Fault Diagnosis of Power Transformers Using Graph Convolutional Network

Liao, Wenlong; Yang, Dechang ; Wang, Yusen ; Ren, Xiang

Published in:
CSEE Journal of Power and Energy Systems

DOI (link to publication from Publisher):
10.17775/CSEEJPES.2020.04120

Publication date:
2021

Document Version
Accepted author manuscript, peer reviewed version

Fault Diagnosis of Power Transformers Using Graph Convolutional Network

Wenlong Liao, Dechang Yang, *Member*, *IEEE*, Yusen Wang, and Xiang Ren

***Abstract*—Existing methods for transformer fault diagnosis either train a classifier to fit the relationship between dissolved gas and fault type or find some similar cases with unknown samples by calculating the similarity metrics. Their accuracy is limited, since they are hard to learn from other algorithms to improve their own performance. To improve the accuracy of transformer fault diagnosis, a novel method for transformer fault diagnosis based on graph convolutional network (GCN) is proposed. The proposed method has the advantages of two kinds of existing methods. Specifically, the adjacency matrix of GCN is utilized to fully represent the similarity metrics between unknown samples and labeled samples. Furthermore, the graph convolutional layers with strong feature extraction ability are used as a classifier to find the complex nonlinear relationship between dissolved gas and fault type. The back propagation algorithm is used to complete the training process of GCN. The simulation results show that the performance of GCN is better than that of the existing methods such as convolutional neural network, multi-layer perceptron, support vector machine, extreme gradient boosting tree, k-nearest neighbors and Siamese network in different input features and data volumes, which can effectively meet the needs of diagnostic accuracy.**

***Index Terms*—Power transformer, fault diagnosis, graph convolutional network, similarity metrics.**

I. Introduction

As one of the most important equipment in substation and power plant, the power transformer has the ability of adjusting multiple voltage levels, and plays an important role both in the transmission side and distribution side. The operational state of the transformer is directly related to the safety and power quality of the whole power system [1]. Once the transformer fails, it may lead to local power failure or even large-scale power failure, which will cause huge economic losses to society. Therefore, it is of great significance for the power system to diagnose the status of the power transformer accurately.

At present, most of the large-scale transformers belong to oil-immersed transformers. When faults occur, oil-immersed transformers will release a large amount of dissolved gas, which are important indexes for fault diagnosis by dissolved gas analysis (DGA) [2]. The existing methods for transformer fault diagnosis based on DGA can be summarized into two categories: distance-based methods and model-based methods. For the first category, it mainly includes case-based reasoning, expert system, k-nearest neighbor (KNN), and Siamese network [3-7], which attempt to calculate the similarity metrics between the samples to be classified or verified and the historical samples. In general, although these distance-based methods make full use of historical data and prior knowledge through similarity metrics, they are difficult to capture the complex nonlinear relationship between dissolved gas and corresponding labels, resulting in their limited accuracy for transformer fault diagnosis. For the second category, traditional model-based algorithms include support vector machine (SVM) [8], multi-layer perceptron (MLP) [9], extreme gradient boosting tree (XGBoost) [10], and light gradient boosting machine (Light GBM) [11]. Generally, although these traditional methods are more suitable for smaller data set, their limited feature extraction ability is difficult to fully explore the potential nature between dissolved gas and corresponding labels.

Deep learning is part of machine learning research. The emergence of deep learning promotes the development of machine learning and the innovation of artificial intelligence [12]. Due to its powerful learning ability, deep learning has been applied to various fields of power system, such as fault diagnosis [13], optimized scheduling [14], and scenarios generation [15]. The new model-based methods of deep learning include convolutional neural networks (CNN), deep belief network, and capsule neural networks [16]. Although these new model-based methods fully exploit the complex nonlinear relationship between dissolved gas and fault types, they don't make good use of the similarity metrics between historical data and current samples, which makes their performance hard to be further improved.

As a branch of deep learning, the graph neural networks have achieved outstanding performance in graph-structured data, such as recommendation system, link prediction, node classification, and protein structure inference [17], [18], which brings a new opportunity for the development of transformer fault diagnosis. Especially, the graph convolutional network

Manuscript received August 18, 2020; revised October 28, 2020; accepted November 18, 2020. Date of online publication August 18, 2020; date of current version August 18, 2020.

Wenlong Liao is with the Department of Energy Technology, Aalborg University, Aalborg 9220, Denmark.

Dechang Yang (corresponding author, e-mail: Yangdechang@cau.edu.cn) is with the college of information and electrical engineering, China Agricultural University, Beijing 100083, China.

Yusen Wang is with the School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, SE-100 44, Sweden.

Xiang Ren is with the North China Electric Power Research Institute Co., Ltd., Beijing 100045, China.

DOI: 10.17775/CSEEJPES.2020.04120

(GCN) is a kind of feed-forward neural network which uses graph convolution to process graph-structured data such as social networks, traffic network, knowledge graphs, and molecules of materials. Traditional CNN can only deal with data in Euclidean domains, which is a special case of graph-structured data [19]. For example, the features of each node included in social networks are data in Euclidean domains. In addition, the social network also has an adjacency matrix to represent the connection relationship of each node. If every node is independent, the adjacency matrix is empty, and the social network (graph-structured data) degenerates into the data in Euclidean domains. At present, GCN has shown excellent performance in many fields such as power load prediction, drug synthesis, few-shot learning, and link prediction [20], [21]. However, its application in the transformer fault diagnosis is still in its infancy. In theory, GCN can not only use graph convolutional layers with strong learning ability to effectively mine the complex nonlinear relationship between fault type and dissolved gas, but also use an adjacency matrix to represent the similarity metrics between unknown samples and labeled samples, so as to improve the accuracy for transformer fault diagnosis. Specifically, how to design a structure of GCN with strong feature extraction ability and high diagnostic accuracy according to the characteristics of dissolved gas data of the transformer needs further research.

