Bidirectional Recurrent Neural Network based on Multi-Kernel Learning Support Vector Machine for Transformer Fault Diagnosis

Xun Zhao[1], Shuai Chen[2*], Ke Gao[3], Lin Luo[4]
School of Information and Control Engineering, Liaoning Petrochemical University, Fushun, China[1, 2, 4]
Offshore Oil Engineering CO.LTD, Tianjin, China[3]

***Abstract*—Traditional neural network has many weaknesses, such as a lack of mining transformer timing relation, poor generalization of classification, and low classification accuracy of heterogeneous data. Aiming at questions raised, this paper proposes a bidirectional recurrent neural network model based on a multi-kernel learning support vector machine. Through a bidirectional recurrent neural network for feature extraction, the features of the before and after time fusion and obvious data are outputted. The multi-kernel learning support vector machine method was carried out on the characteristics of data classification. The study of multi-kernel support vector machines in the weighted average of the way nuclear fusion improves the accuracy of characteristic data classification. Numerical simulation analysis of the temporal channel length for sequential network diagnostic performance, the effects of multi-kernel learning on the generalization ability of support vector machine, the influence on heterogeneous data processing capabilities, and transformer fault data classification experiment verifies the correctness and effectiveness of the bidirectional recurrent neural network based on multi-kernel learning support vector machine model. The experiment result shows that the diagnosis performance of bidirectional recurrent networks based on a multi-kernel learning support vector machine is better, and the prediction accuracy of the model is improved by more than 1.78% compared with several commonly used neural networks.**

Keywords—Multi-kernel learning; support vector machine; bidirectional recurrent neural network; fault diagnosis

I. Introduction

A transformer, one of the key hub equipment in the power grid, act as an important link in energy conversion, distribution, or transmission. Transformer failure will cause huge financial loss and endanger public security. Therefore, timely and accurate diagnosis of transformer failure has important significance to make sure that the power system in safe state [1]. Characteristic gas method is often used for manual diagnosis of traditional transformer faults, but the differences in experience for fault identification often lead to errors. For transformer insulation maintenance, manual judgment requires not only a power failure of tested transformers but also regular maintenance of replaced equipment or parts. And a large part of equipment does not exceed its service life, which often causes a waste of resources and reduces the economy of the power system.

When the LSTM network conducts data analysis on time series feature quantity, the model is too simple and can only consider a single time series. The fault data identification is not ideal and the accuracy of long-time series will decline. Besides, the ability to generalize is not high. In this paper, a Bidirectional Long Short-Term Memory (Bi-LSTM) network is proposed and Multi-Kernel Learning Support Vector Machines (MKL-SVM) combined optimization algorithm. In this method, a reverse sequential LSTM network is added to establish a Bi-LSTM network model, which can consider the feature quantity of the future time. Then, the kernel functions used by SVM are aggregated together by the weighted average method to establish the MKL-SVM model, and MKL-SVM is used to replace the Softmax function to achieve fault classification. The feature of SVM is applied to increase the classification generalization ability, and multi-kernel learning has more excellent characteristics for heterogeneous data classification. A failure in the transformer oil dissolved gas H_2, CH_4, C_2H_6, C_2H_4, C_2H_2, the characteristics of the gas is chosen to verify the effectiveness of multi-kernel learning support vector machine and bidirectional recurrent neural network for transformer fault diagnosis. Compared with several existing prediction models, the model can make a more accurate judgment on a few heterogeneous data, with stronger generalization ability and higher prediction accuracy.

The rest of this paper is consisted of as: Section II presents the related works. The proposed method is described in Section III and Section IV. Experimental results and performance analysis are discussed in Section V. Finally, this paper concludes in Section VI.

II. Related Work

Machine learning is used increasingly frequently and effectively to process and analyze data as a result of technological advancements. Zhang Hang et al. proposed to use support vector machine to improve the accuracy of motor fault diagnosis, but the model has many limitations, including poor generalization ability, low classification efficiency and high interference accuracy [2]. Zhang Xin, Wang Heng et al. come up with a model of a neural network optimized by a sparrow search algorithm, which in view of transformer fault diagnosis to increase the correctness more effectively. However, the number of probabilistic neural network neurons was affected by training samples, and the time it takes to train the model would be increased to get a better model [3]. For increase the stability of SVM, Qingchuan Fan come up with a whale optimization algorithm that introduced the characteristics of a genetic algorithm to improve SVM but sacrificed part of the classification accuracy of SVM [4]. In view of the contradiction between the stability and accuracy of SVM, Yuhan Wu et al. proposed an SVM optimization algorithm that adopts adaptive probability formula to balance the squirrel search algorithm. The iteration cycle and accuracy of the model are increased [5]. Bing Zeng et al. proposed an optimized gray wolf algorithm Let-Squares SVM combining particle swarm optimization and differential evolution. The model got rid of the weakness which it was caught in local optimum frequently and improved the correct rate of the model, but the generalization ability of the model was greatly affected [6].

The emergence of deep learning brings new changes to machine learning. Miao Jianjie et al. proposed to use improved particle swarm optimization algorithm to optimize fuzzy neural network, automatically adjust parameters and accelerate convergence [7]. Gao Xincheng et al. proposed using improved genetic algorithm to optimize convolutional neural networks, which shortened the time to obtain the optimal weight and improved the convergence and accuracy of neural networks [8]. With the emergence of more complex neural networks, more deep-learning methods have been put into use for fault diagnosis. Taha Ibrahim B. M. et al. come up with an improved convolutional neural network suitable for noise environment, which perfected the accuracy of the Convolutional Neural Network (CNN) model. But the CNN model does not have a good classification effect on time series data and cannot effectively extract features from time series data [9]. In order to enable neural networks to better consider timing features, Fan Xiaodong et al. come up with apply Long Short-Term Memory (LSTM) network for transformer fault diagnosis, which has a better effect than CNN. When LSTM is applied to fault diagnosis, it has trouble to ensure the stability of the model and the accuracy of long-time series will decline as the number of training increases [10]. To improve the stability of the model, He Yigang and Wu Xiaoxin put forward a kind of complicated correlation characteristic of bidirectional Recurrent Neural Network (RNN). The stability of the model is higher and can be related to more features. However, when we use models, the ability to generalize is not high with only the specific training being set to get a better training result [11]. Omar Alharbi combine CNN and Bi-LSTM for generating the final features representation to be passed to a linear SVM classifier [12]. This method used for classification of Arabic reviews show that the method achieved superior performance than the two baseline algorithms of CNN and SVM in all datasets. The combination of multiple machine learning methods has a better development prospect.

III. Fault Diagnosis Principle of Transformer with Bidirectional Recurrent Neural Network

Dissolved Gas Analysis (DGA) in oil is often applied to transformer fault diagnosis, which has the merit of easy to operate and strong anti-interference ability. In order to ensure reliable data collection and effective analysis results, the DGA method is adopted in this paper. DGA is a series of classical time series data, which collects dissolved gas data in transformer oil in a fixed period without interruption in time sequence. Therefore, the bidirectional recurrent neural network can be used as the feature extraction part of the model to feature extraction of time sequence data in transformer oil. Make sure that the timing sequence feature extraction ability of RNN and the dependability of the real operation, in the passage adopts an LSTM network with better timing sequence extraction ability and stable operation.

