Hindawi
Security and Communication Networks
Volume 2022, Article ID 5357412, 11 pages
https://doi.org/10.1155/2022/5357412

Research Article

A Power Transformer Fault Prediction Method through Temporal Convolutional Network on Dissolved Gas Chromatography Data

Mengda Xing[1,2,3] **Weilong Ding**[1,3] **Han Li**[1,3] **and Tianpu Zhang**[1,3]

[1]*School of Information Science and Technology, North China University of Technology, Beijing, China*
[2]*Artificial Intelligence on Electric Power System State Grid Corporation Joint Laboratory (GEIRI), Global Energy Interconnection Research Institute Co. Ltd., Beijing 102209, China*
[3]*Beijing Key Laboratory on Integration and Analysis of Large-Scale Stream Data, Beijing, China*

Correspondence should be addressed to Weilong Ding; dingweilong@ncut.edu.cn

Received 24 December 2021; Revised 26 January 2022; Accepted 18 February 2022; Published 11 April 2022

Academic Editor: Yuyu Yin

The power transformer is an example of the key equipment of power grid, and its potential faults limit the system availability and the enterprise security. However, fault prediction for power transformers has its limitations in low data quality, binary classification effect, and small sample learning. We propose a method for fault prediction for power transformers based on dissolved gas chromatography data: after data preprocessing of defective raw data, fault classification is performed based on the predictive regression results. Here, Mish-SN Temporal Convolutional Network (MSTCN) is introduced to improve the accuracy during the regression step. Several experiments are conducted using data set from China State Grid. The discussion of the results of experiments is provided.

1. Introduction

As key equipment, the power transformer is directly related to the system availability and enterprise security of power grid. Dissolved gas analysis (DGA) is one of the most reliable means for condition estimation and fault diagnosis of oil-immersed transformers and is recommended for condition evaluation by standards from the International Electrotechnical Commission (IEC) and the National Energy Administration. Through the gas chromatography online monitoring technology, business analyses such as transformer fault detection can be done in quasi-real time, which improve the safety and stability of power grid [1].

However, it faces challenges in predicting power transformer faults due to inherent limitations in practice. First, the low quality of raw data makes direct data usage infeasible because transmission links may be interrupted and data packets may be lost [2]. Due to the equipment or communication network problems, in-complete, missing, and outlier records exist in gas chromatography data. Availability is usually an essential requirement [3], and such defective data makes fault prediction more difficult. Second, traditional methods like widely used binary classification are not accurate enough, because such threshold-based fault detection technology ignores the data below the threshold and lacks historical trends employment. The oscillatory values around threshold may imply potential fault but cannot be found only by those methods. Third, the model is hard to be learned on small samples. The faults of power transformers appear casually, and related data must be a small proportion, and traditional models trained have to perform poorly due to the fact that too few features can be learned.

In this work, a fault prediction method is proposed for power transformers, which converts the classification problem for power transformers into a regression problem. Our contributions can be summarized as follows: (1) Missing imputation and outlier detection during the data preprocessing step guarantee completeness and continuity for gas chromatography data, which improve data quality obviously. (2) MSTCN proposed during regression step can learn features from data below fault threshold, which avoids overfitting through small sample learning. (3) On real-world data, our work shows convincing benefits and has been adopted in a practical business project.

The rest of this work is organized as follows: Section 2 discusses related work. Section 3 presents research background including motivation and methodology, as well as the transformer fault diagnosis method, AKA the three-ratio rule. Section 4 elaborates transformer fault prediction method based on MSTCN model. Section 5 evaluates the effects in extensive experiments. Section 6 summarizes the conclusion.

2. Related Work

Power transformer fault prediction is significant nowadays, but its discovery still faces challenges in efficiency and accuracy. Many works have adopted deep learning techniques in specific domains [4, 5]. We categorize related work into two technical perspectives: one is traditional algorithms through machine learning methods, and the other is deep learning methods, including recurrent neural networks (RNNs) and Temporal Convolutional Networks (TCNs).

2.1. Machine Learning Method. Machine learning methods can learn the fault occurrence pattern and then predict the possible faults. The literature [6] compared and analyzed MLP (Multilayer Perceptron), RBF (radial basis function), fuzzy logic, and support vector machine (SVM) for fault prediction of power transformers. However, their parameters are mainly selected empirically, which limits the efficiency of modeling. The $F1$-score of these methods is not more than 90% as evaluated by our data set.

Machine learning methods combined with DGA for transformer fault prediction have achieved many results. Dukarm [7] shows how fuzzy logic and neural networks are used to automate standard DGA methods. Furthermore, Wang et al. [8] conducted a combined artificial neural network and expert system tool (ANNEPS) developed for transformer fault diagnosis using dissolved gas-in-oil analysis. Huang et al. [9] introduced an evolutionary programming (EP) based fuzzy logic technique to identify the incipient faults of the power transformers. Yang et al. [10] employed bootstrap and genetic programming to improve the interpretation accuracy for DGA of power transformers. Hellmann [11] applied fuzzy logic (FL) that allows intermediate values to be defined between conventional evaluations like true/false, yes/no, high/low, and so forth. Souahlia et al. [12] applied the support vector machine (SVM) based decision for power transformers fault diagnosis.

However, these works have common problems, including the too small amount of data, few types of data, and only simple classification rules. For example, the fault categories are shallow, including overheating, discharging, and overheating with discharging. Advanced transformer fault prediction is required on Big Data fully using DGA data.

2.2. Deep Learning Method. In recent years, deep learning networks combined with DGA have further improved the accuracy of transformer fault prediction. Recurrent neural networks (RNNs) have been widely adopted in research areas concerned with sequential data, such as text, audio, and video [13]. Among RNNs methods, in particular Long Short-Term Memory (LSTM) [14] and Gated Recurrent Units (GRU) [15] are excellent in fully exploiting the time-varying features of time series data. Although the gradient problem of RNN has been solved to some extent in LSTM and GRU, it will still be tricky for longer sequences [13].

Bai et al. [16] proposed the Temporal Convolutional Networks (TCNs) model, a deep learning model for sequence modeling tasks. TCN combines convolutional neural network (CNN) and recurrent neural network ideas for processing time series type data. Almqvist [17] compared the performance of RNN and TCN for time series forecasting. Instead of using a cell state to preserve information from previous outputs as in LSTMs, TCNs use connection between previous hidden layers configured with two hyperparameters: dilation factor and filter size. Zhang et al. [18] proposed a multiscale temporal convolutional network for fault prediction. They extracted multiscale time-frequency information with the discrete wavelet transform, and each piece of scale data is handled by different TCN, respectively. Zhang et al. [19] presented an attention mechanism enhanced Temporal Convolutional Network for fault prediction. They utilized an attention mechanism to make the TCN-based fault prediction model focus on more essential input variables to enhance the fault prediction performance. Zai et al. [20] put forward a predictive method for dissolved gas content in transformer oil based on Temporal Convolutional Network (TCN) and graph convolutional network (GCN). They designed a GCN to analyze the correlations among all gases and then established a topological graph for their correlations.

