ISSN 13303651 (Print), ISSN 18486339 (Online)
Received: 11 December 2025; Accepted: 13 February 2026
https://doi.org/10.17559/TV-20251211003197
Original scientific paper

Research on Intelligent Transformer Fault Diagnosis Model Based on Multimodal Data Fusion and Deep Learning

Cunjian TIAN

Abstract: To enhance the accuracy of transformer fault diagnosis, this study is dedicated to designing a hybrid intelligent mechanism for fault diagnosis, which organically integrates multimodal data fusion strategies with adaptive deep learning models. The key information required for diagnosis is derived from the analysis of dissolved gases in the oil and serves as the input feature of the deep model. The key to model training lies in the introduction of an adaptive mechanism that can dynamically calibrate the learning rate based on the real-time convergence trend. An adaptive learning mechanism that can dynamically adjust the learning rate during the iterative process is proposed, thereby enhancing the convergence accuracy of the model while improving its training efficiency. Through specific cases, important parameters such as the number of hidden layers and the learning rate adjustment coefficient in the adaptive deep learning model were determined. The experimental results show that the proposed method performs excellently in feature extraction and analysis, featuring a faster convergence speed and higher convergence accuracy, which can significantly improve the accuracy of transformer fault diagnosis. Aiming at the problem that data alignment and series fusion are often ignored in the traditional multimodal data fusion process, this paper further proposes a graph-text multimodal fusion model based on the cross-attention mechanism. This model first uses BERT and ConvNeXt to extract text and image features respectively. Subsequently, with the help of the attention mechanism in the Image Transformer, the detailed information in the feature map output by ConvNeXt is further extracted to obtain higher-level image features and ensure that the image and text features are consistent in dimension. Finally, the alignment and fusion of graphic and text features are achieved through the cross-attention module. Experiments on the three datasets of MSAW-Single, MSAW-Multiple and MMSD show that the classification accuracy of the image-text multimodal fusion model based on cross-attention reaches 75.21%, 73.15% and 85.85% respectively, verifying the effectiveness of this method.

Keywords: adaptive deep learning model; fault diagnosis; learning rate; multimodal analysis; multimodal data fusion; transformer

1 INTRODUCTION

As a core component of the smart grid, the smart substation undertakes key functions such as the automatic collection, control and protection of the operation information of the equipment within the station. Power transformers are core equipment in substations, and their operational status directly affects the overall stability and availability of the power grid. During long-term operation, mechanical structural faults are prone to occur in components such as transformer windings and cores, thereby affecting their operational reliability. Therefore, implementing condition monitoring on transformers and promoting research on intelligent fault diagnosis technology have become key links in achieving the operational goals of smart substations [1].

The implementation of shutdown monitoring first requires the transformer to be taken out of its normal operating state, and then fault detection can be carried out on it [2]. This will cause local power supply interruption and is difficult to meet the development requirements of continuous operation of the smart grid. In online monitoring technology, dissolved gas analysis [3] (DGA) and frequency response analysis [4] (FRA) in oil are commonly used methods. Although the application scope of DGA technology can cover abnormal conditions such as discharge, moisture and overheating inside transformers, it is difficult to provide direct diagnostic basis for mechanical faults such as winding deformation and loose core. Although the FRA method can accurately determine the type and degree of faults, it needs to be connected to the power system and is vulnerable to electromagnetic interference. In recent years, vibration signal analysis technology has been gradually applied to mechanical fault diagnosis [5]. By installing acceleration sensors on the transformer casing, vibration signals are collected in a wired manner to achieve fault identification [6]. The advantage of this method lies in its simplicity and high sensitivity, but it has problems such as a complex measurement system, high cost, and the need for regular maintenance during long-term monitoring. By integrating passive RFID tags with sensing units, the passive RFID sensing tags formed have been actually deployed in engineering practice as a new type of measurement tool [7]. The technical foundation, namely Radio Frequency Identification (RFID), is an advanced means of automatic data collection and identification. This type of tag operates based on the backscattering mechanism and features a simple structure, ease of use and low power consumption. It can run under passive conditions and is particularly suitable for long-term outdoor condition monitoring.