In this paper, it is aimed to design a GCN to improve the accuracy for transformer fault diagnosis. The performance of the proposed method is tested by the actual DGA data set. The key contributions of this paper are as follows:

1) Theoretical innovation: the proposed GCN has the double advantages of the distance-based methods and model-based methods. Specifically, it can not only mine the complex nonlinear relationship between fault types and dissolved gas by using graph convolutional layer, but also represent the similarity metrics between unknown samples and labeled samples through an adjacency matrix.

2) Application innovation: to our best knowledge, this paper is the first to design the GCN for transformer fault diagnosis. The influence of key parameters of GCN (e.g. the size of *k*, optimizer, and the number of graph convolutional layers) on the performance for transformer fault diagnosis is analyzed by simulation on real data set, and the constructive suggestions for the selection of each parameter are given.

3) Extensive experiments on real data set collected from the state grid corporation of China and previous publications are performed to validate the effectiveness of the deep learning framework for transformer fault diagnosis. The simulation results show that the GCN achieves state-of-art performance with superior accuracy in different input features and data volumes for transformer fault diagnosis.

The rest of the paper is organized as follows. Section II explains the principle and structure of the GCN. Section III introduces the steps of transformer fault diagnosis based on the GCN. Section IV discusses the simulations and results. Section V discusses the innovation and generality of the GCN. Section VI summarizes the conclusions.

II. Graph Convolutional Network

A. Definitions of the graph convolutional layer

At present, most graph convolutional layers have a common architecture, since filter parameters are usually shared at all locations in the graph [22]. For these graph convolutional layers, the goal is to learn a function of features on the graph-structured data $G=(V,E)$ which takes a feature matrix X of dissolved gas content and an adjacency matrix A of the samples as input:

$$\text{Input}=(X,A) \tag{1}$$

where X is a $n\times d$ feature matrix that consists of the feature description x_i for each node i. n the is number of nodes (n is the number of samples in the transformer fault diagnosis) and d is the number of input features. The adjacency matrix represents the similarity metrics between historical data and current samples in matrix form.

The output of the graph convolutional layer is a $N\times F$ node-level vector Y, where F is the number of transformer states. Every graph convolutional layer can be written as a non-linear function:

$$H^{(i+1)}=f\left(H^{(i)},A\right),i=0,1,\cdots L \tag{2}$$

where L is the number of graph convolutional layers. When i equals 0, $H^{(0)}$ is identical to X. When i equals L, $H^{(L)}$ is identical to Y. The specific graph convolutional layers differ only in how activation function f is chosen and parameterized.

Now, an example is used to show the following simple form of a layer-wise propagation principle for graph convolutional layers:

$$f\left(H^{(i)},A\right)=\sigma\left(AH^{(i)}W^{(i)}\right) \tag{3}$$

where σ is a non-linear activation function such as rectified linear unit (ReLU) function and $W^{(i)}$ is a weight matrix in the i-th graph convolutional layer.

Although the graph convolutional layer is very powerful, it has two limitations that need to be addressed [17]:

1) Multiplying the adjacency matrix A means that for each node, it totes up the eigenvectors of all adjacent nodes, not the node itself (unless there is a self-loop in the graph-structured data). This limitation can be fixed by enforcing self-loops in the graph-structured data (e.g. the identity matrix is added to the adjacency matrix A):

$$A'=A+I \tag{4}$$

2) The second limitation is that the adjacency matrix A' is not normalized, so the multiplication may change the scale of the eigenvector, which can be examined by checking the eigenvalues of the adjacency matrix A'. To deal with this problem, the adjacency matrix A' should be normalized by the following formula:

$$A''=D^{-\frac{1}{2}}A'D^{-\frac{1}{2}} \tag{5}$$

where D is the diagonal node degree matrix of the adjacency matrix A':

$$D_{ii}=\sum_j A'_{ij} \tag{6}$$

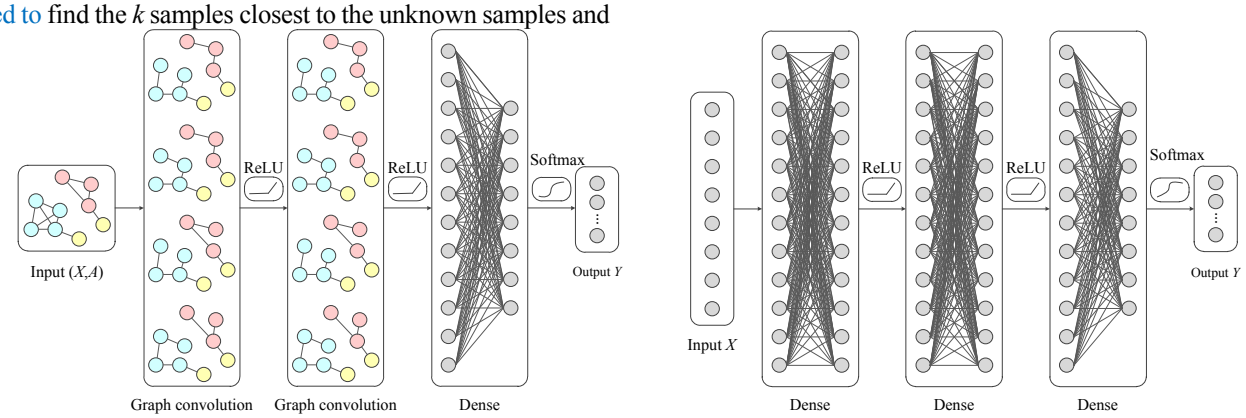
After using these two tricks, the new propagation principle for graph convolutional layers becomes:

$$f\left(H^{(i)},A\right)=\sigma\left(A"H^{(i)}W^{(i)}\right) \tag{7}$$

B. Structure of the GCN

Normally, the neural network is designed to diagnose fault type Y for power transformer by inputting dissolved gas content X. Specifically, the GCN requires an $n\times n$ adjacency matrix A in addition to X. Let n denotes the number of samples in the dataset. The samples in training set and validation set are only linked to samples with the same label. For example, if the *i*-th sample and the *j*-th sample belong to partial discharge, then $A(i,j)=A(j,i)=1$. For the samples in the test set (unknown samples), the Siamese network is used to extract the low dimensional features of input variables, so as to calculate the Euclidean distance between samples [7]. Then, the KNN is utilized to find the *k* samples closest to the unknown samples and consider them to be connected. For example, if *k* is equal to 1, the Euclidean distance between the *i*-th unknown sample and the *j*-th sample is the closest, then $A(i,j)=A(j,i)=1$. In this case, the adjacency matrix A can represent the similarity metrics between historical data and current samples.