However, these models did not solve the problem caused by the rectified linear unit (ReLU) and weight normalization layers. The rectified linear unit (ReLU) based activation function applied in TCN is underutilization of negative values leading to vanishing gradient. Meanwhile, the weight normalization applied in TCN is sensitive to initial values leading to overfitting. Meanwhile, these models used binary classification for fault prediction, which might not thoroughly learn the information below the threshold and lacks historical trends employment.

Inspired by the works in [21, 22], we propose a Mish-SN Temporal Convolutional Network (MSTCN) for dissolved gas regression to predict transformers' fault. We apply the Mish activation function and switchable normalization to MSTCN to solve the problems caused by ReLU and weight normalization. Meanwhile, the dissolved gas regression can explore the numerical fluctuations before the threshold value and learn historical fault feature patterns.

3. Preliminary

3.1. Motivation. Our work is originated from a practical project of China State Grid.

This work utilizes the dissolved gas chromatography data set provided by China State Grid as the data set for experimentation. The data set comes from the gas chromatography online monitoring equipment of the power grid, which is based on an integrated, high-speed two-way

communication network [23]. The data set covers roughly 600 transformers. With the explosive growth in Internet of Things (IoT) devices, applications have also substantially expanded in recent years [24]. Some of the data is a relatively long time series, containing more than 60 months of monitoring data, while others are short, only three or four months of monitoring data. In addition, each data item in the dissolved gas chromatography data is a multidimensional vector rather than a single number in some stock market and house price analysis data sets. The main fields in the dissolved gas chromatography data set of transformer oil are shown in Table 1. The data are all collected and measured automatically through the gas chromatography online monitoring technology.

Definition 1. Status code. In this work, a status code is used to identify each sample's possible fault categorical value. Status code is used as the classification label for the later transformer fault classification. The possible status code is summarized in Table 2.

The appearance of dissolved gas in the transformer oil indicates transformer faults. The gas formation comes from three conditions: overheating, discharge, and moisture. The amount of gas inside the transformer oil can be measured frequently by technical means to keep track of the operating health of the transformer. If any of the gases has a tendency to exceed a notice value, the gas production rate should be observed. However, if all the gases are lower than the notice value, the transformer is considered to be working properly. Based on the recommendations of the data provider, the notice values of our data set in this work are shown in Table 3.

3.2. The Three-Ratio Rule. We apply the three-ratio rule to converse dissolved gas regression to the status code in our proposed method. The three-ratio rule is proposed by the National Energy Administration of China [25]. By studying the trend of the dissolved gas amount in transformer oil, the status of the transformer can be determined based on the gas chromatography combined with the three-ratio rule. The conversion of three-ratio rule is shown in Tables 2 and 4.

Table 4 shows ratio code of two gases. For example, if the ratio of C_2H_2 to C_2H_4 is 0.2, the ratio code for C_2H_2/C_2H_4 is 1. Similarly, the other ratio codes for CH_4/H_2 and C_2H_2/C_2H_6 can be calculated. Table 2 shows the three-ratio codes and their corresponding faults. For example, if the ratio codes of C_2H_2/C_2H_4, CH_4/H_2, and C_2H_2/C_2H_6 are 1, 1, and 2, that is, the three-ratio code 112, the corresponding type of fault is low energy discharge, with the status code 6.

4. Power Transformer Fault Prediction Method

4.1. Overview. The fault prediction method based on dissolved gas regression proposed in this work is shown in Figure 1. Our method is divided into three steps. The first step is data preprocessing. Inspired by the work of Ding et al. [26], we convert data from different sources in the dissolved gas chromatography data set into a uniform format and resolve problems such as missing values and outliers in the data as much as possible. The second step is to predict gas amounts using a deep learning model. We apply MSTCN to dissolved regression gas regression to obtain future gas amounts. The third step is fault classification. The predicted transformer status code is calculated based on regression results of the second step and three-ratio rule mentioned above.

On the basis of transformer fault prediction studies, fault prediction methods usually use machine learning models or statistical tools to predict transformer fault. Instead of directly using deep learning models to predict transformer fault, we add a gas regression step between data preprocessing and fault classification. The usual fault prediction uses binary classification as predicting labels to do classification prediction, and the fault classification is judged based on the threshold value, ignoring the fluctuation of the value before the threshold value. The final prediction model might not learn the prethreshold value fluctuation information.

4.2. Data Preprocessing. In the domain of gas chromatography online monitoring technology of power transformer, there are problems such as network instability and server performance bottlenecks in processing extensive data. We mainly address the problem of missing data and outliers that exist in the dissolved gas chromatography data set.

Definition 2. Missing data. The missing data types include negative number, not a number (NAN), and null. Let $\mathbf{X} \in \mathbb{R}^{n\times m}$ be a feature matrix consisting of n data points and m features of dissolved gases. The t-th data point is denoted as $\mathbf{x}_t$. The j-th feature value of $\mathbf{x}_t$ is denoted as $\mathbf{x}_{tj}$. $\mathbf{x}_{tj}$ is defined as missing data if $\mathbf{x}_{tj} \notin [0, \infty)$.

The definition of outlier points combined with the characteristics of the data set and 3 σ-rule is shown as follows.

Definition 3. Outlier. Let $\mathbf{X} = \{\mathbf{x}_{t-k}, \ldots, \mathbf{x}_{t-1}, \mathbf{x}_{t+1}, \ldots, \mathbf{x}_{t+k}\}$ be defined as a set of the k-nearest neighbors of $\mathbf{x}_t$. Each of $\mathbf{X}$ is recorded at a specific time point $t \subseteq \mathbb{U}^+$ and consists of m observations that could be denoted as $\mathbf{x}_t = (x_{t1}, \ldots, x_{tm})$, each dimension j of m-dimensional vectors at a certain data point t could be denoted as x_{tj}, the expected value of x_{tj} could be denoted as $\widehat{x}_{\mathrm{tj}}$, the Euclidean distance of two data points can be denoted as d, and the highest distance threshold between a true data point and its expected data point could be denoted as 3σ. The outlier could be denoted as

$$\mathbf{x}_t \text{ is an outlier} \Leftrightarrow \left\{\forall \mathbf{x}_t \in X, \exists j \in \{1, \ldots, m\} | d\left(x_{\mathrm{tj}}, \widehat{x}_{\mathrm{tj}}\right) \geq 3\sigma\right\}. \tag{1}$$

With the definitions above, missing data and outlier problems are explicitly defined to be handled. Properly imputed data and corrected outliers could lower the regression errors and further promote fault prediction effectiveness:

Table 1: The main fields in the dissolved gas chromatography data set.

Collection date	C_2H_2	C_2H_4	C_2H_6	CH_4	CO	CO_2	H_2	Status code
2012-11-15	0.0	0.38	3.51	3.11	653.71	3391.83	6.74	0 (normal)
2012-11-16	0.0	0.47	2.885	11.5075	641.685	229097.165	5.555	2 (low temperature overheating)

Table 2: Fault category and status code through ratio code [25].