This paper introduces the theory of deep learning into the transformer fault diagnosis task and proposes an adaptive deep learning model to accelerate the convergence speed of the model and thereby improve the diagnostic efficiency. This method builds a deep learning structure with multiple hidden layers based on a deep belief network and dynamically adjusts the learning rate according to the training process, thereby accelerating model convergence, reducing the interference of initial parameters on the modeling effect, and improving fault diagnosis performance. Verified through specific examples, the results show that the proposed method not only improves the accuracy of fault identification but also significantly accelerates the convergence speed of the model.

To further enhance the accuracy of data analysis and overcome the limitations of single-modal data analysis, researchers are gradually turning to the path of multimodal data analysis. The key to multimodal analysis lies in effectively extracting and integrating the features of heterogeneous modal data, promoting information complementarity among different modalities, and enhancing the representation ability of shared labels. With the advancement and application of deep learning technology, the limitations of traditional manual feature

engineering [7] have been broken through, making modal features more abundant and high-dimensional, thereby enhancing the accuracy of multimodal analysis. The combination of multimodal data and deep learning has strongly promoted the development of human-computer interaction in data analysis. This paper will also conduct research on the current problems such as shallow fusion levels and insufficient analysis accuracy in multimodal fusion analysis, aiming to construct a multimodal fusion analysis model with deeper fusion and higher analysis accuracy. The first part of the article is the introduction, the second part is related work, the third part is the transformer fault adaptive deep learning diagnosis model based on multimodal data fusion, the fourth part is simulation verification, and the fifth part is conclusion.

2 RELATED WORK

Transformer fault diagnosis mainly covers the mechanical structure fault identification of windings and cores, and its methods usually include two key links: feature extraction and classifier construction. Feature extraction refers to the process of extracting effective features that can reflect the operating status of a transformer from the measured vibration signals. At present, the commonly used feature extraction methods in research include multimodal fusion and deep learning analysis [8], S-transform [9], Hilbert-Huang transform [10], fractal analysis based on multimodal fusion and deep learning [11], as well as entropy and kurtosis in signal statistics [12], etc. Among them, multimodal fusion and deep learning analysis can effectively capture the local characteristics of signals, but the selection of their basis functions is rather difficult. In the time-frequency representation of the S transform, the phase of each frequency component remains consistent with the original signal. The drawback is that the analysis window is fixed and not adjustable. It has good adaptability, but there are problems of non-orthogonality and insufficient stability. Fractal analysis based on multimodal fusion and deep learning has advantages in graphic complexity measurement, but it still relies on the setting of basis functions. Entropy and kurtosis in signal statistics are suitable for describing the uncertainty and distribution characteristics of signals, but their extraction effect is poor under noise interference.

According to statistics from relevant international industry organizations, winding faults account for more than 25% of all types of transformer faults, and over half of them are mechanical faults [10]. For instance, insulating pads made of cellulose paper are usually installed between the layers of transformer windings to mitigate the impact of axial short-circuit force and enhance the axial preload. However, as the operating time increases and continuous pressure is applied, the mechanical properties of the insulating pads will gradually decline, leading to permanent deformation and subsequently weakening the axial preload [11]. If an external short-circuit impact occurs at this time, a gap may be formed between the insulating pad and the winding due to the axial short-circuit force, causing the pad to shift or even part of the coil to overturn, which seriously affects the stable operation of the transformer. Therefore, conducting relevant fault diagnosis research has significant practical significance.

