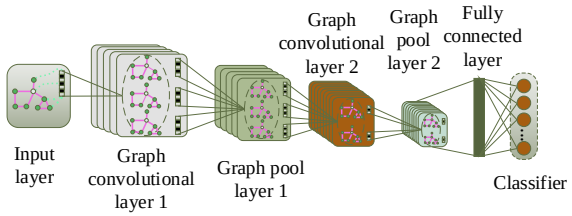

Fig. 1. The structure schematic diagram of GCN.

Assume that the graph data constructed from wind data has *n* nodes. Each node has a T-dimensional sequence of wind data which forms a matrix $\boldsymbol{X} \in \mathbb{R}^{n \times T}$. The edges between nodes form a topological relationship adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$. $\boldsymbol{X}$ and $\boldsymbol{A}$ will be used as inputs to the graph convolution model. GCN is a multilayer neural network, the propagation between layers is GCN is a multilayer neural network and the propagation between layers is as follows:

$$\boldsymbol{H}^{(l+1)} = \sigma(\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{H}^{(l)}\boldsymbol{W}^{(l)}) \quad (1)$$

Where, $\tilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{I}$, $\boldsymbol{I}$ is the array of units. $\boldsymbol{D}$ is the degree matrix, $\boldsymbol{D}_{ii} = \sum_j \tilde{\boldsymbol{A}}_{ij}$. $\boldsymbol{H}^{(l)}$ is the feature of each layer, $\boldsymbol{H}$=$\boldsymbol{X}$ at the input layer. $\boldsymbol{W}^{(l)}$ is the weight matrix. σ is the non-linear activation function Sigmoid. With two layers of GCN and sigmoid and ReLU, the overall propagation formula is as follows:

$$f(\boldsymbol{X}, \boldsymbol{A}) = \sigma[\tilde{\boldsymbol{A}}\,\mathrm{ReLU}(\tilde{\boldsymbol{A}}\boldsymbol{X}\boldsymbol{W}^{0})\boldsymbol{W}^{1}] \quad (2)$$

Where, $\tilde{\boldsymbol{A}} = \boldsymbol{D}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\boldsymbol{D}^{\frac{1}{2}}$, $\boldsymbol{W}^{0} \in \mathbb{R}^{T \times h}$ is the weight matrix from the input layer to the hidden layer, *h* is the number of cells in the hidden layer. $\boldsymbol{W}^{1} \in \mathbb{R}^{h \times t}$ is the weight matrix from the hidden layer to the output layer. $f(\boldsymbol{X}, \boldsymbol{A}) \in \mathbb{R}^{n \times t}$ is the output. Different from CNN, where the GCN convolution kernel acts in the nodes in the topology graph of the traffic network, the GCN also obtains a deep abstracted representation through the stacking of multiple graph convolution layers.

Similarity graph generation strategy

Graph data adjacency matrix generation methods typically use a single distance metric. However, different distance calculations pay attention to different information. A single distance metric may not fully capture the similarity information between samples. To address this shortcoming, this paper proposes a similarity graph generation (SGG) strategy to construct wind power data graph by fusing adjacency matrices constructed from multiple distance metrics. The proposed similarity graph generation strategy can capture the inter-sample similarity information more comprehensively. The wind power data co-similarity graph is represented as a graph $G = (V, E)$. The vertex V corresponds to the samples and the edge E is weighted according to the similarity of the samples. The flowchart of the SGG strategy is shown in Fig. 2.

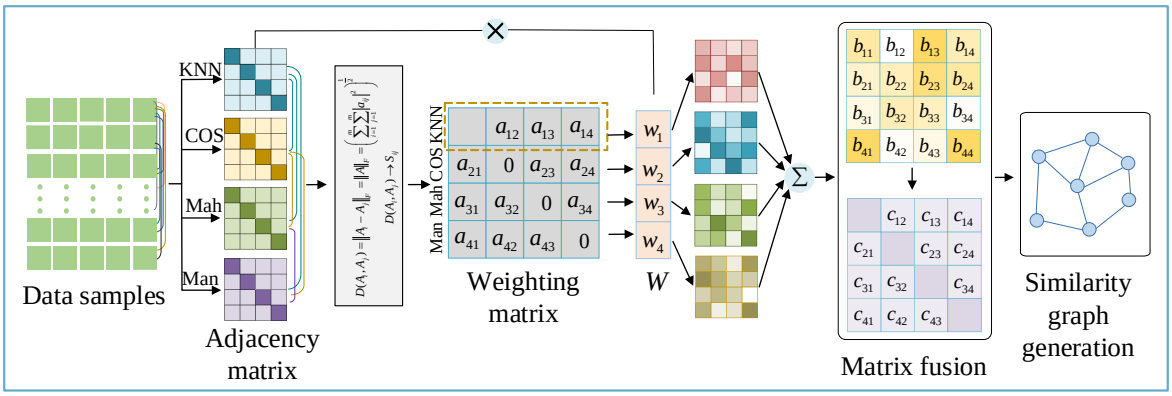

Fig. 2. The structural diagram of the similarity graph generation.