A. Dissolved Gas Analysis Method in the Transformer Oil

Transformer faults are split into electrical faults, thermal faults, mechanical faults. Yet as the frequency of mechanical faults is the lowest with the rest of the several faults, the main analysis object of thermal faults and electrical faults.

When transformer is in the process of fault operation, the insulation oil of the transformer will be oxidized and cracked due to the action of discharge and heat. The main composition of the insulation oil hydrocarbon will produce hydrogen and low molecular alkenes, alkenes, alkynes, and other gases. With the severity of the fault, the rising rate of each gas is different, and the transformer fault type can be roughly judged according to the different types of gas and concentration. Table I lists the characteristic gases of different faults.

TABLE I. The Characteristic Gases of Different Faults

Fault Types	**Main Characteristic Gases**	**Accompany Characteristic Gases**
Oil overheating	Methane CH_4, Acetylene C_2H_2	Hydrogen H_2, Ethane C_2H_6
Oil and paper overheating	Methane CH_4, Ethylene C_2H_4, Carbon monoxide CO	Hydrogen H_2, Ethane C_2H_6, Carbon dioxide CO_2
Partial discharge in oil paper insulation	Hydrogen H_2, Methane CH_4, Carbon monoxide CO	Ethylene C_2H_4 、 Ethane C_2H_6、Acetylene C_2H_2
Spark discharge in oil	Hydrogen H_2, Acetylene C_2H_2	
Arc in oil	Hydrogen H_2, Acetylene C_2H_2, Ethylene C_2H_4	Methane CH_4, Ethane C_2H_6

According to Table I, the characteristics of gas under different fault types, the transformer insulating oil by electrolysis produces several kinds of characteristic gas for parameters. And the characteristics of gas concentration and velocity were analyzed in different operation conditions while the operating state of the transformer was assessed to decision the transformer fault types. The transformer oil is a dissolved gas analysis method. This analysis method has the advantage of supporting live online detection to spare it from being impacted by the signal of electric and magnetic fields and has a simple operation mode. DGA applied in transformer state monitoring and fault diagnosis [13-14].

H_2, CH_4, C_2H_4, C_2H_2, C_2H_6, and the other five gases in the gas concentration detection results are more accurate and can determine the transformer fault type. In this experiment, H_2, CH_4, C_2H_4, C_2H_2, C_2H_6 were chosen as the characteristic gases of experiment by using transformer fault data. Fig. 1 shows the concentration curves of the five characteristic gases.

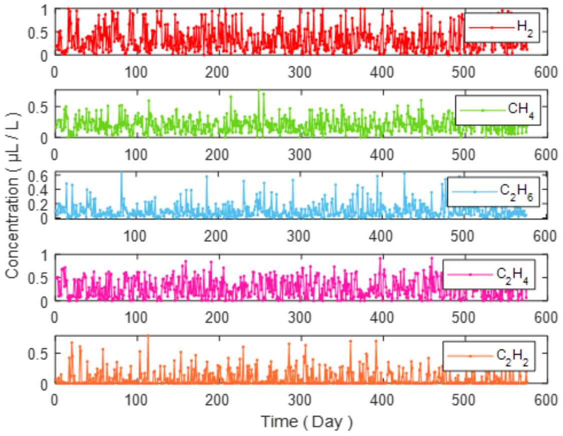


Fig. 1. The concentration curves of the five characteristic gases

B. Long Short-Term Memory Network

The perfect forecast of dissolved gas in oil can more effectively understand the operation state of the transformer and make a timely judgment of the fault. The Neural structure of a Recurrent Neural Network (RNN) has a self-feedback function, which can retain both current and previous information at the same time and can be used to calculate the current output [15]. However, when the RNN model is used to analyze the long time series data, there are defects of gradient explosion or gradient disappearance present to delayed backpropagation during the training process. So the RNN network is not good at analyzing long-series data [16]. LSTM network is a kind of RNN model that can store time sequence information for a long time by adding a gating unit on the basis of a general recurrent neural network. As a way of deep learning, the recurrent neural network can extract data features from time series more efficiently and accurately. The quality of features determines the accuracy of classification, and accurate classification of fault data can make a more timely response to transformer faults. For increasing the precision rate of neural network feature extraction for time series data, Take LSTM model as the prototype, the reverse time-series memory network was introduced to increase the dependence of characteristic gases on time series information and increase the precision rate gas prediction by the model.

The basic unit structure of the LSTM is shown in Fig. 2 is the input of the current moment while being the value of the previous moment in memory unit. The activation function σ usually used the sigmoid function. Compared with RNN, LSTM is characterized by a gating mechanism, which includes three parts, namely the forget gate, input gate, and output gate [17]. The responsibility of forget gate is selectively forgetting the historical information stored in the memory unit. The input gate retains the current external information and integrates it with the historical information. And the output gate decides whether to output. The output of the LSTM is determined jointly by the gating unit and the input.

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_f[\boldsymbol{X}_t;\boldsymbol{Y}_{t-1}]+b_f) \tag{1}$$

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_i[\boldsymbol{X}_t;\boldsymbol{Y}_{t-1}]+b_i) \tag{2}$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_o[\boldsymbol{X}_t;\boldsymbol{Y}_{t-1}]+b_o) \tag{3}$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \cdot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \cdot g(\boldsymbol{X}_t) \tag{4}$$

$$\boldsymbol{Y}_t = \boldsymbol{o}_t \cdot h(\boldsymbol{c}_t) \tag{5}$$

In the above equation, W_f and b_f, $\boldsymbol{W}_i$ and b_i, $\boldsymbol{W}_o$ and b_o are weight and bias for the forget gate, input gate, output gate respectively. $g(\,)$ and $h(\,)$ are mappings of $\boldsymbol{X}_t \rightarrow \boldsymbol{c}_t$ and $\boldsymbol{c}_t \rightarrow \boldsymbol{Y}_t$ respectively represents matrix multiplication. $\boldsymbol{c}_t$ contains both the input state and the history state, which improves the stability of the neural network and solves the problem of gradient disappearance that often occurs in the RNN.

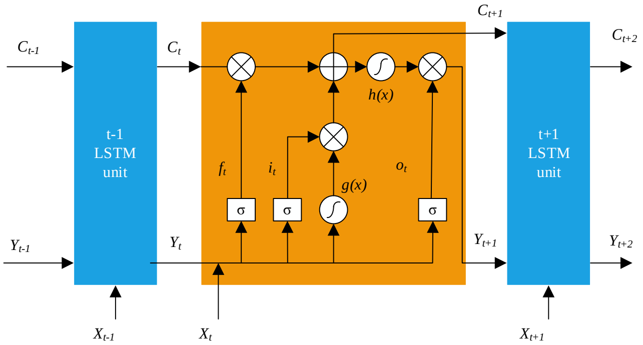


Fig. 2. The time sequence diagram of LSTM

The time sequence diagram of LSTM is shown in Fig. 2. In each epoch of LSTM, the output will affect the gating unit so that the information will not be lost. Because of the existence of the forget gate, LSTM can decide when to discard some unimportant time information, to maintain the updating speed and certain accuracy of the network in the long time series.