Three-ratio code	Nonfault and fault	Status code
000	Nonfault	0
001	Low temperature overheating (below 150°C)	1
020	Low temperature overheating (150~300°C)	2
021	Medium temperature overheating (300~700°C)	3
0* 2	High temperature overheating (above 700°C)	4
010	Partial discharge	5
10* 11*	Low energy discharge	6
12*	Low energy discharge and overheating	7
20* 21*	Arc discharge	8
22*	Arc discharge and overheating	9

*0, 1, and 2 for simplicity.

Table 3: Gas and threshold.

Gas	Threshold
H_2	10000
C_2H_2	10000
C_2H_4	5000
C_2H_6	1000
CH_4	5000
CO	10000
CO_2	20000

Table 4: Ratio code [25].

Gases ratio range	Code of ratio		
	C_2H_2/C_2H_4	CH_4/H_2	C_2H_2/C_2H_6
(0, 0.1)	0	1	0
[0.1, 1)	1	0	0
[1, 3)	1	2	1
[3, +∞]	2	2	2

Missing Data Imputation. For the missing data mentioned earlier, considering the data characteristics of gas amount, in this work, we took a modification of the EM algorithm proposed by Junger [27]. The algorithm comprises the following steps: (i) replace the missing values by estimates; (ii) estimate parameters μ and σ; (iii) estimate the level for each of the univariate pieces of data; (iv) reestimate the missing values using updated estimates of the parameters and the level of the data. These steps are iterated until some convergence criterion is reached.

Let $\mathbf{x}_t$ be the t data point of m features matrix $\mathbf{X}$. After $k+1$ iteration, the revised maximum likelihood estimates $\widetilde{x}_t = \left\{\widetilde{x}_{t1}^{(k+1)}, \mathbf{x}_{t2}\right\}$.

Outlier Correction. $\mathbf{X} = \left\{\mathbf{x}_{t-k}, \ldots, \mathbf{x}_{t-1}, \mathbf{x}_{t+1}, \ldots, \mathbf{x}_{t+k}\right\}$ is a set of the k-nearest neighbors of $\mathbf{x}_t$. Each of $\mathbf{X}$ is recorded at a specific time point $t \in \mathbb{Z}^+$ and consists of m real-valued observations that could be denoted as $\mathbf{x}_t = (x_{t1}, \ldots, x_{\mathrm{tm}})$, each dimension j of m-dimensional vectors at a certain data point t could be denoted as x_{tj}, the expected value of x_{tj} could be denoted as $\widehat{x}_{\mathrm{tj}}$, and the highest distance threshold between a true data point and its expected data point could be denoted as 3σ. In the outlier equation (1), the expected value $\widehat{x}_{\mathrm{tj}}$ and the highest distance threshold 3σ are defined in the two following equations:

$$\widehat{x}_{\mathrm{tj}} = \frac{1}{2k}\left(\sum_{i=1}^{k}\left(x_{t+i,j} + x_{t-i,j}\right)\right), \tag{2}$$

$$\sigma_j = \sqrt{\frac{1}{2k}\sum_{i=1}^{k}\left(\left(x_{t-i,j} - \widehat{x}_{\mathrm{tj}}\right)^2 + \left(x_{t+i,j} - \widehat{x}_{tj}\right)^2\right)}. \tag{3}$$

The expected value $\widehat{x}_{tj}$ is also the corrected value of the outlier $\mathbf{x}_t$.

4.3. Dissolved Gas Regression. In order to fully explore the numerical fluctuations before the threshold value and learn historical fault feature patterns, we proposed a regression model called Mish-SN Temporal Convolutional Networks.

On the other hand, if the predicted value of dissolved gas is obtained from the prediction model, the conversion from the predicted value of dissolved gas to the predicted value of the status code can be achieved with very few calculations.

Figure 2 shows a complete MSTCN map formed by stacking h residual blocks. The input is denoted as $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_t)$. In this work, we use a common technique in RNNs modeling called **time step** to improve predictive accuracy. The time step length could be denoted as L. The number of features is denoted as m. Let $\mathbf{v}_t \in \mathbb{R}^{L*m} = (\mathbf{x}_{t-L+1}, \ldots, \mathbf{x}_t)$ be defined as a new data point. For any $\mathbf{v}_t$, its gas regression label is denoted as $\mathbf{y}_t = (\mathbf{x}_{t+1}, \ldots, \mathbf{x}_{t+L})$.

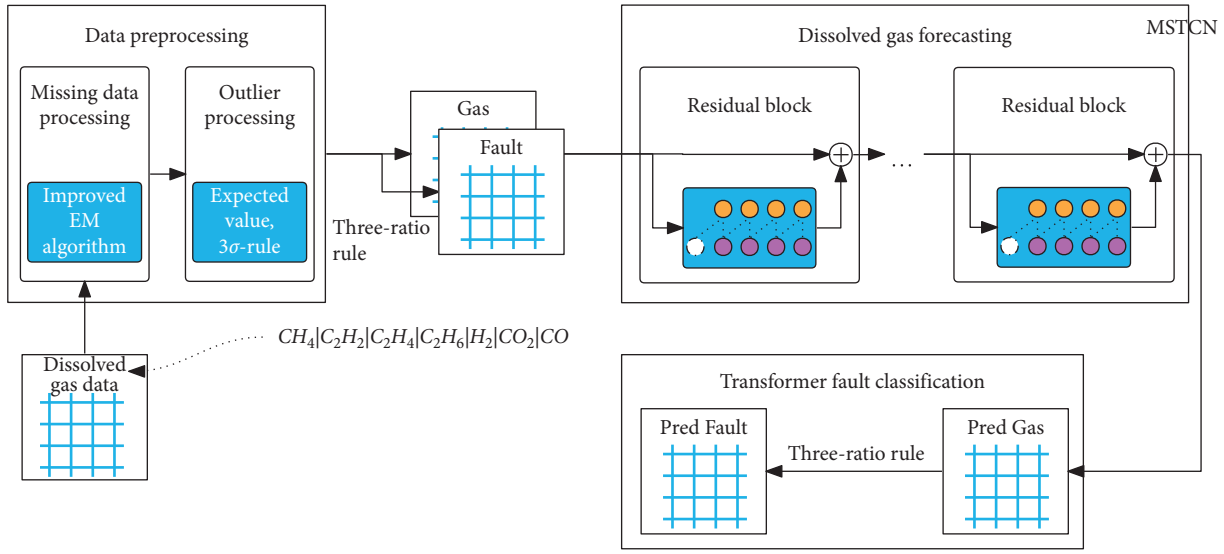

Figure 1: The architecture of our method.

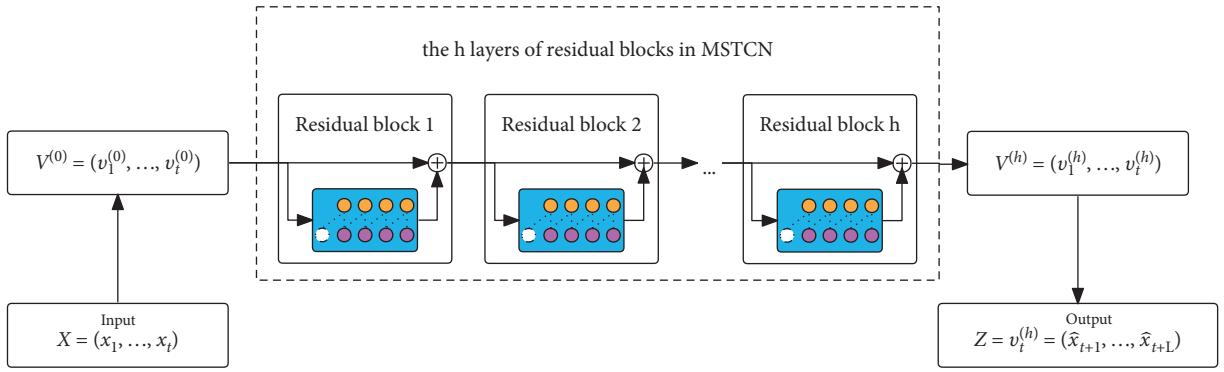

Figure 2: The architecture of MSTCN.