To illustrate the structure of GCN, Fig.1 (a) shows a simple GCN model that includes two graph convolutional layers followed by two dense layers. Obviously, this architecture is different from that of traditional neural networks (e.g. MLP) in Fig. 1(b). The input data (X,A) of the GCN is a graph form and the intrinsic topology is utilized to generate intermediate features, while MLP only uses the dissolved gas content X as input data and ignores the similarity metrics of samples.



(a) Graph convolutional network (b) Multi-layer perceptron

Fig. 1. Structures of GCN and MLP.

As shown in Fig. 1 (a), the graph-structured data (X,A) is fed to the GCN that outputs $H^{(1)}$ on the first layer. Specifically, a hybrid feature matrix $A"\hat{X}$ is obtained, which combines a feature vector of each node with those of adjacent nodes linearly by using weights represented by $A"$. Next, a new set of features $A"\hat{X}W_1+b_1$ are obtained by multiplying a weight matrix W_1 and adding a bias vector b_1. After that, an activation function (e.g. ReLU) is selected for the new vector to obtain the output data of first layer $H^{(1)}$:

$$H^{(1)}=\mathrm{ReLU}\left(A"\hat{X}W_1+b_1\right) \tag{8}$$

Similarly, the output data of the second graph convolutional layer is:

$$H^{(2)}=\mathrm{ReLU}\left(A"H^{(1)}W_2+b_2\right) \tag{9}$$

where W_2 and b_2 is a weight matrix and a bias vector in the second graph convolutional layer, respectively.

There are two dense layers behind the graph convolutional layers. It is necessary to victories the data before inputting $H^{(2)}$. In the third layer, the output data $H^{(3)}$ can be obtained by a weight matrix W_3, a bias vector b_3, and an activation function:

$$H^{(3)}=\mathrm{ReLU}\left(H^{(2)}W_3+b_3\right) \tag{10}$$

The output of the fourth layer by a Softmax function is:

$$Y=\mathrm{Softmax}\left(H^{(3)}W_4+b_4\right) \tag{11}$$

where Y is the fault type of the transformer.

III. Transformer Fault Diagnosis Using GCN

A. Input Variables of the Model

In regular operation, the solid organic insulating materials and insulating oil of the power transformer will age gradually due to the combined action of electric field and thermal field. A small amount of dissolved gas (e.g. hydrogen and low molecular hydrocarbon gas) will be dissolved in transformer oil [23]. If the transformer has a discharge fault or thermal fault, the content of the dissolved gas will increase rapidly. If the rate of generating dissolved gas is greater than that of transformer oil absorbing gas, the excess gas will continue to diffuse and enter the relay, thus triggering an alarm. At present, one commonly used detection technology that can diagnose the fault type of oil-immersed transformer is through analyzing the content of dissolved gas [3]. Furthermore, the IEC ratio codes method, Dornenburg ratio codes method, and Rogers ratio codes method are used to construct a variety of new features, which effectively improve the accuracy of fault diagnosis [24]. Previous works have shown that CO and CO_2 have weak correlation with transformer fault type, while H_2, C_2H_6, CH_4, C_2H_2, and C_2H_4 have strong correlation with transformer fault type [7], [8]. Therefore, the content of dissolved gas (H_2, C_2H_6, CH_4, C_2H_2, and C_2H_4) are selected as the original features, and four new features (CH_4/H_2, C_2H_2/C_2H_4, C_2H_4/C_2H_6, and

C_2H_6/CH_4) constructed by Rogers ratio codes method are further considered as the input variables of the GCN.

Due to the large difference in the numerical values of the nine features, the performance of the model will be adversely affected if they are directly used as input variables, and even the loss function is difficult to converge. Therefore, before being fed into GCN, the nine features should be mapped to the interval [0,1] by the min-max normalization method:

$$x_i^{'} = \frac{x_i - x_{i,\min}}{x_{i,\max} - x_{i,\min}}, i = 1, 2, \cdots 9 \tag{12}$$

where x_i and $x_i^{'}$ represent the i-th feature before and after normalization, respectively. $x_{i,\min}$ and $x_{i,\max}$ represent the minimum and maximum values of the i-th feature, respectively.

B. Output Variables of the Model

Transformer faults can be divided into thermal fault and discharge fault. Specifically, the thermal fault includes the thermal fault of low temperature (LT), thermal fault of medium temperature (MT), and thermal fault of high temperature (HT). The discharge fault includes partial discharge (PD), discharge of low energy (LD), and discharge of high energy (HD) [2]. In order to effectively calculate the cross-entropy loss function in the training process of GCN, the one hot code method is used to preprocess various state types of the transformer, as shown in Table I.

TABLE I
THE ONE HOT CODE FOR TRANSFORMER STATE

Status of transformer	Codes
Normal	1000000
Thermal fault of low temperature	0100000
Thermal fault of medium temperature	0010000
Thermal fault of high temperature	0001000
Partial discharge	0000100
Discharge of low energy	0000010
Discharge of high energy	0000001

C. Process of Fault Diagnosis

The process of transformer fault diagnosis based on GCN is shown in Fig. 2, and the specific steps are as follows:

1) Data import and normalization

The contents of dissolved gas C_2H_2, CH_4, C_2H_4, C_2H_6, and H_2 are regarded as original features and four new features are constructed by the Rogers ratio codes method. The above nine features are used as input variables of GCN. To obtain adjacency matrix A, the Siamese network is used to extract the low dimensional features of input variables, so as to calculate the Euclidean distance between samples. Then, the KNN are utilized to find the k samples closest to the unknown samples and consider them to be connected. Furthermore, the min-max normalization method is used to transfer the input data into values that range from 0 to 1 [25].