The dissolved gas data in transformer oil is typical time series data. The data at a certain moment is closely related to the data at the previous moment, and the general neural network cannot find this connection, thus limiting the accuracy of the model prediction. LSTM can store historical data at a certain time scale due to the existence of memory units. This structure gives LSTM great advantages in processing time series feature data.

IV. Bidirectional Recurrent Neural Network based on Multi-Kernel Learning Support Vector Machine Diagnosis Model

Generally, the recurrent neural network uses Softmax function to classify, but classification effect depends on the data analysis of recurrent neural network and the ability of generalization is bad. For solving this problem, we use SVM to replace the Softmax function in a recurrent neural network. Dissolved gas in transformer oil analysis by recurrent neural network, there are often some heterogeneous data located at the edge of classification space. For single-kernel support vector machines, it is difficult to process heterogeneous data, and they often ignore the heterogeneous data. For making more effective use of the processed heterogeneous data, MKL-SVM is produced by combining multiple kernel functions. The bidirectional recurrent neural network model based on a multi-kernel learning support vector machine is composed of two parts. Feature extraction based on a bidirectional recurrent neural network is an important part. The classifier of the second part uses MKL-SVM.

A. Data Preprocessing

According to the selection of H_2, CH_4, C_2H_4, C_2H_2, and C_2H_6, the five characteristics gases, and the ratio of the total volume input of the model, the data normalization processing is carried out according to the (6).

$$x_{\mathrm{m}}^{(i)} = \frac{x^i}{x^1 + x^2 + x^3 + x^4 + x^5},\; i = 1,2,3,4,5 \qquad (6)$$

In the above equation, $x_{\mathrm{m}}^{(i)}$ is normalized. $x^i\,(i=1,2,3,4,5)$ is the volumes of H_2, CH_4, C_2H_4, C_2H_2, and C_2H_6 before normalization respectively.

B. Training Set Selection

For the volume fraction set of dissolved gas in oil, the K-fold cross-validation method is adopted to partition the data set into training and test two parts. The cross-validation times are 5. The training set is responsible for training model, and the accuracy is evaluated through the test set, which can avoid the phenomenon of over-fitting. For large data sets, a relatively small training set can meet the requirements of model training. For small data set, a relatively large training set is needed to train model. In this experiment, the size of data set is small, so training set and test set account for 80% and 20% each.

Based on the study of various representative characteristic gases of transformer faults in Table I, combined with the faults frequently encountered in actual production, a variety of faults are summarized according to the characteristics of faults to deal with faults faster in actual operation and optimized classification ability. And the operation states of transformers are classified into six categories: (1) Normal, (2) partial discharge, (3) low energy discharge, (4) high energy discharge, (5) medium-low temperature overheating, and (6) high temperature overheating.

C. Bidirectional Circulation Structure

Generally, the data analysis of recurrent neural networks depends on time series. Recurrent neural network need the data of the previous moment to predict the change of future data. When data has a temporal relationship or the data is more time-dependent, such network structure will often produce large errors. In addition, the training of this kind of neural network requires large-scale training samples. If there are few elements in the training sample, it may be difficult to obtain an ideal model.

To solve these problems, a bidirectional recurrent neural network is adopted in this paper. On the basis of original recurrent neural network model, a reverse sequential recurrent neural network is added. Two recurrent neural networks share input and output. Fig. 3 shows the structure of the bidirectional recurrent neural network.

As Fig. 3 show, the recurrent neural network has two hidden layers. The data is simultaneously input into the forward and reverse timing sequence. W2 is the forward timing sequence, and W5 is the reverse timing sequence. W1 enters the forward timing sequence, and W3 enters the reverse timing sequence. W4 and W6 jointly constitute the output layer. Different from the general neural network, the forward cycle layer and reverse cycle layer in the recurrent neural network are not connected, which can effectively prevent the occurrence of the self-cycle phenomenon.

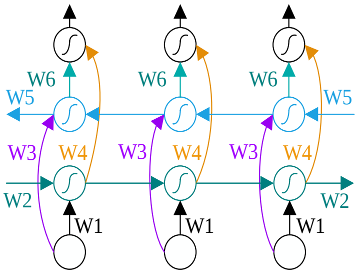


Fig. 3. The structure of the bidirectional recurrent neural network

D. Support Vector Machine

The recurrent neural network uses the Softmax function in fault classification process, but this function cannot perform well in a wide range of faults. When the Softmax function is multi-classified, its effect is similar to the Sigmoid function used for Logistic regression. Only when various types are mutually exclusive, can the Softmax function have good classification ability [18]. SVM has a good classification effect and generalization ability in the classification task and it is a common fault diagnosis method to combine data analysis and classification with a recurrent neural network. The LSTM-SVM model obtained by combining the LSTM network with the SVM classifier can simultaneously give full play to LSTM's ability to process long sequence information and SVM's ability to classify low-dimensional feature data, to reduce modeling difficulty and computational complexity and improve the speed and accuracy of performance degradation prediction [19].

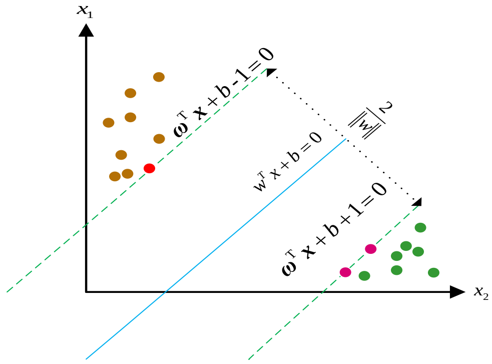

x_1
x_2
$\boldsymbol{\omega}^{\mathrm{T}}\mathbf{x}+\mathrm{b}-1=0$
$\mathrm{w}^{\mathrm{T}}\mathrm{x}+\mathrm{b}=0$
$\boldsymbol{\omega}^{\mathrm{T}}\mathbf{x}+\mathrm{b}+1=0$
$2/\|w\|$

Fig. 4. The diagram of the hyperplane dividing two types of samples

The central theme of the SVM is classification learning. For a given training sample set, a classification hyperplane is found in the sample space as the decision boundary to separate samples of different categories [20]. Fig. 4 shows the diagram of the hyperplane dividing two types of samples.

The hyperplane described by the following linear equation:

$$0=\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}+b \tag{7}$$

Sample points in the sample space satisfy the following equation:

$$y=\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}+b \tag{8}$$

In the above equation, is the weight of the input quantity, b is the bias of the value, is the data set of the input, is the classification hyperplane of the partition class.