Multiple adjacency matrices are generated by KNN, cosine similarity, Mahalanobis distance and Manhattan distance. As shown in Eqs. (3-6).

$$L_{ij} = \left(\sum_{i=1}^{d} \left| h_i^{(L)} - h_j^{(L)} \right|^2 \right)^{\frac{1}{2}} \tag{3}$$
$$A_{ij}^{\mathrm{KNN}} = \mathrm{KNN}(k,\ L_{ij},\ \Omega_i)$$

$$A_{ij}^{\mathrm{COS}} = \cos(h_i, h_j) = \frac{h_i \cdot h_j}{\left|h_i\right|\left|h_j\right|} \tag{4}$$

$$\begin{aligned} A_{ij}^{\mathrm{Mah}} &= D_M(h_i, h_j) \\ &= \sqrt{(h_i - h_j)^T \varepsilon^{-1} (h_i - h_j)} \end{aligned} \tag{5}$$

$$A_{ij}^{\mathrm{Man}} = D_{man}(h_i, h_j) = \sum_{i=1}^{n} \left| h_i - h_j \right| \tag{6}$$

Where, h stands for node. Suppose we have adjacency matrices, denoted as $[A^{\mathrm{KNN}}, A^{\mathrm{COS}}, A^{\mathrm{Mah}}, A^{\mathrm{Man}}]$. We use the Frobenius paradigm to calculate the distance between matrices. The distance between matrices can be defined by Eq. (7).

$$d_{ij} = \left\| A_i - A_j \right\|_F \tag{7}$$

Where, d_{ij} denotes the distance between matrices. The distance is converted to similarity, where the closer the distance between the matrices represents a higher degree of similarity. The conversion of similarity distance can be defined by Eq. (8).

$$S_{ij} = \frac{1}{1 + d_{ij}} \tag{8}$$

Where, S_{ij} denotes the inter matrix similarity. The weight matrix can be defined by Eq. (9).

$$W = \begin{bmatrix} S_{11} & S_{12} & S_{13} & S_{14} \\ S_{21} & S_{22} & S_{23} & S_{24} \\ S_{31} & S_{32} & S_{33} & S_{34} \\ S_{41} & S_{42} & S_{43} & S_{44} \end{bmatrix} \tag{9}$$

Some elements of the matrix may have much larger values than others and will affect the final result. We normalize each element in the weight matrix W. The normalized weight matrix W can be defined by Eq. (10).

$$W_{normalized} = \frac{W - W_{\min}}{W_{\max} - W_{\min}} \tag{10}$$

An all-zero matrix F is created as the fused adjacency matrix and each adjacency matrix is traversed using a loop, which multiplies it

with the weights in the corresponding normalized weights W and adds the result to the fused matrix F. The fusion matrix F can be defined by Eq. (11).

$$F = \sum_{i=1}^{4} \left(W_{normalized_i} \times A_i \right) \qquad (11)$$

The fusion matrix F is transformed into a binary matrix using median thresholding.

Assign weights to each edge based on the fused matrices, the weights can be determined based on the values of the elements in the adjacency matrices. Node features and labels also need to be added to the graph. The graph data constructed from the fused adjacency matrices will be used as input to the deep graph convolutional networks for graph feature extraction. We considered the advantages of each distance metric to fuse the adjacency matrices. This improves the expressiveness and robustness of the graph data.

Framework

In this paper, we propose a similarity graph generation (SGG) strategy for data graph construction of deep graph convolutional networks (DGCN) to realize wind turbine anomaly identification. The detailed steps and the overall framework of the proposed method are shown in Fig. 3 and summarized as follows.

Step 1: Data acquisition and pre-processing. The wind turbine status monitoring data is collected by SCADA system. The sample data is formed after preprocessing.

Step 2: Similarity graph generation. Multiple adjacency matrices are generated by KNN, cosine similarity, Mahalanobis distance and Manhattan distance. The proposed SGG strategy fuses the generated adjacency matrices to generate the sample similarity graph.

Step 3: Node similarity feature extraction. Based on the constructed sample similarity graph, deep graph convolutional networks (DGCN) are used for feature extraction of similar nodes. Based on the extracted features, similar samples are identification.

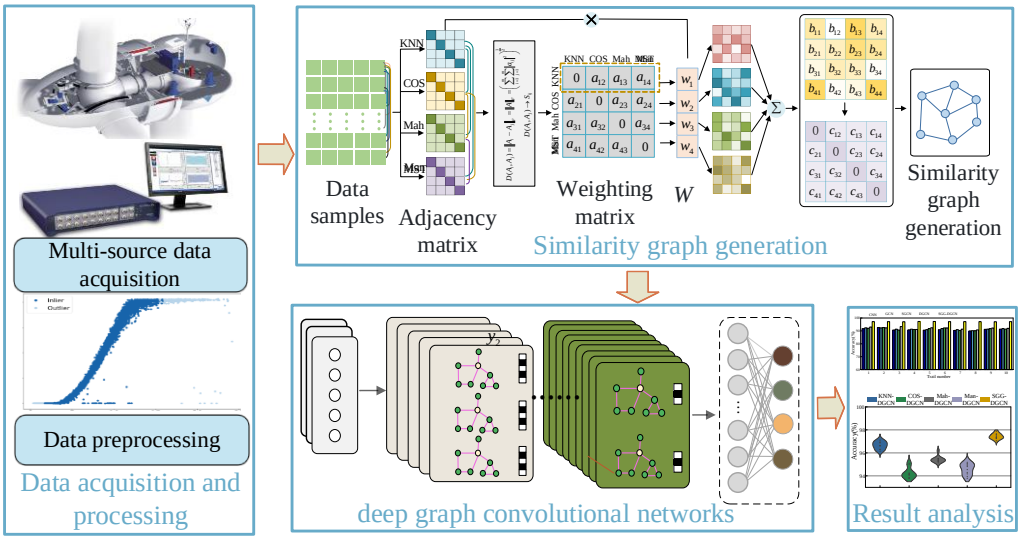

Fig. 3. Overall architecture of SGG-DGCN.