2) Reconstitution and division of data

The adjacency matrix is reshaped into a sparse matrix in coordinate format, since it includes a lot of 0 elements, which leads to a waste of space. In the data set, 75% of the samples are used to train the GCN, and the remaining samples are used to evaluate the performance of the model.

3) Initializing the structure and parameters of GCN

In order to improve the accuracy of the transformer fault diagnosis, it is necessary to explore the best structure and parameters before training GCN. The structure and parameters of GCN mainly include the number of graph convolutional layers, the number of iterations, the size of k for adjacency matrix A, and the selection of optimizer. A simple structure of GCN is shown in Table II. The size of convolutional filters is 8 and 16 respectively. All activation functions of graph convolutional layers are ReLU. To alleviate over-fitting, the graph convolutional layer is followed by a dropout layer with a probability of 0.25. Both the size of convolutional filters and the probability of dropout layer are the optimal values found by many experiments in Case study (detailed discussion in Section B). Finally, the dense layer outputs a 1×7 vector to represent the state of the unknown sample.

4) Training GCN

The back-propagation algorithm is used to complete the training process of the GCN, which mainly includes two steps: forward incentive propagation and backward weight update. For the forward incentive propagation, the input variables are processed by multiple graph convolutional layers and then transferred to the dense layer, and the labels of samples are output by the dense layer. The diagnostic results and real results are used to calculate the loss function (error). For backward weight update, the error is transferred from the output layer to the middle layer by using the chain rule. Then, the weights of each layer are updated by the gradient descent method. When the set number of iterations is reached, the test set is used to evaluate the performance of GCN. In addition, the ensemble technology is used to improve the accuracy of the proposed models in the training process [26].

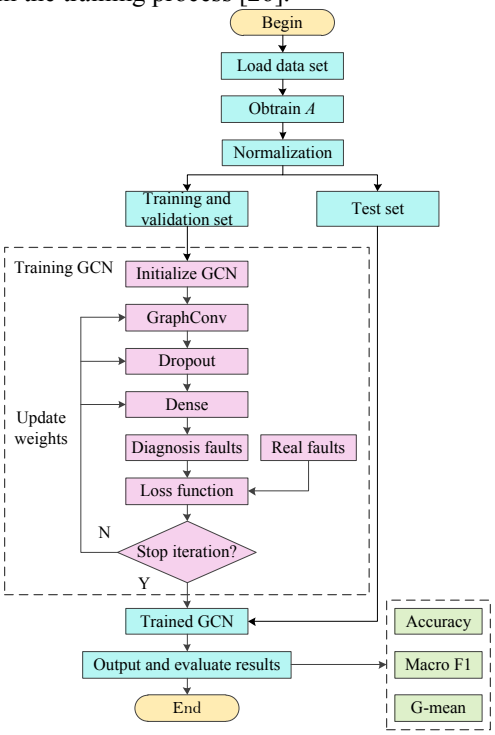


Fig. 2. Process of fault diagnosis based on GCN.

TABLE II
A SIMPLE STRUCTURE OF GCN

Layer	Structure and parameters	Shape
1-st	H0=Input(shape=(X.shape[1],))	1×9
2-rd	H1=GraphConv(filter=16,ReLU)([H0]+Graph)	1×16
3-th	H2=Dropout(rate=0.25)(H1)	1×16
4-th	H3=GraphConv(filter=8,ReLU)([H2]+Graph)	1×8
5-th	H4=Dropout(rate=0.25)(H3)	1×8
6-th	Y=Dense(unit=7, Softmax)(H4)	1×7

5) Evaluating the performance of GCN

For the binary classification problem, the results are either positive class or negative class. The accuracy and recall rate can be used to evaluate the performance of the model, but these indexes are not suitable for multi-classification problems such as transformer fault diagnosis. Normally, the k classification problem can be transformed into k binary classification problems. Besides the accuracy, the Macro F1 and geometric mean of recall (G-mean) are used to evaluate the performance of the model [27]. Macro F1 and G-mean are both positive indicators. In other words, the bigger the indicators, the better the performance.

IV. CASE STUDY

A. Data Description and Simulation Tools

In order to test the performance of GCN for transformer fault diagnosis, the actual data set collected from the state grid corporation of China and previous publications is used for simulation and analysis [28], [29]. The voltage level of these samples is 220 kV. After data cleaning, there are 718 samples left in the data set, which includes 7 state types: normal, thermal fault of low temperature, thermal fault of medium temperature, thermal fault of high temperature, partial discharge, discharge of low energy, and discharge of high energy. The number of training samples accounts for 75%. The remaining data is used as test samples to evaluate the performance of the models. The size of each state type is shown in Table III.

TABLE III
A SAMPLE DISTRIBUTION OF DATA SET

Status	All samples	Training samples	Testing samples
Normal	52	39	13
LT	99	74	25
MT	73	55	18
HT	168	126	42
PD	105	79	26
LD	42	31	11
HD	179	134	45

The proposed methods are all run on the Spyder platform from Anaconda software. The framework of deep learning is Tensorflow 1.12.0 and Keras 2.2.4. Computer hardware configuration: Intel Core i3-3110M CPU @ 2.4GHz dual-core CPU, 6GB memory.

B. Performance evaluation of GCN

In order to clearly observe the training process of GCN, Fig. 3 shows the changing trend of the loss function with the increase of iterations.

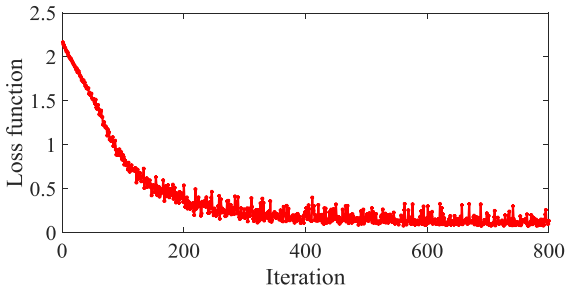


Fig. 3. The training process of the GCN.