The distance that sample arrive the hyperplane in the sample space expressed as follows:

$$r=\frac{|\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}+b|}{\|\boldsymbol{\omega}\|} \tag{9}$$

The nearest sample that arrive the hyperplane in the sample space is called ‘support vector’, and the sum of distance between the support vectors of two classes and the hyperplane is as follows:

$$\gamma=\frac{2}{\|\boldsymbol{\omega}\|} \tag{10}$$

The basic form of a support vector machine can be obtained by maximizing the geometric interval between the data set and hyperplane.

$$\begin{aligned}&\min_{\omega,b}\frac{1}{2}\|\boldsymbol{\omega}\|^2\\&s.t.\ y_i(\boldsymbol{\omega}^T\boldsymbol{x}_i+b)\geq 1,\ i=1,2,\ldots,m.\end{aligned} \tag{11}$$

From (11), the solution of SVM is a convex quadratic programming problem, duality problem got by using the Lagrangian multiplier method. The Lagrangian function of the dual problem can be expressed by the following formula.

$$L(\boldsymbol{\omega},b,\boldsymbol{\alpha})=\frac{1}{2}\|\boldsymbol{\omega}\|^2+\sum_{i=1}^{m}\alpha_i(1-y_i(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}_i+b)) \tag{12}$$

Here, $\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\ldots,\alpha_{\mathrm{m}})$. The partial derivatives of the Lagrangian function L are taken with respect to $\boldsymbol{\omega}$ and $\boldsymbol{b}$, and they are set to zero respectively to get the optimal solution of the function.

$$\boldsymbol{\omega}=\sum_{i=1}^{m}\alpha_i y_i\boldsymbol{x}_i \tag{13}$$

$$0=\sum_{i=1}^{m}\alpha_i y_i \tag{14}$$

According to (11), the dual problem is obtained by substituting (13) and (14) into (12):

$$\begin{aligned}&\max_{\alpha}\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j\boldsymbol{x}_i\boldsymbol{x}_j\\&s.t.\ \sum_{i=1}^{m}\alpha_i y_i=0,\ \alpha_i\geq 0,\ i=1,2,\ldots,m\end{aligned} \tag{15}$$

According to (8), (13), and (14), the model is obtained.

$$f(\boldsymbol{x})=\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}+b=\sum_{i=1}^{m}\alpha_i y_i\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}+b \tag{16}$$

It can be seen from (11) that there are inequality constraints on α_i and training samples $(\boldsymbol{x_i},\ y_i)$ in support vector machines.

$$\begin{cases}\alpha_i\geq 0\\ y_i f(\boldsymbol{x}_i)-1\geq 0\\ \alpha_i(y_i f(\boldsymbol{x}_i)-1)=0\end{cases} \tag{17}$$

The above constraints stipulate that the training samples always have $\alpha_i=0$ or $y_i f(\boldsymbol{x}_i)=1$.

Thinking about nonlinear case, the hyperplane model in the characteristic space is calculated below:

$$f(\boldsymbol{x})=\boldsymbol{\omega}^{\mathrm{T}}\varphi(\boldsymbol{x})+b \tag{18}$$

From (18), $\varphi(\boldsymbol{x})$ is the expression generated after the two-dimensional sample $\boldsymbol{x}$ to a higher dimensional feature space. According to (12), (13), and (14), a new dual problem is as follows:

$$\begin{aligned}&\max_{\alpha}\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j\varphi(\boldsymbol{x}_i)\varphi(\boldsymbol{x}_j)\\&s.t.\ \sum_{i=1}^{m}\alpha_i y_i=0,\ \alpha_i\geq 0,\ i=1,2,\ldots,m\end{aligned} \tag{19}$$

It is necessary to find a $K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\varphi(\boldsymbol{x}_i)\varphi(\boldsymbol{x}_j)$ function, namely the kernel function of SVM, which generally uses linear kernel, Gaussian function kernel, polynomial kernel, and sigmoid kernel.

E. Multi-Kernel Learning

Generally, single-kernel structures are used in support vector machines, which are classified based on a single feature space. The selection of kernel functions needs to be judged according to actual needs and then selected according to experiences. And different parameters are set. Such kernel function design is not convenient, and SVM is difficult to train the specific data in the training sample, resulting in lower accuracy of the classifier.

For solving above flaws of SVM, this paper adopts the combination of multiple kernel functions to establish the MKL-SVM model and improves the adaptability of support vector machine to complex data through multi-kernel learning. MKL integrates multiple sub-kernels into a unified optimization framework to seek the best combination. Using the multi-kernel model optimize the performance of the learning model and obtain explicable decision functions [21]. With the adoption of the multi-kernel learning method, the classification effect of a single support vector machine on a few specific data is changed, and the classification accuracy of fault data is improved.

The kernel function in multi-kernel learning is composed of several basic kernel functions.

$$K^{'}(x_i,x_j)=\sum_{i=1}^{m}\lambda_i K_i(x_i,x_j),\lambda_i \ge 0;\sum_{i=1}^{m}\lambda_i=1 \qquad (20)$$

From (20), m is the number of kernels, and λ_i is the weight of each kernel. Before calculating the weight of each kernel function, we need to obtain the kernel matrix of each kernel function.

$$\boldsymbol{K}_i=\begin{bmatrix} k(x_1,x_1) & \dots & k(x_1,x_m) \\ \vdots & \ddots & \vdots \\ k(x_m,x_1) & \cdots & k(x_m,x_m) \end{bmatrix},i=1,2,\dots,m \qquad (21)$$

After the kernel matrix is obtained, the eigenmatrix ***M*** is calculated according to the eigenvectors that get by data set. Getting the weight of each kernel function, the trace of each matrix tr(***K***) and tr(***M***) is used to characterize the characteristics of each matrix. And then the Euclidean distance of each tr(***K***) and tr(***M***) is calculated as follows.

$$|X_i|=\sqrt{tr(K_i)+tr(M)},i=1,2,\dots,m \qquad (22)$$

Then the kernel matrix is substituted into the feature matrix to obtain the importance of each kernel matrix to the feature.

$$H_i=\frac{\boldsymbol{K}_i\boldsymbol{M}}{|X_i|} \qquad (23)$$

Finally, the weights of each kernel function are obtained.

$$\lambda_i=\frac{H_i}{\sum_{i=1}^{m}H_i} \qquad (24)$$

F. Calculation Steps of the Fault Diagnosis Method

The flow chart of Bidirectional Long Short-Term Memory network based on Multi-Kernel Learning Support Vector Machine (Bi-LSTM-MKL-SVM) is shown in Fig. 5.