Experiments and analysis

Data description

In wind turbine, anomalies are primarily distinguished by three characteristics. The first type of anomaly is the “abandoned wind data”, where wind speeds exceeds the turbine's cut-in speed. This anomaly is characterized by a horizontal line on the power curve, indicating that the wind energy is not being utilized. The second type of anomaly is “overgeneration state data”, which is characterized by a dense band of output power exceeding the rated power. It can occur during periods of operation that exceed the design capacity of the turbine, which may shorten the life of the turbine. The third type of anomaly is “outliers”, which are isolated data points that deviate from the power curve due to sensor malfunctions or errors in data transmission. These anomalies are critical to identify and address for ensuring the optimal performance and longevity of wind turbines.

The data originated from the Penmanshiel wind farm [14]. The total data consisted of 14 turbines and was collected from 1 January 2021 to 1 July 2021, with samples taken every ten minutes. The wind speed, power and generator speed time series distributions are shown in Fig. 4. The correlation between wind speed, power and generator speed is shown in Fig. 5. The 14 turbines have a total of 364,910 data samples. All abnormal samples are selected as experimental samples. Considering that the number of normal samples far exceeds that of abnormal samples, it will lead to a poor classification effect on abnormal samples.

The WT SCADA systems are monitored by hundreds of sensors which include speed, power and temperature. To improve the accuracy of anomaly identification, key variables need to be selected. In this paper, Pearson's correlation coefficient is used to screen the variables. Parameters with strong correlation are retained. Parameters with weak or no correlation are eliminated. The calculation equation of Pearson correlation coefficient is shown in (12).

$$r = \frac{N\sum XY - \sum X \sum Y}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}} \quad (12)$$

where, N denotes the variable sample size. X and Y are sensor variables. By using the Pearson correlation coefficient analysis, 34 parameters are selected. The proportions of the training and test sets are 80% and 20%, respectively. Each experiment is processed 10 times on average to reduce the effect of randomness.

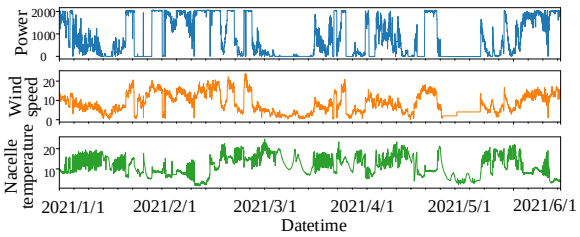

Fig. 4. Wind speed and power time-series distribution.

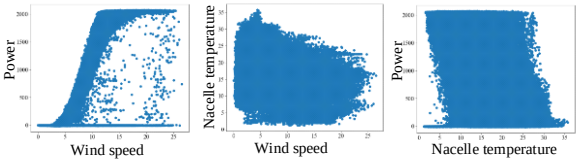

Fig. 5. Monitor the correlation between parameters.

To eliminate differences in magnitude between the data and reduce bias in the data, we normalize the data. Data normalization can be defined by Eq. (13).

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (13)$$

where, x' represents the normalized data. x is the original data. x_{max} and x_{min} represent the maximum and minimum values of the original data, respectively.

Experimental conditions

In this paper, all experiments are conducted on the desktop PC (4.6 GHz Intel i7-11800H processor, 32 GB RAM memory). The environment configuration of the experiment platform is shown in Table 1.

Table 1. Experiment platform and environment configuration

Experiment environment	Configuration
Operating system	Windows 11 system 64-bit
CPU	i7-11800H
GPU	NVIDIA GeForce RTX3080 10G
RAM	32GB
Development environment	Python 3.10

Configuration of SGG-DGCN

As shown in Table 2, the parameters of SGG-DGCN are set. The batch size is set to 42. The dropout is utilized to prevent the overfitting issue. The dropout rate is set to 0.1 and the learning rate is set to 0.01. The DGCN model consists of two convolutional layers with an input dimension of [36,130]. The number of attention heads for the two convolutional layers is 15 and 1, respectively. The cross-entropy loss function is used as the loss function. Adam is used as the optimization algorithm and trained for 300 epochs.

Table 2. Parameter settings of SGG-DGCN model

Description	Value
Convolution layer1	36*15
Convolution layer2	130*1
loss function	Cross-entropy
optimizer	Adam
learning rate	0.01
batch size	42
Activation function	ReLU
epoch	200
dropout rate	0.1

Configuration of benchmark models

To validate the effectiveness and superiority of SGG-DGCN, four commonly used benchmark models, including CNN, GCN, SGCN and DGCN are selected to compare with SGG-DGCN. The experimental details of these benchmarks are as follows:

(1) CNN. The 1D convolutional layer is 64*3. The maximum pooling layer is used with the pooling window size set as 2. The fully connected layer has 64 neurons. The sigmoid is used as the activation function for the binary classification task. The batch size for the training process is set as 42. The number of training iterations is set to 200. Binary cross entropy is used as the loss function for the model. Adam is used as the optimization algorithm and learning rate is set as 0.01.

(2) GCN. The GCN has two Conv layers. The network is trained over 200 epochs. NLL Loss is used as the loss function. Adam is used as the optimization algorithm and learning rate is set as 0.01.

(3) SGCN and DGCN. Sparse graph convolutional networks improve GCN performance through sparsity and low-rank graph structure properties. The parameter setting of SGCN and DGCN is the same with GCN.

(4) KNN-DGCN, COS-DGCN, Mah-DGCN and Man-DGCN. The parameter setting of KNN-DGCN, COS-DGCN, Mah-DGCN and Man-DGCN is the same with SGG-DGCN.