In the early stage of the training process, the loss function of the training set decreases rapidly with the increase of iteration. When the number of iterations is greater than 400, the loss function tends to a constant and does not continue to decline, which indicates that GCN has converged. Generally speaking, the training process of GCN is relatively stable and has fast convergence speed. To ensure the convergence of GCN, the GCN is used to diagnose unknown samples after 800 iterations.

In order to analyze the influence of the number of graph convolutional layers on the performance of GCN, the number of graph convolutional layers is gradually increased, and the indexes of the test set under different layers are counted, as shown in Table IV.

TABLE IV
THE INDEXES OF TEST SET UNDER DIFFERENT LAYERS

layers	Accuracy	Macro F1	G-mean	Time/s	Number of parameters
1	0.714	0.689	0.709	35.27431	80
2	0.793	0.772	0.791	65.98592	152
3	0.780	0.753	0.769	96.35367	224
4	0.699	0.672	0.688	126.7464	296
5	0.683	0.655	0.673	157.6507	368

The following conclusions can be drawn from Table IV: 1) In the early stage, the indexes of the test set increase as the number of graph convolutional layers increases, which indicates that the performance of GCN is gradually becoming stronger. When the number of graph convolutional layers is 2, the accuracy, Macro F1, and G-mean of GCN are the largest, and the performance of fault diagnosis is the best. This phenomenon shows that it is difficult to mine the complex nonlinear relationship between dissolved gas and transformer state by a small amount of graph convolutional layers. Increasing the number of graph convolutional layers can improve the feature learning ability of GCN, thus enhancing the accuracy of fault diagnosis. 2) Furthermore, it is found that the number of parameters to be trained increases linearly with the increase of graph convolutional layers. When the number of graph convolutional layers is greater than 2, the performance of GCN will be worse and worse, if more convolution layers are added to GCN. This is because the number of samples in the dataset is limited. Too many graph convolutional layers will not only increase the parameters of GCN to be trained, but also consume a lot of training time, and it is easy to over-fitting and reduces the accuracy of diagnosis. 3) In general, the number of graph convolutional layers should be determined according to the size of the data set. If the number of samples is small, the GCN can get better performance by setting the graph convolutional layer to 2 layers.

The size of k determines the number of samples from training set connected with each unknown sample, which will directly

affect the adjacency matrix A. In order to explore the influence of the size of k on the performance of GCN, k is set from 1 to 20, and the performance of GCN on the test set is counted, as shown in Fig. 4

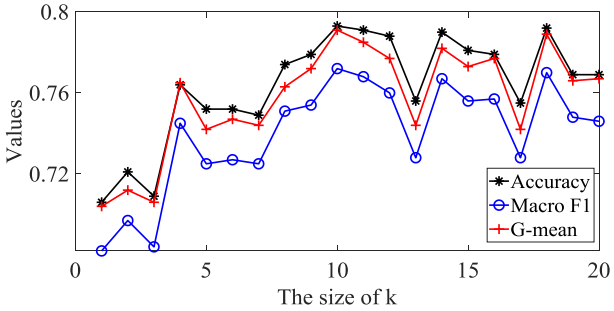


Fig. 4. The indexes of the test set in different sizes of k.

When the size of k is very small, the unknown samples are only connected with the nearest sample, which makes it difficult to make full use of the similarity metrics between samples, resulting in limited accuracy. By contrast, if the size of k is very large, the unknown samples are connected with many samples, which may cause the unknown samples to connect with different types of samples, resulting in noise and limited the performance of GCN. Therefore, with the increase of k value, the accuracy, Macro F1, and G-mean first increased and then decreased. When the k value is 10, the accuracy, Macro F1, and G-mean of the test set are the maximum, and the performance of GCN in fault diagnosis is the optimized.

After initializing the structure and parameters of GCN, a gradient descent method is needed to optimize the neural network. The popular methods for gradient descent include SGD, RMSprop, Adagrad, Adadelta, Adam, Adamax, and Nadam. At present, the mainstream deep learning libraries (e.g. Keras, Tensorflow, and Pytorch) include implementations of these methods to reduce loss function. Normally, these methods are often used as black boxes, because their principles are too complex to be explained in practical engineering. In order to find the optimizer suitable for GCN in transformer fault detection, these popular optimizers are set up and simulated, and the indexes of test set are counted, as shown in Table V.

TABLE V
THE INDEXES OF TEST SET UNDER DIFFERENT OPTIMIZER

Optimizers	Accuracy	Macro F1	G-mean
SGD	0.652	0.625	0.644
RMSprop	0.782	0.760	0.779
Adagrad	0.653	0.628	0.651
Adadelta	0.613	0.586	0.606
Adam	0.793	0.772	0.791
Adamax	0.781	0.761	0.780
Nadam	0.774	0.750	0.768

Table V shows that GCN has good performance when RMSprop, Adagrad, Adamax, Nadam, and Adam algorithms are used as optimizers. Specifically, the accuracy, Macro F1, and G-mean of Adam algorithm are slightly higher than those of the first four optimizers, which indicate that Adam algorithm is the most suitable optimizer for GCN in transformer fault detection. Furthermore, the corresponding accuracies of Adagrad, Adadelta, and SGD algorithms are all lower than 0.7, which shows that they are not suitable for transformer fault diagnosis based on GCN.