Therefore, the regression result of $\mathbf{v}_t$ is denoted as $\hat{y}_t = (\hat{x}_{t+1}, \ldots, \hat{x}_{t+L})$. The convoluted result of the h-th residual block layer is denoted as $\mathbf{V}^{(h)} = (\mathbf{v}_1^{(h)}, \ldots, \mathbf{v}_t^{(h)})$. However, to solve the real-world problem, we are only interested in the last case $\mathbf{v}_t$. $\mathbf{v}_t^{(h)} = (\hat{x}_{t+1}, \ldots, \hat{x}_{t+L})$ represents L regression points of input $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_t)$. The output $\mathbf{Z}$ of MSTCN regression result is shown as follows:

$$\begin{aligned}\mathbf{Z} &= \mathbf{V}^{(h)}[:\ -1] = \left(\mathbf{v}_1^{(h)}, \ldots, \mathbf{v}_t^{(h)}\right)[:\ -1] \\ &= \mathbf{v}_t^{(h)} = \left(\hat{x}_{t+1}, \ldots, \hat{x}_{t+L}\right).\end{aligned} \tag{4}$$

Residual Block. In order to solve the vanishing gradient problem, in a deep convolution network, a well-known technique called residual blocks is applied in MSTCN shown in Figure 3. Residual blocks have been proven to be an effective method for training deep networks, which enables the network to transmit information in a cross-layer manner.

In Figure 3, the upper branch of the residual block presents dilated causal convolution $\mathscr{H}(\cdot)$ with the input $\mathbf{V}^{(h)} = (\mathbf{v}_1^{(h)}, \ldots, \mathbf{v}_t^{(h)})$. The lower branch is the skip connections added to solve the vanishing gradient problem. In this work, we replace weight normalization with switchable normalization. Through the switchable normalization self-learning method, let the MSTCN decide which normalizer to use to obtain the best prediction effect. The MSTCN also introduces the Mish activation function to replace the ReLU for solving the dead ReLU problem in order to make the activation function smooth and derivable at 0 points and to improve the generalization of the model. Let $\delta(\cdot)$ be the activation layer. The output $\mathbf{V}^{(h)}$ could be expressed as

$$\mathbf{V}^{(h)} = \delta\left(\mathscr{H}\left(\mathbf{V}^{(h-1)}\right) + \mathbf{V}^{(h-1)}\right). \tag{5}$$

Dilated Casual Convolution. Figure 4 presents the structure of the dilated causal convolution stack from a residual block with filter size $k = 2$ and dilation factor $d = 3$. In Figure 4, the other layers and skip connection are omitted. The input of dilated casual convolution is denoted as $\mathbf{V}^{(h-1)} \in \mathbb{R}^{t*L} = (\mathbf{v}_1^{(h-1)}, \ldots, \mathbf{v}_t^{(h-1)})$. The output of dilated casual convolution is denoted as $\mathbf{V}^{(h)} \in \mathbb{R}^{t*L} = (\mathbf{v}_1^{(h)}, \ldots, \mathbf{v}_t^{(h)})$. Inspired by the idea of

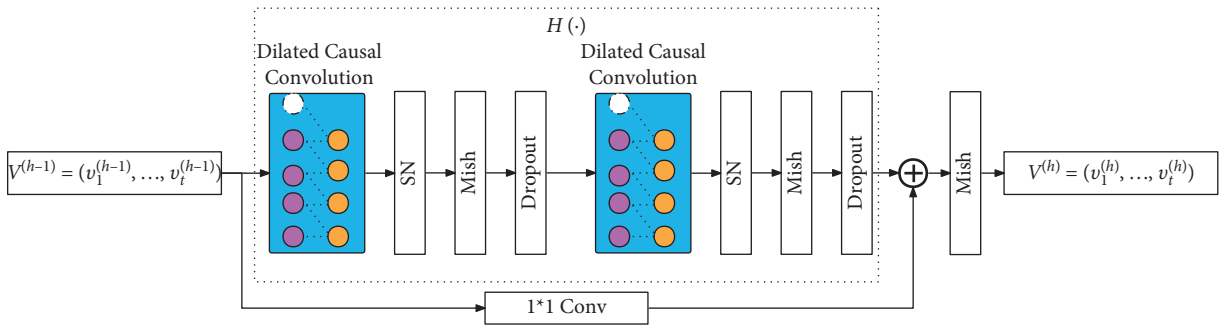

Figure 3: The architecture of the residual block in MSTCN.

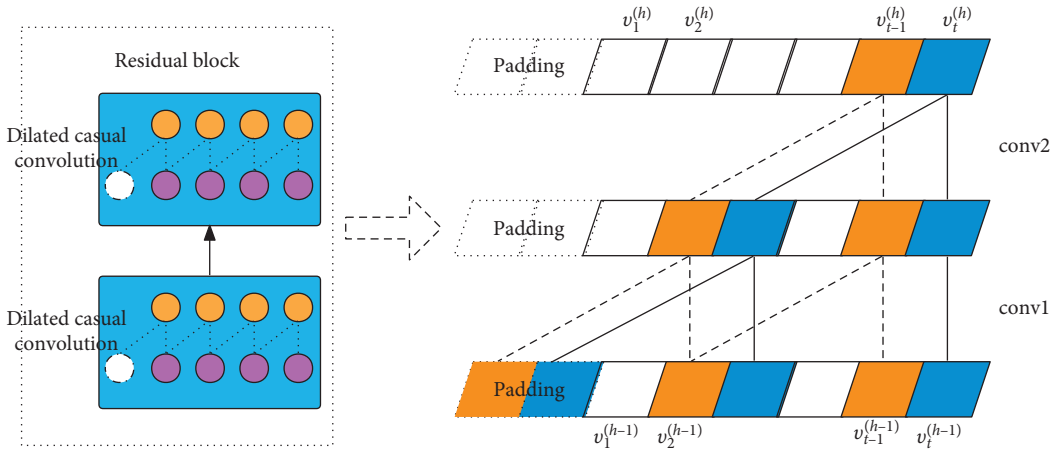

Figure 4: A dilated causal convolution with dilation factor $d = 3$ and filter size $k = 2$.

dilated convolution [28] and casual convolution [29], we set a constraint according to concept of casual convolution that any $\mathbf{v}_t^{(h)}$ only depends on $\mathbf{v}_1^{(h-1)}, \ldots, \mathbf{v}_t^{(h-1)}$ and not on future $\mathbf{v}_{t+1}^{(h-1)}, \ldots$. Meanwhile, to enlarge receptive field without deepening the structure, we apply concept of dilated convolution to the residual block.