Evaluation index

Four evaluation indexes are adopted to evaluate the performance of the identification model in this paper. They are Accuracy, Precision, Recall and F1-score respectively.

The formula for the accuracy rate is shown in Eq. (14).

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \quad (14)$$

The formula for the precision is shown in Eq. (15).

$$\text{Precision} = \frac{TP}{TP+FP} \quad (15)$$

The formula for the recall is shown in Eq. (16).

$$\text{Recall} = \frac{TP}{TP+FN} \quad (16)$$

The formula for the F1-score is shown in Eq. (17).

$$\text{F1-score} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \quad (17)$$

where, TP denotes the number of samples that are actually positive and labeled as positive. FP denotes the number of samples that are actually negative but labeled as positive. FN denotes the number of samples that are actually positive but labeled as negative. TN denotes the number of samples that are actually negative and labeled as negative.

Results and discussion

Similar information

The similarity information derived using a single method and fusing the four methods is shown in Fig. 6. Each non-diagonal element represents the weight of the connection between two nodes. The change in color in the figure indicates the magnitude of the connection weights. The red color represents high weight. The blue color represents low weight. As can be concluded from Fig. 6, each distance method captures different aspects of the data information. By fusing these four methods, the advantages of each method are fused.

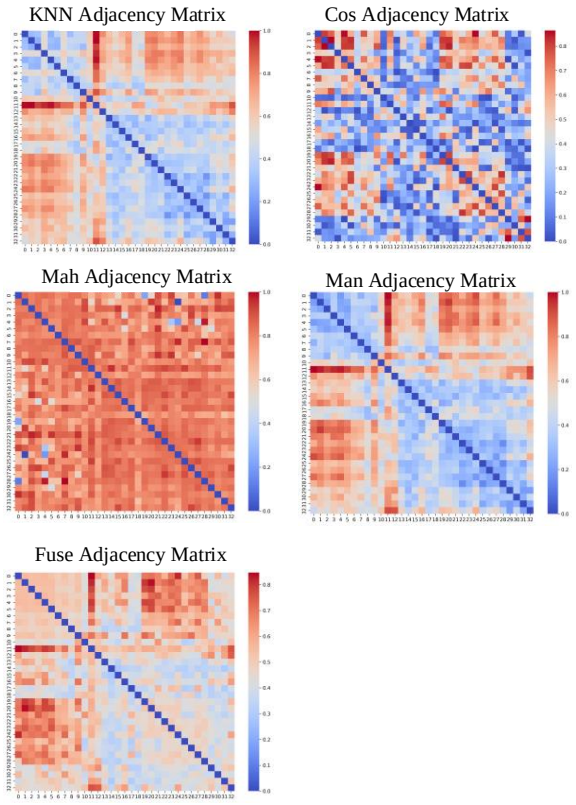

Fig. 6. Similarity information derived from the single method and fusing the four methods.

Comparative analysis of models

The testing data is input into the trained benchmark models and the trained SGG-

DGCN for comparison. Each trial is repeated ten times to reduce the randomness of the experimental results. The identification accuracies of the ten trials are shown in Fig. 7. The confusion matrix for each model is shown in Fig. 8. Besides that, the mean accuracy and standard deviation of utilized models are illustrated in Table 3. Other assessment indicators are shown in Table 4.

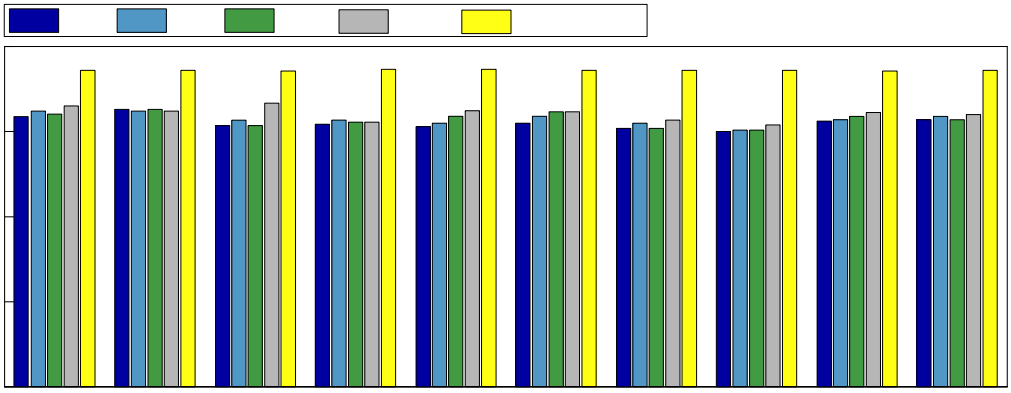

Fig. 7. Accuracies of different models in ten trials.