C. *Comparison under different input features*

In order to illustrate the effectiveness of GCN, the common distance-based methods (e.g. KNN and Siamese network) and model-based methods (e.g. CNN, MLP, XGBoost, and SVM) are used as the baselines. The indexes of the test set are compared under different input characteristics. After many experiments, the optimal parameters and structures of various algorithms are found as follows:

1) For the KNN, the size of k is 7. 2) For the Siamese network, it includes two CNNs with the same weights, which are utilized to calculate the Euclidean distance of input data (the input data of the Siamese network is a pair of samples). Specifically, the filters of the two convolutional layers are 16 and 36, and the size of the convolutional kernels is 2×2. The size of the maximum pooling layer is 2×2. A dropout layer is inserted behind the two convolutional layers to alleviate over-fitting. The probability of the dropout layer is set to 0.25. The numbers of neurons in the dense layers are 8 and 1. The activation function of all layers is the ReLU function. 3) For CNN, it includes two convolutional layers, two max-pooling layers, two dropout layers, and two dense layers. The probabilities of dropout layers are 0.25. The size of the kernel in the convolutional layers is 3. The size of the pool in max-pooling layers is 2×2. The activation functions of convolutional layers are the ReLU function. The numbers of neurons in the dense layers are 14 and 7, respectively. 4) For MLP, the number of neurons in the input layer is 9, and the number of neurons in the middle layer is 9 and 7, respectively. The number of neurons in the output layer is equal to the number of categories. To alleviate over-fitting, a dropout layer is inserted between each dense layer. 5) For XGBoost, The max depth is 5, and the gamma value is 0.2. The subsample rate is 0.6 and the min child weight is 3. 6) For SVM, the fitcecoc function from MATLAB2018a is used to classify fault type of transformer.

The above algorithms are trained under different input features, and the simulation results of the test set are shown in Table VI and Table VII.

TABLE VI
RESULTS OF GCN TRAINED BY FIVE ORIGINAL FEATURES

Method	Accuracy	Macro F1	G-mean
GCN	0.781	0.758	0.776
CNN	0.74	0.717	0.734
MLP	0.633	0.602	0.62
XGBoost	0.634	0.606	0.626
SVM	0.642	0.616	0.635
KNN	0.669	0.644	0.664
Siamese Network	0.751	0.725	0.741

TABLE VII
RESULTS OF GCN TRAINED BY NINE ORIGINAL FEATURES

Method	Accuracy	Macro F1	G-mean
GCN	0.793	0.772	0.791
CNN	0.756	0.732	0.748
MLP	0.672	0.649	0.669
XGBoost	0.641	0.613	0.631
SVM	0.688	0.662	0.683
KNN	0.700	0.677	0.695
Siamese Network	0.771	0.745	0.763

The following conclusions can be drawn from the Tables: 1) Accuracy represents the probability that the model can correctly identify the positive and negative classes. If the data

set has the unbalanced problem, the accuracy will be greatly affected by unbalanced classes. Therefore, Macro F1 and G-mean are selected to evaluate the performance of the model. It can be seen from Tables that the values of the accuracy, Macro F1 and G-mean are similar, which indicates that the probability of correct diagnosis of various fault types by the model is very close. In addition, the values of G-mean are all greater than 0, indicating that there is no missing detection for each fault type in the model. 2) The accuracy, Macro F1, and G-mean of each algorithm in Table VII are larger than those in Table VI, which shows that the four features constructed by Rogers ratio method can improve the performance of each algorithm in transformer fault diagnosis to a certain extent. 3) GCN has a higher diagnostic performance than the other algorithms in the case of two different input features. Table VII is illustrated as an example. The accuracy, Macro F1 and G-mean of GCN are 79.3%, 77.2%, and 79.1%, respectively. Compared with CNN, MLP, XGBoost, SVM, KNN, and Siamese networks, the accuracy of GCN increased by 3.7%, 12.1%, 15.2%, 10.5%, 9.3%, and 2.2%, respectively; Macro F1 of GCN increased by 4.0%, 12.3%, 15.9%, 11.0%, 9.5%, and 2.7%, respectively; G-mean of GCN increased by 4.3%, 12.2%, 16.0%, 10.8%, 9.6%, and 2.8%, respectively. 4) Last but not least, both CNN and GCN use convolutional layers to extract the features of input data. The difference is that the latter also takes into account the similarity metrics between unknown samples and labeled samples through an adjacency matrix. The performance of GCN is better than that of CNN, which shows that CNN ignores the similarity metrics between samples, which limits the accuracy of fault diagnosis. GCN can not only use the graph convolutional layer to explore the relationship between features and fault types, but also take into account the similarity metrics between samples, resulting in a more accurate diagnose of the fault type in transformer.

D. *Comparison under different data volumes*

In order to further illustrate the effectiveness of GCN, the number of samples in the training set is expanded to 2 ~ 4 times of the original number by using the generative adversarial network [30], and the performance of each algorithm under different data volume is compared. The statistical results of the test set are shown in Table VIII.

The following conclusions can be drawn from Table VIII: 1) When the number of samples in the dataset is 718, the accuracy, Macro F1, and G-mean of GCN are not high, and the error of fault diagnosis is large. When the number of samples is extended to 2332, the performance of GCN is significantly improved. At this time, the accuracy is improved from 79.3% to 89.7%. The corresponding Macro F1 is improved from 77.2% to 88.5%, and the G-mean is improved from 79.1% to 89.6%. 2) In addition to GCN, the performance of the remaining six methods is improved with the increase of data volume. This phenomenon shows that with the increase of the number of samples in the training set, the information learned by various methods from the training samples is richer, thus strengthening the fault diagnosis ability of the models. 3) By comparing the accuracy, Macro F1, and G-mean under different data volumes, it is found that GCN performs better than other methods in different data volumes, which shows that GCN is not only suitable for the small data set, but also has light advantages over other algorithms in the case of large data volumes.

TABLE VIII
THE RESULT OF EACH METHOD WITH DATA VOLUMES

Method	Number of samples	Accuracy	Macro F1	G-mean
GCN	718	0.793	0.772	0.791
	1256	0.839	0.821	0.837
	1794	0.870	0.853	0.866
	2332	0.897	0.885	0.896
CNN	718	0.756	0.732	0.748
	1256	0.814	0.795	0.809
	1794	0.831	0.815	0.830
	2332	0.871	0.855	0.870
MLP	718	0.672	0.649	0.669
	1256	0.723	0.699	0.719
	1794	0.768	0.746	0.764
	2332	0.800	0.778	0.797
XGBoost	718	0.641	0.613	0.631
	1256	0.743	0.717	0.734
	1794	0.747	0.720	0.737
	2332	0.782	0.763	0.783
SVM	718	0.688	0.662	0.683
	1256	0.727	0.702	0.722
	1794	0.761	0.740	0.757
	2332	0.790	0.767	0.786
KNN	718	0.700	0.677	0.695
	1256	0.757	0.736	0.756
	1794	0.770	0.748	0.766
	2332	0.791	0.770	0.789
Siamese network	718	0.771	0.745	0.763
	1256	0.824	0.804	0.823
	1794	0.841	0.822	0.835
	2332	0.872	0.854	0.865