Sum the output values at the current time to get $\boldsymbol{Y}_t$.

$$\boldsymbol{Y}_t=\boldsymbol{Y}_t^1+\boldsymbol{Y}_t^2 \qquad (25)$$

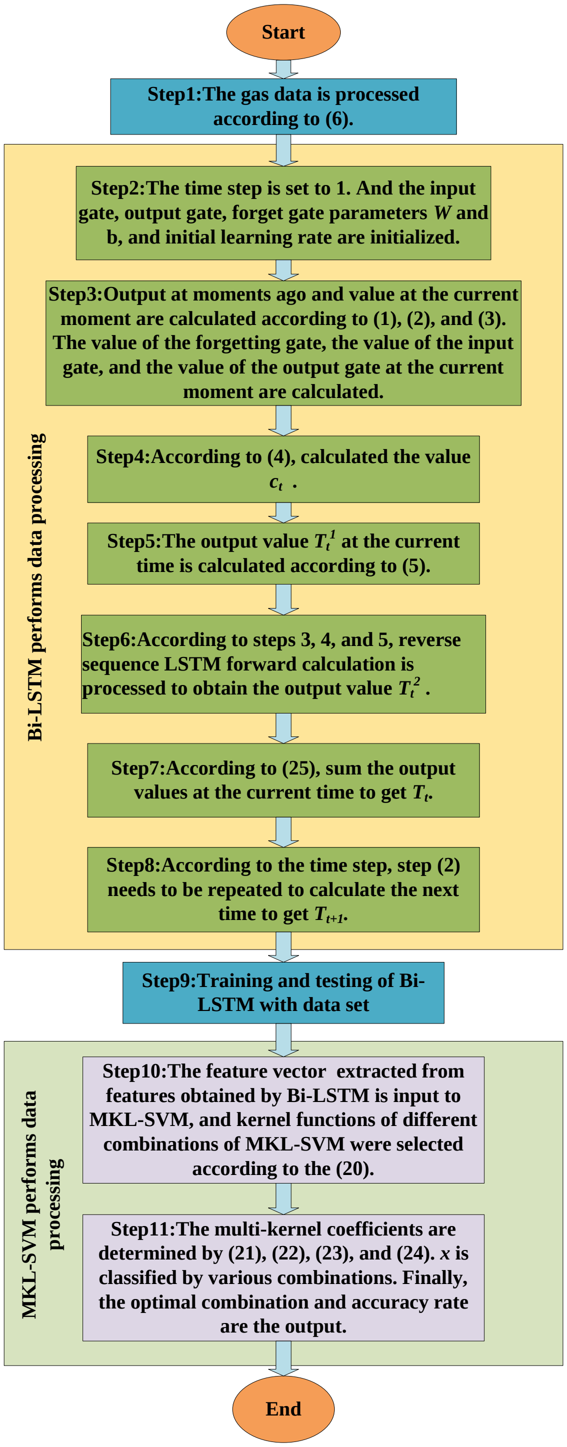

Start
Step1:The gas data is processed according to (6).
Bi-LSTM performs data processing
Step2:The time step is set to 1. And the input gate, output gate, forget gate parameters *W* and b, and initial learning rate are initialized.
Step3:Output at moments ago and value at the current moment are calculated according to (1), (2), and (3). The value of the forgetting gate, the value of the input gate, and the value of the output gate at the current moment are calculated.
Step4:According to (4), calculated the value c_t .
Step5:The output value T_t^1 at the current time is calculated according to (5).
Step6:According to steps 3, 4, and 5, reverse sequence LSTM forward calculation is processed to obtain the output value T_t^2 .
Step7:According to (25), sum the output values at the current time to get T_t.
Step8:According to the time step, step (2) needs to be repeated to calculate the next time to get T_{t+1}.
Step9:Training and testing of Bi-LSTM with data set
MKL-SVM performs data processing
Step10:The feature vector extracted from features obtained by Bi-LSTM is input to MKL-SVM, and kernel functions of different combinations of MKL-SVM were selected according to the (20).
Step11:The multi-kernel coefficients are determined by (21), (22), (23), and (24). *x* is classified by various combinations. Finally, the optimal combination and accuracy rate are the output.
End

Fig. 5. The flow chart of Bi-LSTM-MKL-SVM

V. Experimental Analysis

A. Selection of Training Times

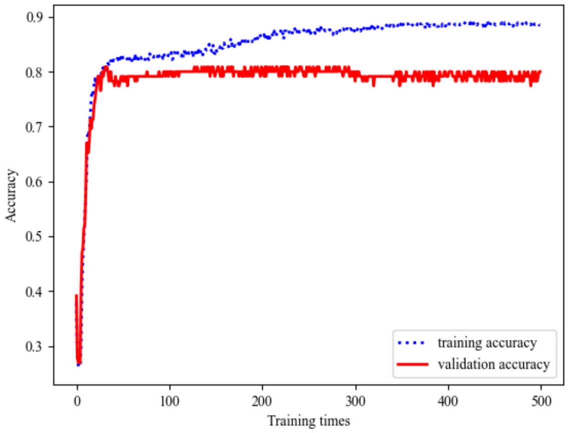


Fig. 6. The accuracy of the training set and test set of Bi-LSTM

The accuracy of the training set and test set of Bi-LSTM is shown in Fig. 6. After 300 training times, the accuracy of the test set decreases obviously, and the model training has an overfitting phenomenon. Thus the number of sessions is set to 300 for the best results.

B. Feature Extraction Performance Analysis of Time Series Data

In Table II, the model related parameters for using method in this paper are presented. It mainly includes the model mentioned in the paper and the parameters used in the model.

TABLE II. Description of Model Related Parameters

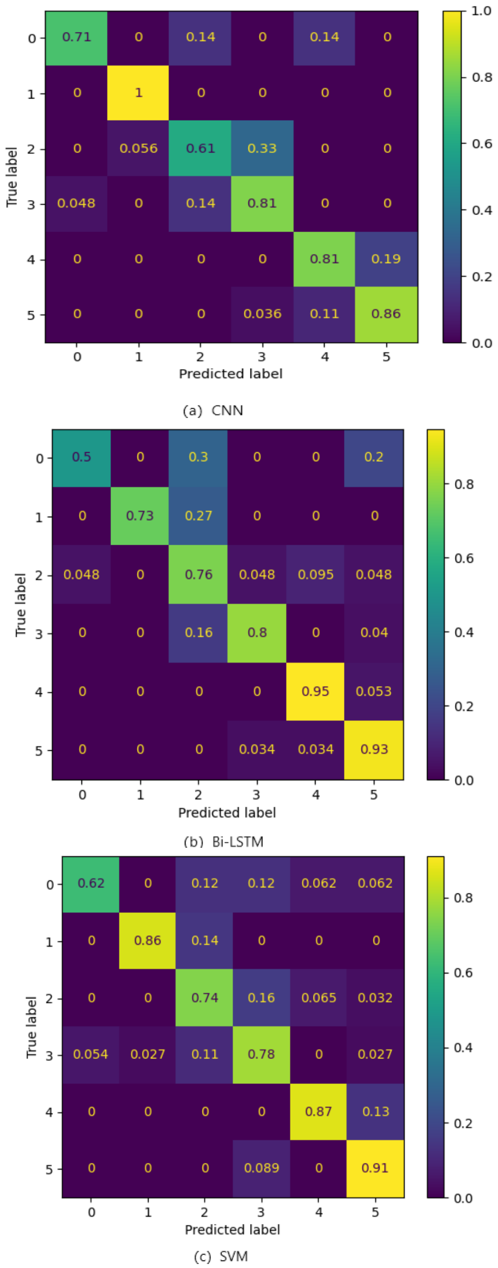
Abbreviation	Description
Bi-LSTM	It represents an LSTM with an additional reverse sequence
Bi-LSTM-SVM	Bi-LSTM is connected with SVM, and SVM is used to replace the classification function in Bi-LSTM
MKL-SVM	SVM is used by stacking multiple kernel functions according to a certain weight
Bi-LSTM- MKL-SVM	Bi-LSTM is connected with MKL-SVM, and MKL-SVM is used to replace the classification function in Bi-LSTM
Time step	The step size of each computation in LSTM
forget gate	It is responsible for controlling the persistence of memory cell
input gate	It is responsible for controlling the input of the immediate state into memory cell
output gate	It is responsible for controlling whether the current value is printed
Memory cell	It is responsible for storing long-term state