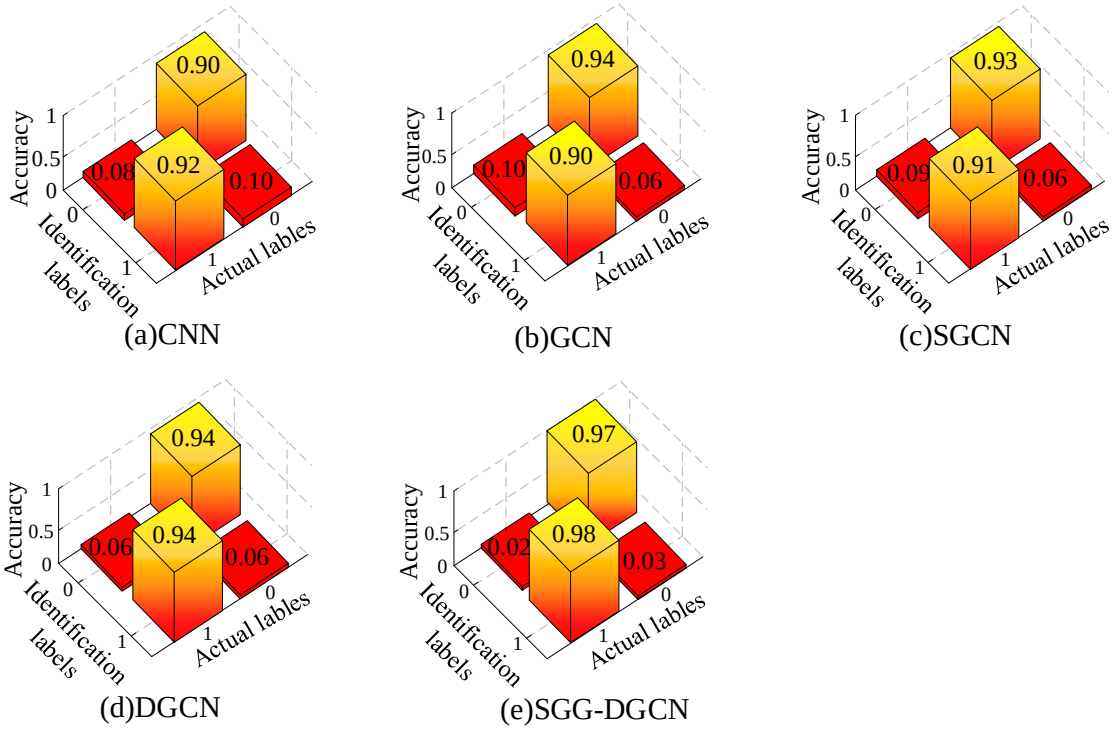

Fig. 8. Three-dimensional confusion matrices in the testing data.

Table 3. Accuracy and standard deviation of different models

Accuracy \ Model	CNN	GCN	SGCN	DGCN	SGG-DGCN
Mean	91.06%	92.10%	92.55%	94.10%	97.44%
Std	2.58	2.56	2.28	1.43	0.83

Table 4. Evaluation indexes of five models

Method	Precision	Recall	F1-score
CNN	90.39%	91.87%	91.12%
GCN	90.67%	93.88%	92.25%
SGCN	91.59%	93.71%	92.68%
DGCN	93.79%	94.44%	94.02%
SGG-DGCN	97.74%	97.12%	97.43%

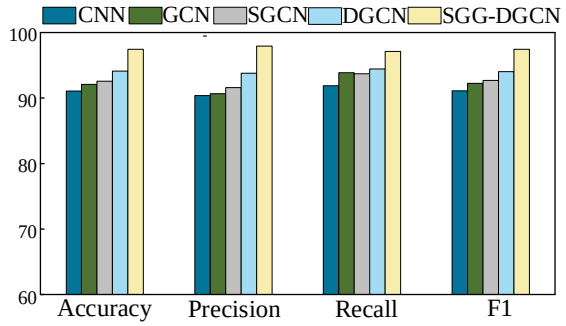

Fig. 9. Evaluation indexes for different models on the testing data.

We compare the proposed SGG-DGCN with four benchmark models such as CNN, GCN, SGCN, DGCN and SGG-DGCN to analyze the performance of the model. The results in Fig. 7 show that compared to the other four benchmark models, the SGG-DGCN model has the highest identification accuracy in ten trials. As shown in Fig. 8, the confusion matrix visualizes the recognition results, where the columns and rows represent the true and predicted states of the samples, respectively. The analysis shows that the SGG-DGCN method recognizes more than 97% of the anomalous categories.

As can be seen in Fig. 9, the SGG-DGCN has greatly improved the recall and F1-score compared to the other four models. The SGG-DGCN model has the best identification performance.

To more quantitatively evaluate the performance of the benchmark models, other indexes of the benchmark models are calculated. In terms of accuracy, the average identification accuracy of CNN, GCN, SGCN, DGCN and SGG-DGCN are 91.06%, 92.10%, 92.55%, 94.10% and 97.44%, respectively. Compared with CNN, GCN, SGCN and DGCN, the average identification accuracy of SGG-DGCN is improved by 6.38%, 5.34%, 4.89% and 3.34%, respectively. The standard deviation of CNN, GCN, SGCN, DGCN and SGG-DGCN are 2.58, 2.56, 2.28, 1.43 and 0.83, respectively. Compared with CNN, GCN, SGCN and DGCN, the SGG-DGCN has the smallest standard deviation. Above results demonstrate that the SGG-DGCN model has the best identification accuracy and robustness compared with the other four benchmark models.

In terms of precision, the precision of CNN, GCN, SGCN, DGCN and SGG-DGCN are 90.39%, 90.67%, 91.59%, 93.79% and 97.74%, respectively. Compared with CNN, GCN, SGCN and DGCN, the precision of

SGG-DGCN is improved by 7.35%, 7.07%, 6.15% and 3.95%, respectively. In terms of recall, the recall of CNN, GCN, SGCN, DGCN and SGG-DGCN are 91.87%, 93.88%, 93.71%, 94.44% and 97.12%, respectively. Compared with CNN, GCN, SGCN and DGCN, the recall of SGG-DGCN is improved by 5.25%, 3.24%, 3.41% and 2.68%, respectively. In terms of F1-score, the F1-score of CNN, GCN, SGCN, DGCN and SGG-DGCN are 91.12%, 92.25%, 92.68%, 94.02% and 97.43%, respectively. Compared with CNN, GCN, SGCN and DGCN, the F1-score of SGG-DGCN is improved by 6.31%, 5.18%, 4.75% and 3.41%.