V. DISCUSSION

The objective of this paper is to propose a new method based on the GCN to diagnosis the transformer fault. In this paper, the effectiveness of the proposed GCN has been tested on the actual data set collected from the state grid corporation of China and previous publications. The simulation results show that the GCN achieves state-of-art performance with superior accuracy in different input features and data volumes for transformer fault diagnosis. However, the GCN needs to be retrained if the number of samples changes, because the size of the adjacency matrix A depends on the number of samples. In addition, there is a possible solution to avoid retraining GCN: for a new sample, an old sample which is the closest to this new sample is found. Then, the old sample is replaced with the new sample. In other words, it considers that the new sample and the old sample have the same connection relationship with others. In this case, GCN can diagnose new samples without repeated training.

GCN has shown superior performance over convolutional neural networks in some graph-structured data. The use of GCN is not limited to classification. For example, GCN may also be suitable for solving the reactive power optimization model of active distribution networks, because the input data of power flow calculation belongs to the graph-structured data. In addition, the GCN can be used to capture spatial correlations among multiple adjacent wind farms to improve the accuracy of short-term prediction.

VI. Conclusion

In order to improve the accuracy of the transformer fault diagnosis, this paper proposes a transformer fault diagnosis method based on GCN. Through the simulation analysis on real data, the following conclusions are obtained:

1) The size of k, optimizer, the number of iteration and graph convolutional layers, and the structure of GCN have a great influence on the performance of transformer fault diagnosis. In short, it is necessary to determine the number of graph convolutional layers according to the size of the data set. Using Adam algorithm as the optimizer GCN achieves the highest performance. If the size of k is too large or too small, the result of GCN will be worse. When the size of k is set to 10, GCN can get better performance.

2) Compared with CNN, MLP, XGBoost, SVM, KNN, and Siamese network, GCN has the highest diagnostic accuracy under different input features. CNN ignores the similarity metrics between samples, which limits the accuracy of fault diagnosis. GCN can not only use the graph convolutional layer to explore the relationship between features and fault types, but also take into account the similarity metrics between samples, so as to diagnose the fault type of the transformer more accurately.

3) In both small and large data volumes, the performance of GCN is better than that of CNN, MLP, XGBoost, SVM, KNN, and Siamese network.

Acknowledgment

This manuscript is supported by the China Scholarship Council. The authors are very grateful for their help.

References

[1] C. R. Gao, B. Qi, Y. Gao, Z.W Zhu, and C. R. Li, "Kerr electro-optic sensor for electric field in large-scale oil–pressboard insulation structure," *IEEE Transactions on Instrumentation and Measurement*, vol. 68, no. 10, pp. 94520-94532, Oct. 2019.

[2] B. Qi, P. Zhang, Z. H. R, J. Y. Wang, C. R. Li, and J.X. Chen, "Rapid transformer health state recognition through canopy cluster-merging of dissolved gas data in high-dimensional space," *IEEE Access*, vol. 7, pp. 94520-94532, Jul. 2019.

[3] J. Faiz and M. Soleimani, "Dissolved gas analysis evaluation in electric power transformers using conventional methods a review," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 24, no. 2, pp. 1239-1248, Apr. 2017.

[4] C. X. Guo, B. Wang, Z. Y. Wu, M. Ren, Y. F. He, R. Albarracin, and M. Dong, "Transformer failure diagnosis using fuzzy association rule mining combined with case-based reasoning," *IET Generation, Transmission & Distribution*, vol. 14, no. 11, pp. 2202-2208, May. 2020.

[5] S. Sarkar, T. Sharma, A. Baral, B. Chatterje, D. Dey, and S. Chakravorti, "An expert system approach for transformer insulation diagnosis combining conventional diagnostic tests and PDC, RVM data," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 21, no. 2, pp. 882-891, Apr. 2014.

[6] P. B. Thote, M. B. Daigavane, P. M. Daigavane, and S. P. Gawande, "An intelligent hybrid approach using KNN-GA to enhance the performance of digital protection transformer scheme," *Canadian Journal of Electrical and Computer Engineering*, vol. 40, no. 3, pp. 151-161, Aug. 2017.

[7] R. J. Zhu, D. G. Hao, and S. F. Hu. (2020, Jun.). Transformer fault diagnosis based on convolutional Siamese network with small samples. *Proceedings of the CSU-EPSA*. [Online]. pp. 1–7. Available: https://kns.cnki.net/KCMS/detail/12.1251.TM.20200629.1628.009.html

[8] Y. Benmahamed, M. Teguar, and A. Boubakeur, "Application of SVM and KNN to Duval Pentagon 1 for transformer oil diagnosis," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 24, no. 6, pp. 3443-3451, Dec. 2017.

[9] K. Chatterjee, S. Dawn, V. K. Jadoun, and R. K. Jarial, "Novel prediction-reliability based graphical DGA technique using multi-layer perceptron network & gas ratio combination algorithm," *IET Science, Measurement & Technology*, vol. 13, no. 6, pp. 836-842, Jul. 2019.

[10] N. Li, B. L. Li, and L. Gao, "Transient stability assessment of power system based on XGBoost and factorization machine," *IEEE Access*, vol. 8, pp. 28403-28414, Jan. 2020.

[11] R. Punmiya, S. Choe, "Energy theft detection using gradient boosting theft detector with feature engineering-based preprocessing," *IEEE Transactions on Smart Grid*, vol. 10, no. 2, pp. 2326-2329, Mar. 2019.

[12] Y. Li, F. Y. Yu, Q. Cai, K. Yuan, R. Z. Wan, X. Y. Li, M. Y. Qian, P. F. Liu, J. W. Guo, J. Yu, T. Zheng, H. Yan, P. Hou, Y. M. Feng, S. Y. Wang, and L. Ding, "Image fusion of fault detection in power system based on deep learning," *Cluster Computing*, vol. 22, pp. 9435-9443, Feb. 2019.