At present, common detection methods for mechanical faults in transformer windings include short-circuit impedance method, low-voltage pulse method and frequency response analysis method, etc. [12, 13] The detection of mechanical faults in the core mainly relies on methods such as oil chromatography analysis and insulation resistance testing. Most of these methods belong to offline detection, which requires power-off operation, has a long maintenance cycle and high cost [14-16]. Vibration analysis, as a non-electrical quantity detection method, can identify winding deformation and changes in the clamping force of the core and windings after analysis. It has the advantages of no direct electrical connection with the power system, high safety and fast detection speed. In terms of vibration modeling, Garcia and others from Carlos III University of Madrid established a transformer vibration model that takes into account the influence of temperature and ignores partial discharge [16]. Another study adopted the finite element method to model and analyze the vibration characteristics of transformers [17]. Based on the established vibration signal mixing model, some scholars have proposed the TIFROM-BSS separation algorithm suitable for non-independent signal analysis [18]. In terms of fault discrimination, existing methods include directly analyzing the spectral characteristics of vibration signals [19, 20], integrating multiple vibration signal analysis techniques for state judgment [21, 22], adopting multimodal fusion and deep learning package energy entropy methods [23], constructing discrimination models based on feature matrices and their similarity [24], improve the construction method of the vibration signal feature matrix [2], etc.

Intelligent fault diagnosis technology integrates traditional fault diagnosis with artificial intelligence methods. It can mine potential patterns based on real-time and historical status data collected by sensors and provide diagnostic strategies with statistical scientific basis through intelligent algorithms. This technology can still maintain good classification performance and high accuracy when dealing with complex and nonlinear diagnostic problems. Support Vector Machine (SVM) is one of the typical intelligent diagnostic algorithms [4]. As a relatively new machine learning method, SVM essentially belongs to a binary classification model. Its core idea is to combine the principles of structural risk minimization and maximum margin, aiming to find a hyperplane (or hypersurface) that can effectively separate the two types of data and maximize the minimum distance from the data on both sides to this surface. SVM requires a relatively small sample size during the training process. The resulting classifier has good generalization ability and can effectively handle nonlinear classification problems, demonstrating significant advantages.

In cross-modal fusion research, to obtain the feature representation after fusion, the direct concatenation method is one of the earliest widely adopted strategies. For instance, some studies have utilized LSTM (Long Short-Term Memory) to extract text features from Glove word embedding vectors, while simultaneously employing pre-trained CNNS (Connectionless Node Network Service) to extract image features. Ultimately, the two types of features are concatenated as a fusion representation [25].

Another scholar has proposed a goal-oriented multimodal fusion model. By extracting lexical and text features respectively through BERT and Bi-LSTM, and then combining the image features extracted by the residual network (ResNet) [26], the three are linearly fused to complete the sentiment classification task. For this reason, some studies have introduced tensor fusion methods, such as Tensor fusion networks (TFN), which achieve fusion through the Cartesian product of different modal features, thereby dynamically learning the relationships within and between modalities. With the development of the attention mechanism, multimodal fusion methods have gradually expanded. Some people hold the view that in practical applications, the representational ability of images is usually weaker than that of text. Therefore, images are used as auxiliary information for text. Through the text-to-image attention mechanism, image features are integrated into text representation, and then combined with self-attention to extract context information for sentiment classification. In social media scenarios, there are often one-to-many image-text relationships, and the VistaNet model emerged as a result. This model is text-driven, with visual information used for alignment and the attention mechanism employed to locate key sentences in the text, thereby completing sentiment analysis. In addition, to address the common situation of embedded text in images, some studies have also proposed a comprehensive framework method that integrates text and image analysis.

3 TRANSFORMER FAULT ADAPTIVE DEEP LEARNING DIAGNOSIS MODEL BASED ON MULTIMODAL DATA FUSION

3.1 Transformer Fault Diagnosis Input Quantity Selection and Fault Status Coding

Taking characteristic gases such as hydrogen, methane, ethane, ethylene and acetylene as the objects of analysis is an effective means for diagnosing transformer faults. The principle lies in the fact that different operating conditions inside the transformer (such as discharge and overheating) will trigger the cracking of insulating oil, which in turn generates dissolved gases of various compositions and contents, providing a direct basis for fault diagnosis. This paper selects the above-mentioned gas as the input quantity for transformer fault diagnosis.

To construct a diagnostic model, the study selected several common types of faults, including low-energy discharge, high-energy discharge, and overheating from low temperature to high temperature (including both discharge and overheating), as the objects of analysis. A key step in data processing is to digitally map these fault categories using coding methods. The detailed corresponding codes are shown in Tab. 1.