From all the above evaluations, it is clear that the performance of the DGCN model using the distance formula to build the adjacency matrices is improved compared to the basic DGCN model. To summarize, the proposed SGG-DGCN model has the best precision, recall and F1-score. The proposed model can be well applied to the anomaly detection of wind turbines.

Analysis of ablation experiments

In order to further validate the effectiveness of the proposed SGG strategy, the four model are used to compare with the proposed SGG-DGCN model. Each trial is repeated ten times to reduce the randomness of the experimental results. The identification accuracies of the ten trials are shown in Fig. 10. The confusion matrix for each model is shown in Fig. 11. Besides that, the mean accuracy and standard deviation of utilized models are illustrated in Table 5. Other assessment indicators are shown in Table 6.

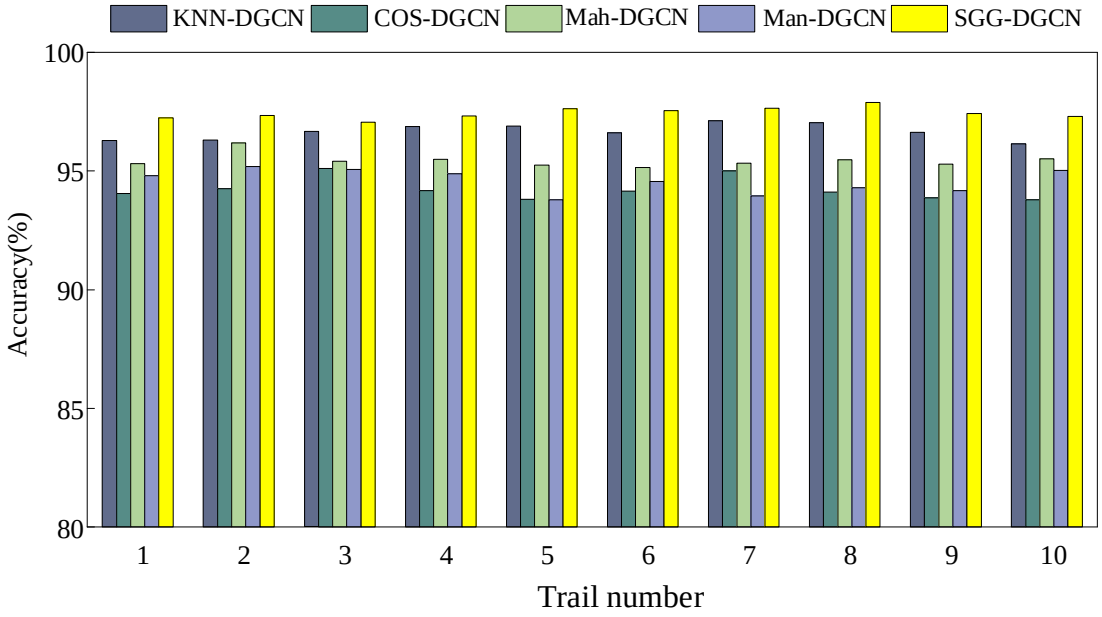

Fig. 10. Accuracies of different models in ten trials.

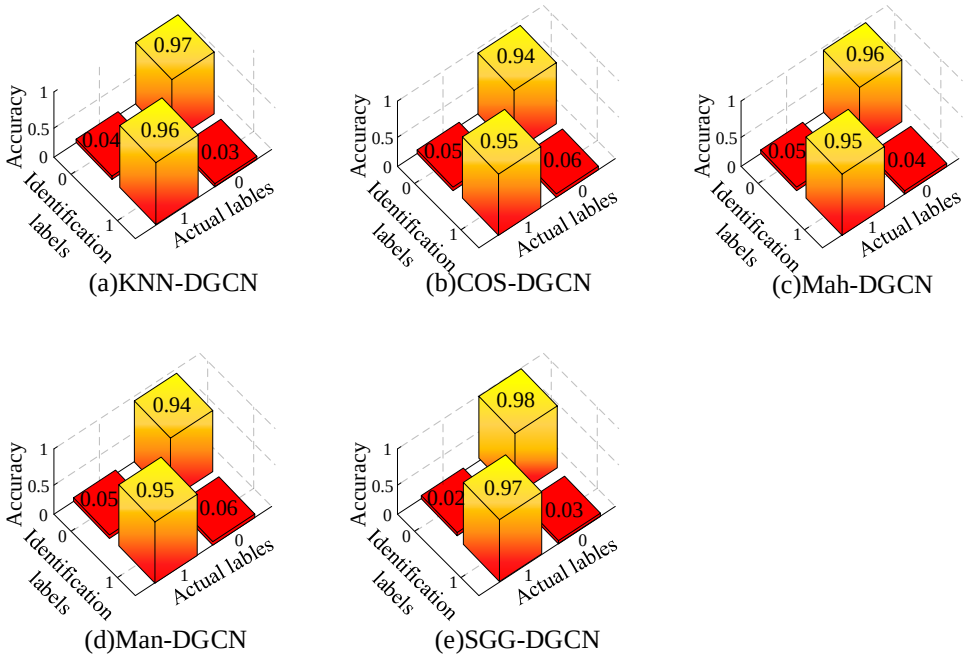

Fig. 11. Three-dimensional confusion matrices in the testing data.