[13] B. Luo, H. T. Wang, H. Q. Liu, B. Li, and F. Y. Peng, "Early fault detection of machine tools based on deep learning and dynamic identification," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 1, pp. 509-518, Jan. 2019.

[14] H. C. Hua, Y .C. Qin, C. T. Hao, and J. W. Cao, "Optimal energy management strategies for energy Internet via deep reinforcement learning approach," *Applied Energy*, vol. 239, pp. 598-609, Apr. 2019.

[15] L. J. Ge, W. L. Liao, S. X. Wang, B. B. Jensen, and J. R. Pillai, "modeling daily load profiles of distribution network for scenario generation using flow-based generative network," *IEEE Access*, vol.8, pp. 77587-77597, Apr. 2020.

[16] N. Mahmoudi, M. H. Samimi, and H. Mohseni, "Experiences with transformer diagnosis by DGA: case studies," *IET Generation, Transmission & Distribution*, vol. 13, no. 23, pp. 5431-5439, Nov. 2019.

[17] Y. B. Liu, Y. Liu, and C. Yang, "Modulation recognition with graph convolutional network," *IEEE Wireless Communications Letters*, vol. 9, no. 5, pp. 624-627, May. 2020.

[18] M. Eisen and A. Ribeiro, "Optimal wireless resource allocation with random edge graph neural networks," *IEEE Transactions on Signal Processing*, vol. 68, pp. 2977-2991, Apr. 2020.

[19] Q. Zhang, J. L. Chang, G. F. Meng, S. B. Xu, S. M. Xiang, and C. H. Pan, "Learning graph structure via graph convolutional networks," *Pattern Recognition*, vol. 95, pp. 308-318, Nov. 2019.

[20] Z. X. Yan, J. G. Ge, Y. L. Wu, L. X. Li, and T. Li, "Automatic virtual network embedding: a deep reinforcement learning approach with graph convolutional networks," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 6, pp. 1040-1057, Jun. 2020.

[21] C. Kim, K. Kim, P. Balaprakash, and M. Anitescu, "Graph convolutional neural networks for optimal load shedding under line contingency,"in *2019 IEEE Power & Energy Society General Meeting*, 2019, pp.1–5.

[22] R. Wang, B. C. Li, S. W. Hu, W. Q. Du, and M. Zhang, "Knowledge graph embedding via graph attenuated attention networks," *IEEE Access*, vol. 8, pp. 5212-5224, Dec. 2019.

[23] J. Jiang, R. Y. Chen, M. Chen, W. H. Wang, and C. H. Zhang, "Dynamic fault prediction of power transformers based on hidden markov model of dissolved gases analysis," *IEEE Transactions on Power Delivery*, vol. 34, no. 4, pp. 1393-1400, Aug. 2019.

[24] J. Faiz and M. Soleimani, "Assessment of computational intelligence and conventional dissolved gas analysis methods for transformer fault diagnosis," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 25, no. 5, pp. 1798-1806, Oct. 2018.

[25] L. J. Ge, W. L. Liao, S. X. Wang, B. B. Jensen, and J. R. Pillai, "Modeling daily load profiles of distribution network for scenario generation using flow-based generative network," *IEEE Access*, vol. 8, pp. 77587-77597, Apr. 2020.

[26] W. K. Yu and C. H. Zhao, "Online fault diagnosis for industrial processes with Bayesian network-based probabilistic ensemble learning strategy," *IEEE Transactions on Automation Science and Engineering*, vol. 16, no. 4, pp. 1922-1932, Oct. 2019.

[27] X. Dong, H. L. Gao, L. Guo, K. S. Li, and A. Duan, "Deep cost adaptive convolutional network: a classification method for imbalanced mechanical data," *IEEE Access*, vol. 8, pp. 71486-71496, Apr. 2019.

[28] J. L. Yin, X. S. Zhou , Y. J. Ma, Y. J. Wu, and X. N. Xu, "Power transformer fault diagnosis based on multi-class multi-kernel learning relevance vector machine," in *2015 IEEE International Conference on Mechatronics and Automation*, 2015, pp.1–5.

[29] Enwen Li. (2019, Jan.). Dissolved gas data in transformer oil---fault diagnosis of power transformers with membership degree. China. [Online].Available:https://www.ieee-dataport.org/documents/dissolved-

gas-data-transformer-oil-fault-diagnosis-power-transformers-membership-degree#

[30] Y. P. Liu, Z. Q. Xu, J. H. He, Q. Wang, S. G. Gao, and J. Zhao, "Data augmentation method for power transformer fault diagnosis based on conditional Wasserstein generative adversarial network," *Power System Technology*, vol. 44, no. 4, pp. 1505-1513, Feb. 2020.

Wenlong Liao received the B.S. degree from China Agricultural University, Beijing, China, in 2017. He received the M.S. degree from Tianjin University, Tianjin, China, in 2020. He is currently pursuing the Ph.D. degree in Aalborg University, Aalborg, Denmark. His current research interests include smart grids, machine learning, and renewable energy.

Dechang Yang received Ph.D. degree in Dortmund University of technology, Germany, in 2012. He is currently an associate professor in college of information and electrical engineering, China Agricultural University, China. His current research interests include smart grids, machine learning, and optimization and operation of distribution network.

Yusen Wang received the B.S. degree from China Agricultural University, Beijing, China, in 2017. He received the M.S. degree from KTH Royal Institute of Technology, Stockholm, Sweden, in 2019. He is currently pursuing the Ph.D. degree in KTH Royal Institute of Technology, Stockholm, Sweden. His current research interests include machine learning, and fault diagnosis for transformer and power quality.

Xiang Ren received the B.S. degree from China Agricultural University, Beijing, China, in 2017. He received the M.S. degree from Tianjin University, Tianjin, China, in 2020. He is currently working for North China Electric Power Research Institute Co., Ltd., Beijing, China. His current research interests include Transformer fault diagnosis and relay protection.