Table 1 Transformer fault state coding

Fault type	Fault coding
Normal	(1，0，0，0，0，0，0)
Low-energy discharge	(0，1，0，0，0，0，0)
High-energy discharge	(0，0，1，0，0，0，0)
Low-temperature overheating	(0，0，0，0，1，0，0)
Medium-temperature overheating	(0，0，0，0，1，0，0)
High temperature overheating	(0，0，0，0，0，1，0)
Discharge and overheating	(0，0，0，0，0，0，1)

1) Select the characteristic gas data under typical fault states of the transformer as the training sample set and encode each fault state of the transformer.

2) Taking the characteristic gas as the input quantity, set the number of hidden layers of the DBN (Deep Belief Network), the initial weight of the DBN, the initial bias vector, the initial learning rate and other parameters, and determine the number of iterations in the DBN parameter training process. Based on the training sample set, the DBN parameters are trained and learned to establish an adaptive deep learning model for transformer fault diagnosis.

3) Taking the characteristic gas of the test sample as the input quantity of the adaptive deep learning model for transformer fault diagnosis, the output quantity is obtained, and the transformer fault diagnosis result is derived based on the transformer fault state coding table.

3.2 Adaptive Deep Learning Model

When constructing a deep learning model based on the DBN theory, the learning rate ε needs to be set during the RBM parameter training. ε has a significant impact on the learning performance of RBM (Restricted Boltzmann Machine) parameters. Increasing ε can accelerate the convergence speed and shorten the training time, but it can lead to an increase in the reconstruction error. Reducing ε can decrease the reconstruction error, but it will lead to a decrease in convergence speed and an increase in training time. The traditional deep learning model based on BBN adopts a fixed learning rate, that is, ε is a constant in each iteration. However, when the learning rate is set globally to a constant, it is impossible to dynamically select the training speed and training accuracy according to the characteristics of the training process itself, which in turn affects the efficiency and recognition rate of transformer fault diagnosis. For this purpose, this paper proposes an adaptive deep learning method, which makes the following improvements to the determination method of the learning rate:

$$\varepsilon_{ij}(t+1)=\begin{cases}(1+\alpha)\varepsilon_{ij}(t), \Delta>0\\(1-\alpha)\varepsilon_{ij}(t), \Delta<0\end{cases} \tag{1}$$

In the formula: α is the learning rate adjustment coefficient, and there is $0<\alpha<1$; Δ is the product of the changes in DBN parameters over two consecutive iterations, which can be expressed as:

$$\Delta=\left(v_i(t)h_j(t)_{\text{data}}\right)-\left(v_i(t)h_j(t)_{\text{recon}}\right) \tag{2}$$

If the result of Eq. (2) is positive, it indicates that the update direction of the RBM parameters is the same in two consecutive iterations, which means that the RBM update process has not yet reached the optimal value. The learning rate can be appropriately increased to accelerate the iterative update. If the result of Eq. (2) is negative, it indicates that the RBM parameter update directions are opposite in two consecutive iterations. At this time, the current learning rate is still maintained, which is highly likely to skip the accurate value and increase the error. By adopting the adaptive learning rate described in this paper,

the learning rate can be adaptively adjusted according to the update trend of the RBM parameters, thereby enabling the update process of the RBM parameters to take into account both the evolutionary speed and the evolutionary accuracy.

Similarly, during the reverse fine-tuning stage, if the parameter update directions are the same for two consecutive times, it indicates that the parameter update in this reverse fine-tuning stage has not reached the optimal value. To accelerate the update rate, the learning rate can be increased, suggesting that the two consecutive iterative processes may have skipped the optimal value. If the current learning rate is maintained, the error will increase. Reference [15] shows that in the reverse fine-tuning stage, the parameter evolution direction can be represented by Eq. (11). Therefore, the adaptive learning rate can also be used for the update of weights, bias vectors, etc. in the reverse fine-tuning stage, so that during the reverse fine-tuning stage, the update of DBN parameters can be adaptively adjusted according to the evolution direction.