In order to validate the bidirectional recurrent neural network to improve the effectiveness of time-series analysis, training set input Bi-LSTM respectively with LSTM, RNN, and CNN model and then validated the accuracy of test set. The results of four kinds of models of time-series data feature extraction accuracy are shown in Table III.

TABLE III. The Results of Four Kinds of Models of Time-Series Data Feature Extraction Accuracy

Table No	Models	Accuracy (%)
1	CNN	80.18
2	RNN	81.34
3	LSTM	82.83
4	Bi-LSTM	84.08

To ensure no interference to the sequential feature extraction ability of the models, the same Softmax function was used as the classifier for the four models. According to the data in Table III, RNN has higher accuracy than CNN in time sequence data because of its memorability. Compared with RNN, the LSTM network has higher accuracy in data analysis due to the gated unit, while RNN has the problem of gradient explosion due to the rugged search path. So pruning operation is required to decrease learning rate and increase training time. Among above recurrent neural networks, Bi-LSTM has the highest accuracy, which is more than 84%. In the process of model training, Bi-LSTM model not only speeds up the convergence rate of the model but also deepens the time sequence at the same time for the correlation of each parameter. It also corrects the training model, so that the model has better accuracy in prediction. The volatility of its accuracy is less than other recurrent neural networks and CNN. The experiment proves that the Bi-LSTM network can better complete the feature extraction task in time series data by adding a reverse time series recurrent neural network. MKL-SVM has a better classification effect on heterogeneous data. By comparing the confusion matrix between CNN and Bi-LSTM, we verify whether the heterogeneous data generated by Bi-LSTM is more suitable for MKL-SVM on the basis of Bi-LSTM having a better feature extraction ability on time series data. SVM, as a classical classifier, is compared with Bi-LSTM to verify the influence of feature extraction of time series data on data sample classification. The confusion matrix of CNN, Bi-LSTM, and SVM is shown in Fig. 7.



(a) CNN

(b) Bi-LSTM

(c) SVM

Fig. 7. The confusion matrix of CNN, Bi-LSTM, SVM

Both CNN and Bi-LSTM use the Softmax function as model classifier. CNN model only has a good classification effect on partial discharge fault type, but a poor classification effect on other fault types. The Bi-LSTM model has a good classification effect on multiple fault types. According to Fig. 7, the distribution range of Bi-LSTM classification results is wider, indicating that Bi-LSTM will generate more heterogeneous data in feature extraction of data and affect the classification effect of the classifier. SVM and Bi-LSTM have similar fault diagnosis effects, but SVM has more misjudgment of fault types and the classification effect is not stable, indicating that the data extracted by feature can be more easily classified. The confusion matrix shows the classification performance of Bi-LSTM is significantly better than CNN and SVM. After feature extraction, the Bi-LSTM time series data generates more heterogeneous data and has a better combination with MKL-SVM.

C. *Diagnosis Result Analysis of Bidirectional Recurrent Neural Network Model based on a Multi-Kernel Support Vector Machine*

In this section, MKL-SVM model will be established through kernel fusion to form a multi-kernel. In multi-kernel learning, multiple kernel functions can be used for fusion, or the same kernel function with different parameters can be used for fusion to ensure the diversity of kernel fusion. Fig. 8 shows the workflow of MKL-SVM.

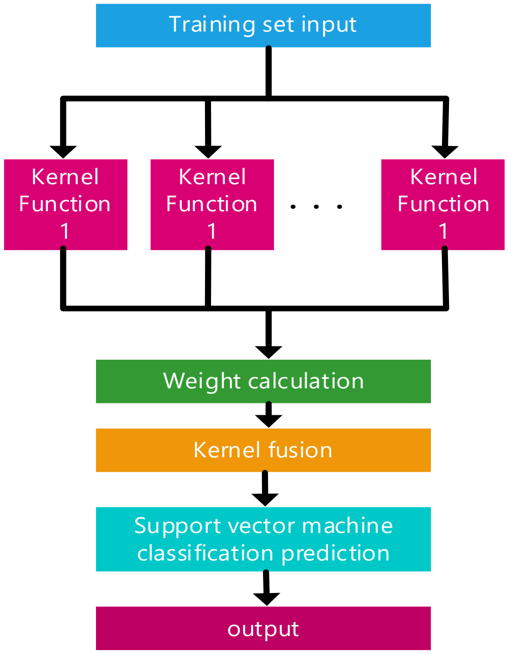


Fig. 8. The workflow of MKL-SVM

To prove the effectiveness of Bi-LSTM-MKL-SVM network in improving the exactitude of time series analysis and prediction, SVM and Bidirectional Long Short-Term Memory network based on Support Vector Machines (Bi-LSTM-SVM), LSTM and Bi-LSTM and Bi-LSTM-MKL-SVM are used to compare and analyze the predicted results of test sets.

The micro-average PR curve of the five models is shown in Fig. 9. On the whole, under the influence of heterogeneous data, the classification performance of Bi-LSTM-SVM is not different from Bi-LSTM and LSTM. Bi-LSTM-MKL-SVM has better classification performance, while the SVM has the worst classification effect.