A constraint according to concept of casual convolution that any $\mathbf{v}_t^{(h)}$ only depends on $\mathbf{v}_1^{(h-1)}, \ldots, \mathbf{v}_t^{(h-1)}$ and not on future $\mathbf{v}_{t+1}^{(h-1)}, \ldots$ is shown in Figure 4. To enlarge receptive field without deepening the structure, the MSTCN introduces dilated convolution.

4.4. Fault Classification. The guidelines [25] stipulate that, in the oil chromatographic analysis, if the content of each gas has a tendency to increase or exceeds a notice value, the gas production rate should be observed, and the gas production rate should be observed based on the three-ratio rule; it could be preliminarily judged that there is an overheating fault or a discharging fault, according to the three-ratio rule of gas chromatography in Table 4.

Let $\mathbf{U} \in \mathbb{R}^L = \{u_t, \ldots, u_{t+L}\}$ be defined as the status code. Let $\mathbf{Z} \in \mathbb{R}^{L*n_{\text{gas}}} = \{\hat{x}_t, \ldots, \hat{x}_{t+L}\}$ be defined as a set of L regression results. Let gas $\in \{C_2H_2, C_2H_4, C_2H_6, H_2, CH_4\}$; $\{r_1, r_2, r_3\} \in \mathbb{Z}^3$. Let n_{gas} be denoted as the feature numbers of $\hat{x}_t$. Let $\hat{x}_{t,\text{gas}} \in \mathbb{R}$ be denoted as the regression value of dissolved gas. The fault classification algorithm is defined in Algorithm 1.

5. Evaluation

5.1. Setting. The experiments in this work are running on a server with CentOS 7 operating system installed with Intel Core i7-6700 CPU, 16 GB RAM, and 1 TB storage. The experiments are written in Python 3.9.6, implementing JupyterLab 3.1.11, TensorFlow 2.5.0, and Matplotlib 3.3.4.

The data set was collected from oil-immersed power transformers in different substations in China, with 200,000 records covering the period from 2012 to 2017. The data set contains 7 fault-related gases ($H_2, C_2H_2, C_2H_4, C_2H_6, CH_4, CO, CO_2$), the time of collection, other gases (N_2, O_2), substation and transformer information, and so forth. The distribution of 7 faulty gases is shown in Figure 5. The horizontal axis is the date. The vertical axis is the amount of gas. The blue curve indicates no fault on the corresponding date. The orange curve indicates a fault status on the corresponding date because the CO_2 amount has reached its notice value.

We selected a subset composed of 100 transformers of about 170,000 records from the data set, divided into 80 training sets, 10 validation sets, and 10 test sets. The time range of the subset is from November 2012 to September 2017. The reason for the selection is that it has high data

Input: Regression result $\mathbf{Z} = \{\hat{x}_{t+1}, \ldots, \hat{x}_{t+L}\}$
Output: Status code $\mathbf{U} = \{u_{t+1}, \ldots, u_{t+L}\}$
(1) **for** t + 1 to $t + L$
(2) compute three ratios: $r_1 = \hat{x}_{tC_2H_2}/\hat{x}_{tC_2H_4}$, $r_2 = \hat{x}_{tCH_4}/\hat{x}_{tH_2}$, $r_3 = \hat{x}_{tC_2H_4}/\hat{x}_{tC_2H_6}$
(3) look up Table 4 to convert the gas ratio to ratio code
(4) combine the three-ratio code get combination $r = 100 \cdot r_1 + 10 \cdot r_2 + r_3$ and look up Table 2 to get the status code u_t
(5) **end for**
(6) **return U**

Algorithm 1: Fault classification algorithm.

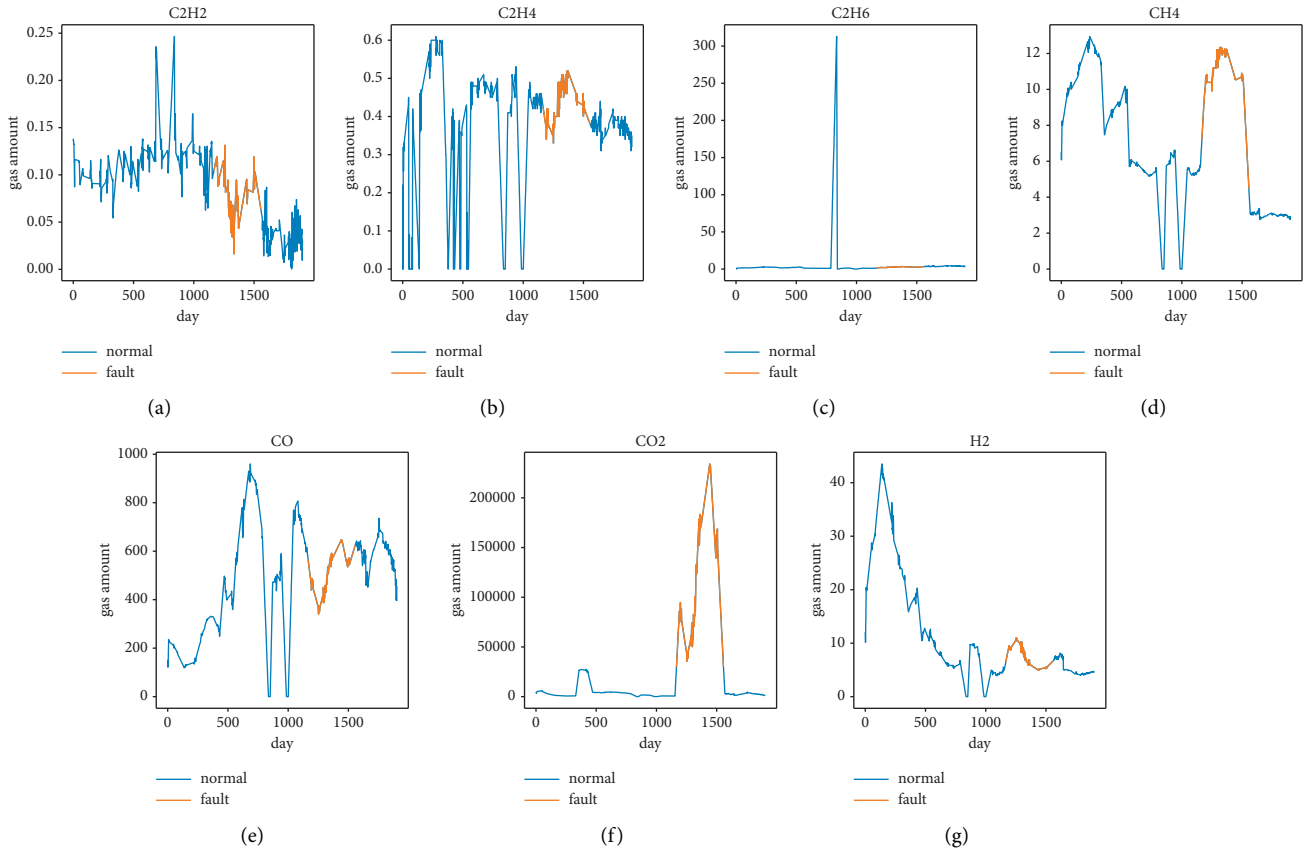

Figure 5: The gas distribution from 2012-08-21 to 2017-11-05 of a transformer. (a) C_2H_2. (b) C_2H_4. (c) C_2H_6. (d) CH_4. (e) CO. (f) CO_2. (g) H_2.

integrity and few missing values. In every transformer sequence of this data set, each record has attributes of collection date, 7 different dissolved gas values, and the status code label according to the three-ratio rule as shown in Table 1.