In conclusion, by adopting the method proposed in this paper, the learning rate can be dynamically adjusted according to the real-time changing trend of DBN parameters, which improves the convergence speed and convergence accuracy of the DBN network.

3.3 Multimodal Information Data Fusion Processing

Infrared thermometers sense the changes in the radiation energy of objects due to temperature variations and decode temperature information from the amplitude of the image signal. The quantitative relationship of the amplitude Us of the detector output signal during this process is described as:

$$U_s = k_1 \varepsilon T^s \int_{\lambda_1}^{\lambda_2} f(\lambda T) R(\lambda) d_\lambda \tag{3}$$

In the formula, k_1 is a constant; ε is the spectral emissivity of the target being measured; T represents the thermodynamic temperature of the measured part of the transformer. In the formula, the working band λ_1 to λ_2, the Boltzmann constant b, and the system's spectral response $R(\lambda)$ are the key parameters. Thus, Us can be expressed as the product of a constant part ($k_1\varepsilon$) and a functional part $F(T)$ that varies only with the target temperature T. Within the wavelength range from λ_1 to λ_2, the spectral radiation emitted by a radiation source (such as the sun or a black body) at a temperature T, after being modulated by $R(\lambda)$, the total radiation energy or the magnitude of the detector output signal Us.

The constant k_1 is calibrated within the effective range by the system gain, and $F(T)$ is a function of the blackbody temperature. Based on this, the temperature measurement system completed the mapping from the temperature signal to the electrical signal, and the amplitude of the electrical signal became the basis for inverting the surface temperature and drawing the infrared thermal image. In the image, the brightness of the casing area is positively correlated with the temperature. The higher the brightness, the greater the probability of an overheating fault.

In addition, the sound intensity of partial discharge is directly proportional to the discharge energy, and this relationship is given by the following formula:

$$I = uE \tag{4}$$

In the expression of this function, I corresponds to the sound intensity, E is the discharge energy, and μ is the proportionality constant.

If the changes in air density and sound velocity are ignored, the sound intensity I can be considered to follow the following relationship with the sound pressure P:

$$I = \frac{p^2}{\rho v} \tag{5}$$

In the formula, ρ represents the density of air; υ is the speed of sound, P represents sound pressure, which is the pressure variation caused by the propagation of sound waves through a medium.

Based on the physical relationship revealed by Eq. (4) and Eq. (5), the intensity of partial discharge can be indirectly evaluated by detecting the intensity of the ultrasonic sound pressure signal, and the ultrasonic detection spectrum presented in Fig. 1b can be generated accordingly. The amplitude of the waveform in this spectrum directly reflects the intensity of the discharge inside the transformer. Generally speaking, the higher the amplitude, the greater the possibility of a high-energy discharge fault occurring.

Ultra-high frequency (UHF) detection technology enables the determination of the location and intensity of the discharge source by deploying dedicated sensors to capture electromagnetic wave signals excited by partial discharges. This method has a strong anti-interference ability and can effectively filter out external electromagnetic noise caused by corona discharge in the air.

Fig. 1c shows the pulse phase distribution PRPD (Phase Resolved Partial Discharge) spectrum of partial discharge, where the horizontal axis represents the phase of discharge occurrence, the vertical axis represents the signal amplitude, and the color depth of each point in the figure is used to characterize the density of discharge occurrence. By analyzing the PRPD spectrum, key information such as the phase distribution, amplitude fluctuation and aggregation degree of the discharge signal can be obtained intuitively. The denser the discharge poles shown in the figure, the higher the probability of partial discharge activity in the transformer.

For the above four types of data with different sources, this study has respectively formulated corresponding preprocessing strategies:

The DGA data are numerical characteristics. Referring to industry norms, the measured concentrations of five characteristic gases, namely H_2, CH_4, C_2H_6, C_2H_4, C_2H_2 were selected and normalized as model input. The remaining three types of image data are processed as follows respectively:

To avoid redundant recognition of the entire infrared (IR) image, the original four-channel color image is converted into a single-channel grayscale image and cropped into an equal-sized area containing only the

high-voltage bushing as valid input. An example of this is shown in Fig. 1a.