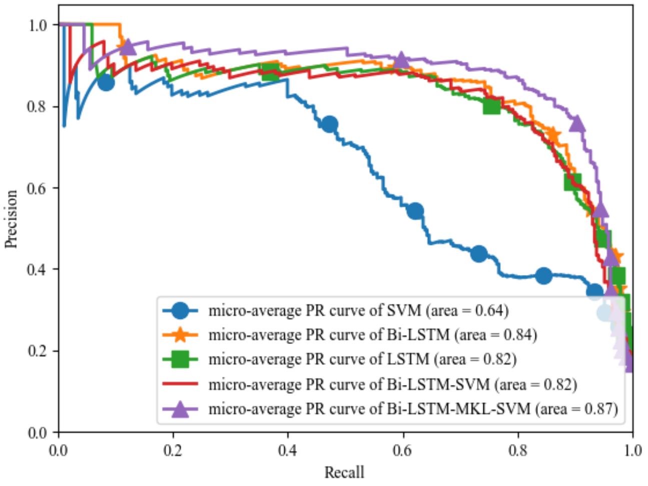


Fig. 9. The micro-average PR curve of the five models

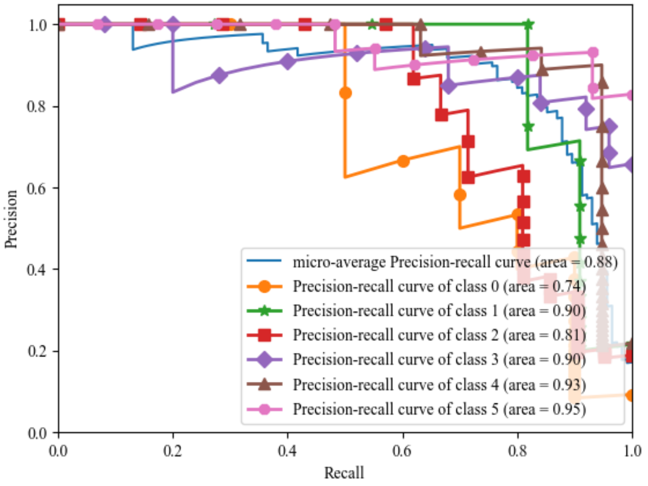
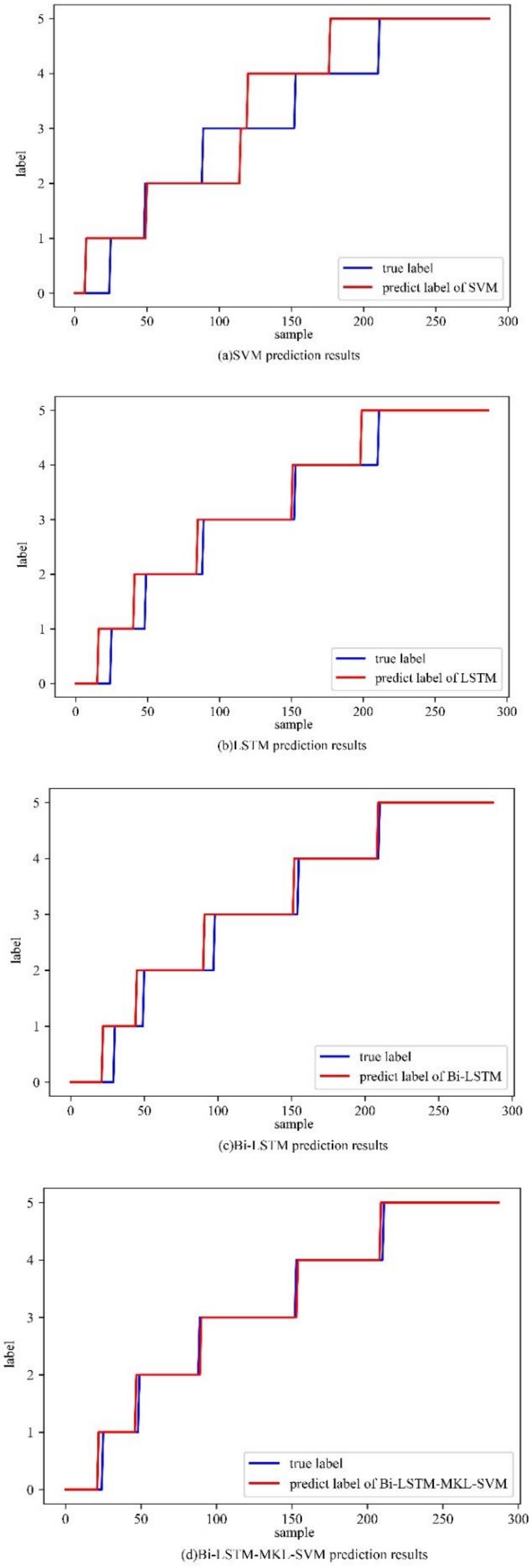


Fig. 10. The PR curve of the Bi-LSTM-MKL-SVM model

The PR curve of the Bi-LSTM-MKL-SVM model is shown in Fig. 10. Bi-LSTM-MKL-SVM has good classification performance for each fault category under good overall classification performance. Compared with the SVM model, the combined model that extract time series information by the recurrent neural network can effectively improve the classification efficiency of the classifier. For the LSTM model, SVM classification can significantly optimize neural network performance. Feature extraction of time series data by Bi-LSTM will generate a lot of heterogeneous data. Using general SVM can not only improve the performance of the classifier but also cause interference with the classification ability of SVM. MKL-SVM is more effective for the classification of heterogeneous data. Bi-LSTM-MKL-SVM is used to solve the influence of heterogeneous data on model judgment better, improve the model's utilization ability of time-series data, and enhance the overall generalization ability of model, as well as improves its stability.



(a)SVM prediction results

(b)LSTM prediction results

(c)Bi-LSTM prediction results

(d)Bi-LSTM-MKL-SVM prediction results

Fig. 11. The comparison between the true label and predicted label

Fig. 11 shows the comparison between the true label and the predicted label. The data fitting and classification effect are shown in Fig. 11; the SVM model has a poor fitting effect on test samples. This shows that simple classification algorithm is difficult to complete accurate classification of complex data. The Bi-LSTM model has better fitting accuracy than the LSTM model. This indicates that adding a reverse timing network can effectively improve the accuracy of LSTM. Compared with the other three models, the Bi-LSTM-MKL-SVM model has the best classification and fitting effect on test samples, indicating that the Bi-LSTM-MKL-SVM model has a stronger ability to extract time series features. It proves that these optimizations are effective. SVM can significantly improve the generalization of neural networks. Multi-kernel learning can effectively optimize the classification effect of SVM on complex data.

TABLE IV. THE ACCURACY OF THE FIVE MODELS

Table No.	Models	Accuracy (%)
1	SVM	81.50
2	LSTM	82.83
3	Bi-LSTM-SVM	83.58
4	Bi-LSTM	84.08
5	Bi-LSTM-MKL-SVM	85.86

The accuracy of the five models is shown in Table IV. The Bi-LSTM network model is used to process time series data, which improves the classification efficiency of MKL-SVM, improves the accuracy and stability of the model as a whole, and effectively solves the shortcomings of neural networks and SVM used alone. It can be seen by comparing with the model prediction accuracy, MKL-SVM does better in heterogeneous data for more accurate judgment, and it can save time for SVM to adjust the parameter and reduce the possibility of error when choosing parameters based on experience. More basic kernel function combinations optimize the effect of MKL-SVM and its applicability.