5.2. Experiment. In order to accurately evaluate the performance of the proposed transformer fault prediction model based on the MSTCN, we carried out dissolved gas regression and transformer fault classification experiments on the dissolved gas chromatography data set. First, we verify the average accuracy of dissolved gas regression based on the improved MSTCN. Second, we verify our proposed method by comparing it with other fault classification models based on binary classification and analyze the effectiveness of the models.

Experiment 1. Dissolved Gas Regression. The experiment applies MSTCN, TCN, LSTM, and GRU, respectively, on the test set to verify the effectiveness of the MSTCN model. The final parameters of MSTCN are defined in configuration: number of epochs is 100, batch size is 32, time step is 12, and learning rate is 0.001. The final parameters of residual block in MSTCN are shown in Table 5. For the TCN, LSTM, and GRU methods, they have roughly the same parameters as MSTCN, considering the rigour of the experiment. This work applies the root mean square error (RMSE) as the loss function shown in equation (6) and Adam as the

Table 5: Hyperparameter of the residual block.

Layer	Parameter	Value
conv1, conv2	Filters	10
	Kernel size	3
	Stride	1
	Padding	Same
	Dilation	1, 2, 4, 8, 16
dropout1, dropout2	Dropout rate	0.5

optimization algorithm. m represents the total m records, y_i represents the actual gas amount of record i, and $\hat{y}_i$ represents the predicted gas amount.

m represents the total m records, y_i represents the actual gas amount of record i, $\overline{y}$ represents the average value of actual gas amount, and $\hat{y}_i$ represents the predicted gas amount. The minimum of RMSE, MAE, and MAPE is 0, and the closer the metric is to 0, the better the predictive effect is. The maximum of R^2 is 1; the closer to 1 the better.

In order to measure the predictive performance of the models, RMSE, MAE, MAPE, and R^2 are used as the models' metrics. The calculation formulas of those metrics are shown in equation (6).

Figure 6 shows the actual gas amount curves and the regression curves predicted by different models, including MSTCN, TCN, LSTM, and GRU. It can be seen from Figure 6 that the fit curve of MSTCN is more accurate than the curves of the other models. Although, in the predictions from (g) H_2, LSTM performed better, overall, the MSTCN error is smaller than those in other models.

We have calculated above metrics of MSTCN, TCN, LSTM, and GRU. The results are listed in Table 6.

Table 6 shows the comparison of the MSTCN model and other deep learning models (TCN, LSTM, and GRU) as regards gas regression effect. MSE, MAE, and MAPE are used to measure the error between the true value and the predicted value of the data; R^2 is also used to measure the difference between the true value and the predicted value of the data and to standardize this difference to $[0, 1]$. For the predictions of $C_2H_4, C_2H_6, CH_4, CO, CO_2, H_2$, MSTCN has achieved a relatively good evaluation index. Although the prediction of C_2H_2 TCN has more minor prediction errors (RMSE), MSTCN is overall significantly better than other models.

$$\begin{aligned}
\mathrm{RMSE} &= \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2},\\
\mathrm{MAE} &= \frac{1}{m}\sum_{i=1}^{m}\left|\hat{y}_i - y_i\right|,\\
\mathrm{MAPE} &= \frac{100\%}{m}\sum_{i=1}^{m}\left|\frac{\hat{y}_i - y_i}{y_i}\right|,\\
R^2 &= 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y})^2}{\sum_{i=1}^{m}(y_i - \overline{y})^2}.
\end{aligned} \tag{6}$$

Experiment 2. Fault Classification. In order to further verify the superiority of the transformer fault prediction method proposed in this work, this experiment uses the regression value of the previous experiment as input. It converts the predicted gas amount to the status code according to the three-ratio rule. The control group uses TCN, LSTM, and GRU models and uses actual gas amount as input to directly classify the fault of the transformer.

In order to measure the accuracy of fault classification under different models, according to the confusion matrix, this experiment denotes faulty status as positive (P) and normal status as negative (N). Therefore, the correct fault classification could be denoted as true positive (TP) and true negative (TN), and the incorrect prediction could be denoted as false positive (FP) and false negative (FN) [30]. This experiment introduces three metrics to measure the model's accuracy on the test set. The precision, recall, and $F1$-score are expressed in equations (7)–(9). The $F1$-score is a harmonic mean of precision and recall, whose value is also between 0 and 1, as well as precision and recall. The more the three metrics are close to 1, the better predictive effect the model has.

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{7}$$

$$\mathrm{recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{8}$$

$$\mathrm{F1-score} = 2 \times \frac{\mathrm{precision} \times \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}. \tag{9}$$

Comparing the accuracy of fault classification under different models, the evaluation metrics are shown in Table 7.

Table 7 shows the comparison of transformer fault classification results between the MSTCN model and other deep learning models (TCN, LSTM, and GRU). The first column indicates the transformers participating in the experiment. Each transformer is an independent and complete experiment. The second column presents the evaluation metrics. The following columns are a comparison of the four model evaluation metrics. Overall, the prediction results of each model are satisfactory. This is caused by the faulty gas three-ratio algorithm and the gas attention value. For example, although the model has a deviation between the predicted gas value and the true value, it is still in the same ratio range, or the failure attention value is not reached at all, so the final predicted failure state will not change easily, resulting in an excellent overall prediction effect. For different transformers, the difference in fault prediction effect is more significant than the difference between models. The effect of model prediction is more affected by the data set than the model difference. For different models, the difference in failure prediction effects is relatively small. Overall, MSTCN is slightly higher than other models.