The acoustic emission (AE) waveform graph eliminates the interference of irrelevant information outside the coordinate area, extracting only the coordinate area containing the waveform curve and converting it into a grayscale image as input. The processing result is shown in Fig. 1b.

In the ultra-high frequency (UHF) detection data, the three-dimensional real-time spectra that are not convenient for two-dimensional analysis are discarded, and only the intuitive PRPD planar analysis spectra are selected as input information, as shown in Fig. 1c.

The preprocessing procedures of the above three types of image data are summarized and presented.

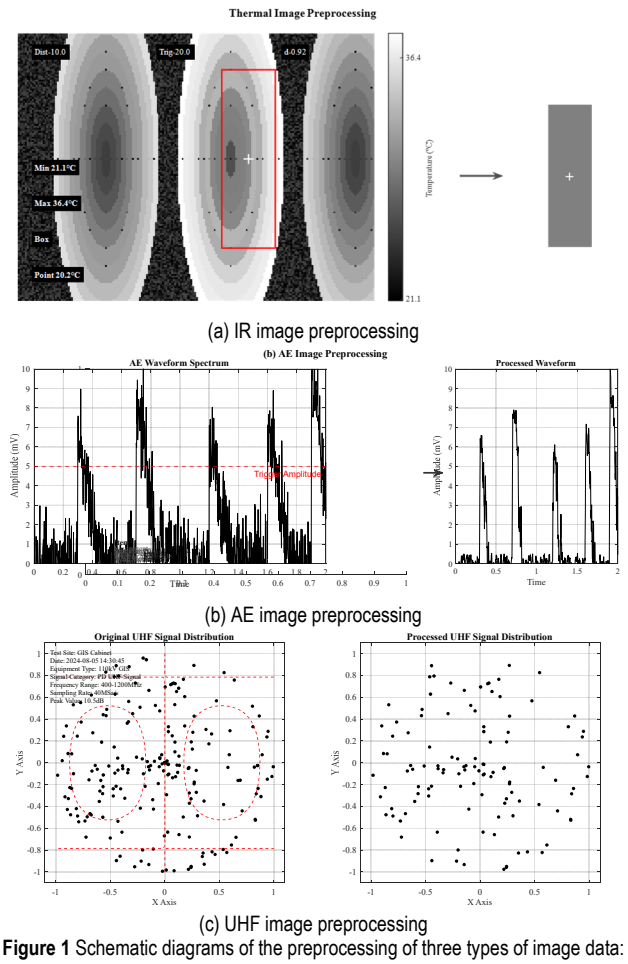

Figure 1 Schematic diagrams of the preprocessing of three types of image data: IR, AE, and UHF

3.4 Establishment of an Intelligent Transformer Fault Diagnosis Model Based on Multimodal Feature Information Fusion

After the data preprocessing is completed, the obtained data can be used as the input of the intelligent diagnosis model. The subsequent work focuses on constructing the corresponding diagnostic model by using machine learning algorithms.

For the numerical and image mixed data involved in this paper, a parallel hybrid network structure is designed. This structure deeply integrates deep neural networks (DNN) and convolutional neural networks (CNN) to achieve multi-source feature extraction, fusion and classification. The specific modeling process is as follows:

(1) DNN modeling of DGA numerical branches:

Input the normalized DGA data into the network input layer;

Four layers of fully connected neural networks are constructed in sequence, with the number of neurons in each layer being 8, 16, 16, and 8 respectively. The final output is a one-dimensional feature vector F_1.

(2) CNN Modeling of image branches (taking infrared image IR as an example):

The input is a normalized IR image with a size of 32 × 32.

Convolutional layer C_1 uses 32 3 × 3 convolutional kernels for feature extraction, and the output feature map size is 32 × 30 × 30.

Pooling layer P_1 uses 2 × 2 Max pooling and reduces the dimension of the output feature map to 32 × 15 × 15.