VI. CONCLUSION

In allusion to transformer fault diagnosis, the deep learning method has problems with the stability and the effective use of data samples. This paper comes up with a kind of model called Bi-LSTM-MKL-SVM and takes Bi-LSTM as the feature extraction part of model. MKL-SVM is used to replace the Softmax function in Bi-LSTM to classify the data after feature extraction. By adding reverse time sequence, the Bi-LSTM-MKL-SVM network model can consider more time sequence factors, achieve better prediction effects for time sequence data, and carry out more effective feature extraction for time sequence data. It has better classification ability for heterogeneous data and reduces the time used for reference adjustment depending on experience. The feature extraction ability and generalization ability of the Bi-LSTM-MKL-SVM network model were verified by the confusion matrix, PR curve, and accuracy comparison. The accuracy of the Bi-LSTM-MKL-SVM network model was higher. The Bi-LSTM-MKL-SVM network model can improve the accuracy of transformer fault diagnosis and reduce the misjudgment of faults in the operation of transformers, thus reducing the economic losses caused by transformer faults.

The model proposed in this paper has certain limitations on the analysis of gas data. Different fault types pay different attention to gas data. More accurate selection of the weight of each gas in different faults can effectively improve the accuracy of classification. Transformer failure is closely related to environmental factors, such as ambient temperature and humidity. Considering more influencing factors to improve the universality of the model is one of the future research directions.

ACKNOWLEDGMENT

This work was supported by the Scientific Research Project of the Liaoning Provincial Department of Education (No. LJKZ0398) and the Scientific Research Foundation of Liaoning Petrochemical University (No. 2017XJJ-044). The authors also would thank reviewers for improving this article.

REFERENCES

[1] Y. H. Wang, and Z. Z. Wang, “Transformer fault identification method based on RFRFE and ISSA-XG Boost,” Journal of Electronic Measurement and Instrumentation, vol: 35, pp. 142-150, 2021.

[2] H. Zhang, and S. X. Zhu, “Application of support vector machine classification in asynchronous motor fault diagnosis,” Journal of Suzhou University of Science and Technology Engineering and Technology edition, vol: 32, pp. 70-74, 2019.

[3] X. Zhang, H. Wang, Y. P. Wei, S. L. Wang, and Y. H. Su, “Fault diagnosis of power transformer based on SSA-PNN,” Industrial Instrumentation & Automation, vol: 1, pp. 86-90, 2022.

[4] Q. C. Fan, F. Yu, and M. Xuan, “Transformer fault diagnosis method based on improved whale optimization algorithm to optimize support vector machine,” Energy Reports, vol: 7, pp. 856-866, 2021.

[5] Y. H. Wu, X. B. Sun, Y. Zhang, X. J. Zhong, and L. Cheng, “A Transformer Fault Diagnosis Based on Improved Squirrel Search Algorithm and Support Vector Machine,” Journal of Physics: Conference Series, vol: 2203, 012067, 2022.

[6] B. Zeng, J. Guo, W. Q. Zhu, Z. H. Xiao, F. Yuan, and S.X. Huang, “A Transformer Fault Diagnosis Model Based On Hybrid Grey Wolf Optimizer and LS-SVM,” Energies, vol: 12, pp. 1-18, 2019.

[7] J. J. Miao, D. B. Li, H. J. Li, and P. Y. Liu, “Research on Furnace Slagging Prediction based on Improved Particle Swarm Optimization and Fuzzy Neural Network,” Journal of Engineering for Thermal Energy and Power, vol: 37, pp. 104-114, 2022.

[8] X. C. Gao, Q. Li, L. L. Wang, G. X. Du, and X. Ke, “Adaptive Convolutional Neural Network Based on Modified Genetic Algorithm,” Computer Technology and Development, vol: 32, pp. 132-136+142, 2022.

[9] I. B. M. Taha, S. Ibrahim, and D. E. A.Mansour, “Power Transformer Fault Diagnosis Based on DGA Using a Convolutional Neural Network with Noise in Measurements,” IEEE ACCESS, vol: 9, pp. 111162-111170, 2021.

[10] X. D. Fan, W. P. Fu, Z. L. Zhao, S. T. Zu, L. S. Zhang, and W. T. Hu, “Study on Fault Diagnosis of Oil-Immersed Transformer Based on Long-Short Term Memory Network,” Transformer, vol: 58, pp. 27-32, 2021.

[11] X. X. Wu, Y. G. He, J. J. Duan, H. Zhang, and Z. R. Zeng, “Bi-LSTM-based transformer fault diagnosis method based on DGA considering complex correlation characteristics of time sequence,” Electric Power Automation Equipment, vol: 40, pp. 184-193, 2020.

[12] O. A. Alharbi, “Deep Learning Approach Combining CNN and Bi-LSTM with SVM Classifier for Arabic Sentiment Analysis,” International Journal of Advanced Computer Science and Applications, vol: 12, pp. 165-172, 2021.

[13] K. Wang, J. Z. Li, and S. Q. Zhang, “New Features Derived from Dissolved Gas Analysis for Fault Diagnosis of Power Transformers,” Proceedings of the CSEE, vol: 36, pp. 6570-6578+6625, 2016.

[14] X.D.Pei, L.Luo, and S. Chen,“A Convolutional Neural Network Diagnosis Method for Dissolved Gas in Power Transformer Oil,”Journal of Liaoning Petrochemical University,vol:40,pp.79-85,2020.

[15] K. Gregor, I. Danihelka, and A. Graves, et al. , “DRAW: a recurrent neural network for image generation,” IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, 2015, pp. 1462-1471.

[16] S. X. Liu, S. Z. Gao, Y. Liu, J. Li, and Y. D. Cao, “Residual Life Prediction of AC Contactor Based on LSTM,” High Voltage Engineering, 2022, pp. 1-11.

[17] X. Chang, Y. B. Li, and M. Tian, et al. , “Reinforcement Learning Algorithm Based on One-dimensional Convolutional Recurrent Network,” Computer Measurement & Control, vol: 30, pp. 258-265, 2022.

[18] N. S. Kiruthika, and G. Thailambal, “Dynamic Light Weight Recommendation System for Social Networking Analysis Using a Hybrid LSTM-SVM Classifier Algorithm,” Optical Memory and Neural Networks, vol: 31, pp. 59-75, 2022.

[19] G. Z. Huang, W. J. Li, and Y. H. Deng, “Modeling of Performance Decay Prediction Based on LSTM-SVM Flexible Circuit Board Processing Roll,” Journal of Physics: Conference Series, vol: 1828, pp. 012027, 2021.

[20] K. Hasegawa, and K. Funatsu, “Non-linear modeling and chemical interpretation with aid of support vector machine and regression,” Current computer-aided drug design, vol: 6, pp. 24-36, 2010.

[21] L. S. Nie, F. Y. Chang, X. Z. Chang, C. Liu, Y. W. Jin, G. S. Liu, J. S. Fu, and X. S. Han, “A Novel Self-adaptive Multiple Kernel Learning Algorithm,” Journal of Jilin University(Science Edition), vol: 59, pp. 1212-1218, 2021.