Table 5. Accuracy and standard deviation of different models

Accuracy \ Model	KNN-DGCN	COS-DGCN	Mah-DGCN	Man-DGCN	SGG-DGCN
Mean	96.65%	94.23%	95.44%	94.57%	97.44%
Std	0.98	1.32	1.04	1.39	0.83

Table 6. Evaluation indexes of five models

Method	Precision	Recall	F1-score
KNN-DGCN	95.88%	95.89%	95.27%
COS-DGCN	93.88%	94.66%	94.87%
Mah-DGCN	95.64%	95.21%	95.42%
Man-DGCN	94.11%	95.10%	94.60%
SGG-DGCN	97.94%	97.12%	97.43%

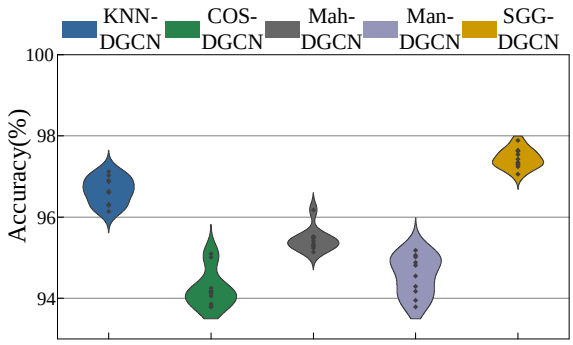

Fig. 12. Accuracy distribution of different models in ten trials.

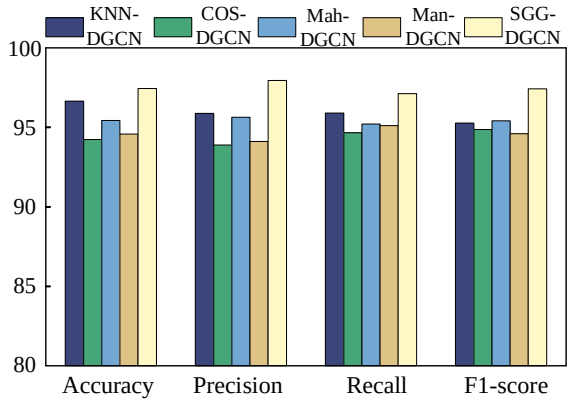

Fig. 13. Evaluation indexes for different models on the testing data.

The KNN-DGCN, COS-DGCN, Mah-DGCN and Man-DGCN are used to compare with the proposed SGG-DGCN model. The results in Fig. 10 show that compared to the five models, the SGG-DGCN model has the highest identification accuracy in ten trials. The confusion matrix for each model is shown in Fig. 11. The analysis shows that the SGG-DGCN method recognizes more than 97% of the anomaly categories and has the highest recognition accuracy in each category compared to other models. The Fig. 12 shows that the proposed model with SGG strategy has the smallest fluctuation range of accuracy value compared with the other four models. In Fig. 13, the results show that the evaluation indexes of the proposed model with SSG strategy are higher than all models.

As shown in Table 5, in terms of accuracy, the average identification accuracy of KNN-DGCN, COS-DGCN, Mah-DGCN, Man-DGCN and SGG-DGCN are 95.65%, 94.23%, 95.44%, 94.57% and 97.44%, respectively. Compared with KNN-DGCN, COS-DGCN, Mah-DGCN and Man-DGCN, the average identification accuracy of SGG-DGCN is improved by 1.79%, 3.21%, 2% and 1.87%, respectively. The standard deviation of KNN-DGCN, COS-DGCN, Mah-DGCN, Man-DGCN and SGG-DGCN are 0.98, 1.32, 1.04, 1.39 and 0.83, respectively. Compared with KNN-DGCN, COS-DGCN, Mah-DGCN and Man-DGCN, the standard deviation of SGG-DGCN is reduced by 0.15, 0.49, 0.21 and 0.56, respectively.

In terms of precision, the precision of KNN-DGCN, COS-DGCN, Mah-DGCN, Man-DGCN and SGG-DGCN are 95.88%, 93.88%, 95.64%, 94.11% and 97.44% respectively. Compared with KNN-DGCN, COS-DGCN, Mah-DGCN and Man-DGCN, the precision of SGG-DGCN is improved by 1.56%, 3.56%, 1.8% and 3.33%, respectively.

The recall of KNN-DGCN, COS-DGCN, Mah-DGCN, Man-DGCN and SGG-DGCN are 95.89%, 94.66%, 95.21%, 95.10% and 97.12%, respectively. Compared with KNN-DGCN, COS-DGCN, Mah-DGCN and Man-DGCN, the recall of SGG-DGCN is improved by 3.78%, 1.58%, 3.53%, 3.74% and 0.64%, respectively.

Apart from this, the F1-score of KNN-DGCN, COS-DGCN, Mah-DGCN, Man-DGCN and SGG-DGCN are 94.58%, 95.89%, 94.72%, 94.62%, 96.84% and 97.20%, respectively. Compared with KNN-DGCN, COS-DGCN, Mah-DGCN and Man-DGCN, the SGG-DGCN is improved by 1.23%, 2.46%, 1.91% and 2.02%, respectively.

The identification performance of the DGCN model is further improved by using the proposed SGG strategy with the highest values in all evaluation metrics compared to a single distance metric. For wind turbine anomaly identification, the DGCN model with the proposed SGG strategy has the best identification accuracy and stability.