With the proposed transformer fault prediction method in this work in Figure 2, it can reduce or eliminate the impact of low accuracy of classification caused by threshold-based

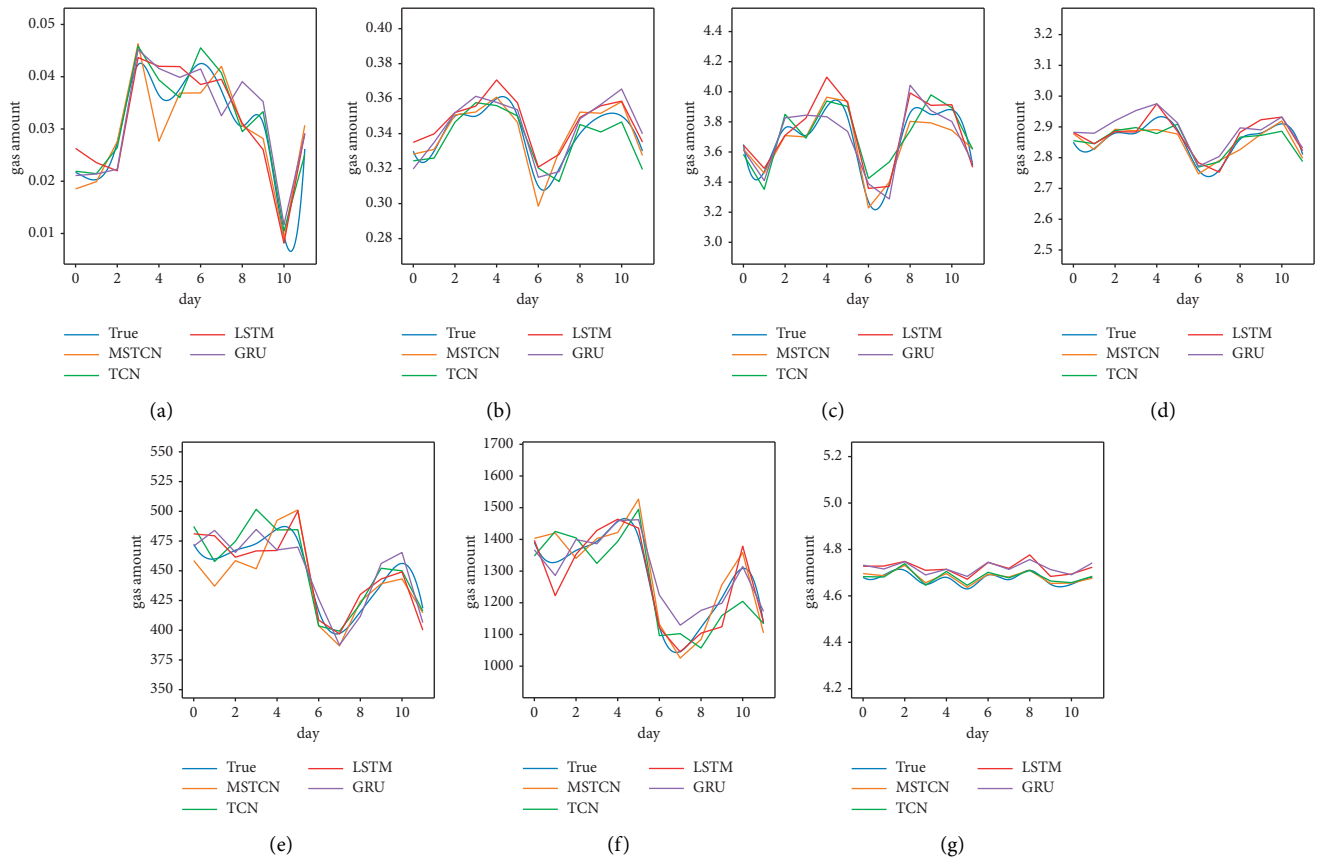

Figure 6: Comparation of gas regression between MSTCN and other models. (a) C_2H_2. (b) C_2H_4. (c) C_2H_6. (d) CH_4. (e) CO. (f) CO_2. (g) H_2.

Table 6: Gas regression results of transformer no. 1.

Gas	Metrics	MSTCN	TCN	LSTM	GRU
C_2H_2	RMSE	0.0051	0.0026	0.0065	0.0042
	MAE	0.0043	0.0020	0.0055	0.0036
	MAPE	14.07%	5.80%	17.96%	11.66%
	R^2	0.7091	0.9254	0.5218	0.7975
C_2H_4	RMSE	0.0056	0.0058	0.0089	0.0065
	MAE	0.0045	0.0045	0.0071	0.0052
	MAPE	1.32%	1.31%	2.05%	1.51%
	R^2	0.8503	0.8393	0.6161	0.7967
C_2H_6	RMSE	0.0494	0.0567	0.0768	0.0721
	MAE	0.0450	0.0491	0.0646	0.0630
	MAPE	1.24%	1.33%	1.76%	1.73%
	R^2	0.9419	0.9237	0.8599	0.8764
CH_4	RMSE	0.0138	0.0200	0.0284	0.0350
	MAE	0.0116	0.0155	0.0230	0.0314
	MAPE	0.41%	0.54%	0.81%	1.10%
	R^2	0.9294	0.8516	0.7019	0.5483
CO	RMSE	12.2925	12.3431	13.9061	16.6197
	MAE	8.8837	10.1510	11.9101	13.2390
	MAPE	2.03%	2.26%	2.65%	2.87%
	R^2	0.8095	0.8079	0.7562	0.6518
CO_2	RMSE	44.8718	50.1873	47.7348	53.6491
	MAE	31.9983	40.4018	40.1977	41.3562
	MAPE	2.56%	3.11%	3.07%	3.25%
	R^2	0.8856	0.8569	0.8705	0.8364
H_2	RMSE	0.0083	0.0108	0.0535	0.0468
	MAE	0.0062	0.0095	0.0524	0.0458
	MAPE	0.13%	0.20%	1.12%	0.98%
	R^2	0.8810	0.7981	−3.9776	−2.8074

Table 7: Classification metrics of different models.

Transformers	Metrics	MSTCN	TCN	LSTM	GRU
Transformer no. 1	Precision	99.20%	98.59%	98.55%	98.55%
	Recall	99.12%	98.29%	98.24%	98.24%
	*F*1-score	0.9914	0.9837	0.9831	0.9831
Transformer no. 2	Precision	99.33%	99.08%	98.87%	98.95%
	Recall	99.18%	98.76%	98.35%	98.53%
	*F*1-score	0.9922	0.9885	0.9850	0.9865
Transformer no. 3	Precision	98.67%	98.40%	98.23%	98.21%
	Recall	98.06%	97.41%	96.94%	96.88%
	*F*1-score	0.9823	0.9770	0.9733	0.9728
Transformer no. 4	Precision	98.16%	98.03%	97.94%	97.84%
	Recall	97.12%	96.76%	96.53%	96.24%
	*F*1-score	0.9741	0.9713	0.9694	0.9671

binary classification. It can use the data information before the threshold and enhance the usage of historical fault data because the proposed method is based on the dissolved gas regression value. At the same time, this classification step does not introduce additional errors because it uses the same judgment criteria as the existing fault diagnosis methods.

6. Conclusions

In this work, we propose a power transformer fault prediction method based on dissolved gas regression, which cleverly converts the transformer fault prediction problem into a regression problem for dissolved gas amount. First, through data preprocessing, we overcome the difficulties in directly using raw data. Second, by dissolved gas regression, we achieve more efficient learning of the data below threshold than binary classification and avoid small sample learning caused by a large amount of preventive maintenance. Compared with the traditional binary-based classification fault prediction model, the fault prediction method based on gas amount prediction has better results with *F*1-score more than 0.9741. This novel method provides new insights for power transformer fault prediction.

In summary, the fault prediction method based on dissolved gas regression using MSTCN has excellent potential. In future work, we will continue to research this concept and shorten the training time with more advanced deep learning techniques. In addition to the fault prediction method proposed, we plan to tune the procedure to simplify the method.