Conclusion

In this paper, a similarity graph generation (SGG) strategy is proposed. Combining SGG strategy and DGCN for anomaly identification of wind turbines. By adopting the SGG strategy, the feature extraction ability of the proposed SGG-DGCN model is greatly improved to realize the high anomaly identification accuracy of wind turbines. A lot of experiments are conducted to validate the effectiveness of SGG-DGCN. The experimental results indicate that compared with other benchmark methods, the proposed SGG-DGCN method has the highest identification accuracy and stability. The accuracy of our proposed method can reach up to 97.2%. Compared with CNN, GCN, SGCN and DGCN, the average identification accuracy of SGG-DGCN is improved by 6.38%, 5.34%, 4.89% and 3.34%, respectively. The SGG-DGCN model has the highest anomaly identification accuracy. In addition, ablation experiments to validate the effectiveness of the proposed SGG strategy are also conducted. Ablation experiments demonstrate that by utilizing the proposed SGG strategy, the average identification accuracy of SGG-DGCN can be effectively improved by 1.79%, 3.21%, 2% and 1.87% compared with KNN-DGCN, COS-DGCN, Mah-DGCN and Man-DGCN, respectively. The proposed SGG strategy can effectively improve the accuracy and reliability of anomaly identification and realize the stable anomaly identification of wind turbines. It provides a guarantee for the safe and reliable operation of wind turbines. The study's constraints lie in its focus solely on sample similarity, neglecting a deeper analysis of individual data patterns. Additionally, the method's efficiency could be enhanced, warranting further investigation into optimizing its lightweight design.

Data availability

Data will be made available on request.

CRediT authorship contribution statement

Xiaomin Wang: Writing – original draft, Methodology, Validation, Investigation, Data curation. **Xiao Zhuang**: Data curation, Funding acquisition. **Jian Ge**: Data curation. **Jiawei Xiang**: Funding acquisition, Project administration, Writing – review & editing. **Di Zhou**: Conceptualization, Methodology, Formal analysis, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This work is supported by National Natural Science Foundation of China (Nos. U52305124, U62201399), the Zhejiang Natural Science Foundation of China (Nos. LQ23E050002), the Basic Scientific Research Project of Wenzhou City (Nos. G2022008, G2023028), the General Scientific Research Project of Educational Department of Zhejiang Province (Nos. Y202249008, Y202249041), China Postdoctoral Science Foundation (Nos.

2023M740988), Zhejiang Provincial Postdoctoral Science Foundation (Nos. ZJ2023122), the Master's Innovation Foundation of Wenzhou University (Nos. 3162024004106).

References

[1] Javed, M. S., Ma, T., Jurasz, J., & Amin, M. Y, "Solar and wind power generation systems with pumped hydro storage: Review and future perspectives," Renewable Energy **Vol,** 148, 176-192 (2020).

[2] Liu, Z., & Zhang, L, "A review of failure modes, condition monitoring and fault diagnosis methods for large-scale wind turbine bearings," Measurement **Vol,** 149, 107002 (2020).

[3] Yeter, B., Garbatov, Y., & Soares, C. G, "Risk-based maintenance planning of offshore wind turbine farms," Reliability Engineering & System Safety **Vol,** 202, 107062 (2020).

[4] Li, M., Yu, D., Chen, Z., Xiahou, K., Ji, T., & Wu, Q. H, " A data-driven residual-based method for fault diagnosis and isolation in wind turbines," IEEE Transactions on Sustainable Energy **Vol,** 10(2), 895-904 (2018).

[5] Kong, Z., Tang, B., Deng, L., Liu, W., & Han, Y, "Condition monitoring of wind turbines based on spatio-temporal fusion of SCADA data by convolutional neural networks and gated recurrent units," Reliability Engineering & System Safety **Vol,** 146, 760-768 (2020).

[6] Chen, H., Liu, H., Chu, X., Liu, Q., & Xue, D, "Anomaly detection and critical SCADA parameters identification for wind turbines based on LSTM-AE neural network," Renewable Energy **Vol,** 172, 829-840 (2021).

[7] Zhang, C., Hu, D., & Yang, T, "Anomaly detection and diagnosis for wind turbines using long short-term memory-based stacked denoising autoencoders and XGBoost," Reliability Engineering & System Safety **Vol,** 222, 108445 (2022).

[8] Wen, X., & Xu, Z, "Wind turbine fault diagnosis based on ReliefF-PCA and DNN," Expert Systems with Applications **Vol,** 178, 115016 (2021).

[9] Zhang, Z., & Wu, L, "Graph neural network-based bearing fault diagnosis using Granger causality test," Expert Systems with Applications **Vol,** 242, 122827 (2024).

[10] Gao, Yiyuan, Mang Chen, and Dejie Yu, "Semi-supervised graph convolutional network and its application in intelligent fault diagnosis of rotating machinery," Measurement **Vol,** 186, 110084 (2021).

[11] Zhu, R., Ji, X., Yu, D., Tan, Z., Zhao, L., Li, J., & Xia, X, "KNN-based approximate outlier detection algorithm over IoT streaming data," IEEE Access **Vol,** 8, 42749-42759 (2020).

[12] Zheng, L., Jia, K., Bi, T., Fang, Y., & Yang, Z, "Cosine similarity based line protection for large-scale wind farms," IEEE Transactions on Industrial Electronics **Vol,** 68(7), 5990-5999 (2020).

[13] Si, Y., Chen, Z., Sun, J., Zhang, D., & Qian, P, "A data-driven fault detection framework using mahalanobis distance based dynamic time warping," IEEE Access **Vol,** 8, 108359-108370 (2020).

[14] https://zenodo.org/record/5946808#.Yg-pAmvso-V5.