Data Availability

The oil chromatography data used to support the findings of this study were supplied by China State Grid under license and so cannot be made freely available. Requests for access to these data should be made to the corresponding author for an application of joint research.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the science and technology project of State Grid Corporation of China: "Research on Data Governance and Knowledge Mining Technology of Power IOT Based on Artificial Intelligence" (Grant No. 5700-202058184A-0-0-00).

References

[1] N. Muhamad, B. Phung, T. Blackburn, and K. Lai, "Comparative study and analysis of DGA methods for transformer mineral oil," in *Proceedings of the 2007 IEEE Lausanne Power Tech*, pp. 45–50, IEEE, Lausanne, Switzerland, July 2007.

[2] H. Gao, C. Liu, Y. Yin, Y. Xu, and Y. Li, "A hybrid approach to trust node assessment and management for VANETs cooperative data communication: historical interaction perspective," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–10, 2021.

[3] Y. Huang, H. Xu, H. Gao, X. Ma, and W. Hussain, "SSUR: an approach to optimizing virtual machine allocation strategy based on user requirements for cloud data center," *IEEE Transactions on Green Communications and Networking*, vol. 5, no. 2, pp. 670–681, 2021.

[4] dH. Faria Jr, J. G. S. Costa, and J. L. M. Olivas, "A review of monitoring methods for predictive maintenance of electric power transformers based on dissolved gas analysis," *Renewable and Sustainable Energy Reviews*, vol. 46, pp. 201–209, 2015.

[5] S. Yu, D. Zhao, W. Chen, and H. Hou, "Oil-immersed power transformer internal fault diagnosis research based on probabilistic neural network," *Procedia Computer Science*, vol. 83, pp. 1327–1331, 2016.

[6] K. Bacha, S. Souahlia, and M. Gossa, "Power transformer fault diagnosis based on dissolved gas analysis by support vector machine," *Electric Power Systems Research*, vol. 83, no. 1, pp. 73–79, 2012.

[7] J. J. Dukarm, "Transformer oil diagnosis using fuzzy logic and neural networks," in *Proceedings of the Canadian Conference on Electrical and Computer Engineering*, pp. 329–332, IEEE, Vancouver, BC, Canada, September 1993.

[8] Z. Wang, Y. Liu, and P. J. Griffin, "A Combined ANN and Expert System Tool for Transformer Fault Diagnosis," in *IEEE Transactions on Power Delivery*vol. 13, no. 4, , pp. 1261–1269, IEEE, 2000.

[9] Y. C. Huang, H. T. Hong-Tzer Yang, and C. L. Ching-Lien Huang, "Developing a new transformer fault diagnosis system through evolutionary fuzzy logic," *IEEE Transactions on Power Delivery*, vol. 12, no. 2, pp. 761–767, 1997.

[10] X. Yang, W. Chen, A. Li, C. Yang, Z. Xie, and H. Dong, "BA-PNN-based methods for power transformer fault diagnosis," *Advanced Engineering Informatics*, vol. 39, pp. 178–185, 2019.

[11] M. Hellmann, "Fuzzy logic introduction," *Université de Rennes*, vol. 1, pp. 1–9, 2001.

[12] S. Souahlia, K. Bacha, and A. Chaari, "SVM-based decision for power transformers fault diagnosis using Rogers and Doernenburg ratios DGA," in *Proceedings of the 10th International Multi-Conferences on Systems, Signals & Devices 2013 (SSD13)*, pp. 1–6, IEEE, Hammamet, Tunisia, March 2013.

[13] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: LSTM cells and network architectures," *Neural Computation*, vol. 31, no. 7, pp. 1235–1270, 2019.

[14] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: continual prediction with LSTM," *Neural Computation*, vol. 12, no. 10, pp. 2451–2471, 2000.

[15] K. Cho, B. Van Merriënboer, C. Gulcehre et al., "Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation," https://arxiv.org/abs/1406.1078.

[16] S. Bai, J. Z. Kolter, and V. Koltun, "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling," *CoRR*, vol. abs/1803.01271, 2018, https://arxiv.org/abs/1803.01271.

[17] O. Almqvist, "A comparative study between algorithms for time series forecasting on customer prediction: an investigation into the performance of ARIMA," *RNN, LSTM, TCN and HMM*, , Thesisvol. 52, 2019.

[18] J. Zhang, Y. Wang, J. Tang, J. Zou, and S. Fan, "A multiscale temporal convolutional network for fault diagnosis in industrial processes," in *Proceedings of the 2021 American Control Conference (ACC)*, May 2021.

[19] J. Zhang, Y. Chang, J. Zou, S. Fan, and T. C. N Ame, "Attention mechanism enhanced temporal convolutional network for fault diagnosis in industrial processes," in *Proceedings of the 2021 Global Reliability and Prognostics and Health Management (PHM-Nanjing)*, IEEE, Nanjing, China, October 2021.

[20] H. Zai, W. Chen, H. He et al., "Prediction for dissolved gas in power transformer oil based on temporal convolutional and graph convolutional network," in *Proceedings of the 2021 IEEE/IAS Industrial and Commercial Power System Asia (I&CPS Asia)*, pp. 1160–1168, IEEE, Chengdu, China, July 2021.

[21] M. D. Mish, "A self regularized non-monotonic neural activation function," vol. 4, p. 2, 2019, https://arxiv.org/abs/1908.08681.

[22] P. Luo, J. Ren, Z. Peng, R. Zhang, and J. Li, "Differentiable learning-to-normalize via switchable normalization," in *Proceedings of the International Conference on Learning Representation (ICLR)*, 2019.

[23] W. Ding, Z. Wang, Y. Xia, and K. Ma, *An Efficient Interpolation Method through Trends Prediction in Smart Power Grid*, Springer, Berlin, Germany, pp. 79–92, 2021.

[24] H. Gao, X. Qin, R. J. D. Barroso, W. Hussain, Y. Xu, and Y. Yin, "Collaborative learning-based industrial IoT API recommendation for software-defined devices: the implicit knowledge discovery perspective," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 6, no. 1, 2020.

[25] oN. E. A. China, "Guide to the analysis and the diagnosis of gases dissolved in transformer oil," *Tech. Rep.*, 2014.

[26] W. Ding, X. Wang, Z. Zhao, and S. T. A. R Co, "CO-STAR: a collaborative prediction service for short-term trends on continuous spatio-temporal data," *Future Generation Computer Systems*, vol. 102, pp. 481–493, 2020.

[27] W. L. Junger and A. Ponce de Leon, "Imputation of missing data in time series for air pollutants," *Atmospheric Environment*, vol. 102, pp. 96–104, 2015.

[28] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," https://arxiv.org/abs/1511.07122.

[29] T. J. Brazil, "Causal-convolution-a new method for the transient analysis of linear systems at microwave frequencies," *IEEE Transactions on Microwave Theory and Techniques*, vol. 43, no. 2, pp. 315–323, 1995.

[30] W. Ding, Z. Wang, J. Chen, Y. Xia, J. Wang, and Z. Zhao, "Potential trend discovery for highway drivers on spatio-temporal data," *Wireless Networks*, vol. 27, no. 5, pp. 3407–3422, 2021.