Repeat the above convolution and pooling operations, and further extract features through the C_2 and P_2 layers in sequence.

The feature map output by P_2 is flattened by the fully connected layer F to generate a one-dimensional feature vector F_2.

(3) Other image modal modeling:

For the acoustic imaging (AE) and ultra-high frequency (UHF) image data, CNN models with the same structure as in (2) were constructed respectively. The corresponding operations were repeated to obtain the feature vectors F_3 and F_4 respectively.

(4) Multimodal feature fusion

The three feature vectors F_2, F_3, and F_4 from the image branch are concatenated and then merged with the vector F_1 from the DGA branch to form the fused joint feature vector V.

(5) Construction of classification modules:

Taking the fusion feature V as the input, it is further connected to a 4-layer fully connected network, with the number of neurons in each layer being 16, 32, 32, and 16 respectively. Finally, the classification layer outputs the fault diagnosis result.

In summary, the overall process of the multimodal information fusion transformer intelligent diagnosis model constructed in this paper is shown in Fig. 2.

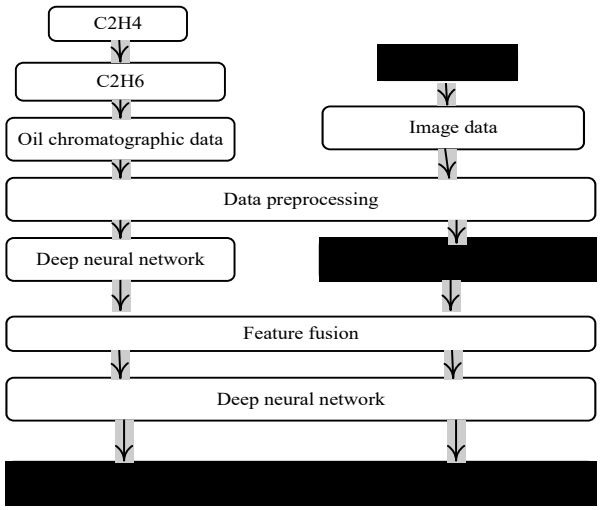

Figure 2 Framework of the intelligent diagnosis model for transformers based on multi-modal feature information fusion

Based on industry standards and taking into account typical faults of transformers comprehensively, this study has established a classification system consisting of six states. This system covers five specific types of faults subdivided under the two major fault types of overheating and discharge, and integrates the normal operating state of no fault (NF). All samples are classified according to this standard, and the specific categories are shown in Tab. 2.

Table 2 Transformer fault classification labels

Category	No faults	Medium and low temperature overheating	Higher than warm	Low-energy discharge	High-energy discharge	Partial discharge
Code	NF	LT	HT	LD	HD	PD

Among the multimodal data adopted in this research, infrared (IR) images mainly reflect the overheating faults of transformers and can further effectively identify low-temperature (LT) and high-temperature (HT) overheating. Acoustic emission (AE) and ultra-high frequency (UHF) images mainly correspond to discharge type faults. Among them, AE images can be used to distinguish low-energy (LD) and high-energy (HD) discharges, while UHF images are mainly used for the identification of partial discharge (PD) type faults.

For the six-category machine learning task constructed in this paper, there is an imbalance in the distribution of training samples among various categories. Among them, the number of normal state (NF) and partial discharge (PD) samples is significantly less than that of the other four categories. To alleviate the impact of category imbalance on model training, a weighted cross-entropy was adopted as the loss function during the network optimization process, and the Kappa coefficient was selected as the evaluation index of the model's classification performance to more objectively reflect the model's discriminative ability under complex distributions.

Weighted cross-entropy loss function

Cross Entropy (CE), as a core concept in information theory, is often used in neural networks as a loss function to measure the difference between the predicted probability distribution and the true label. Its basic form is defined as:

$$\mathrm{CE} = \sum_{k} p(k) \cdot \log \frac{1}{q(k)} \tag{6}$$

In the given formula, k represents the category index, p represents the true category probability distribution, and q is the category distribution predicted by the model.

If the traditional cross-entropy loss function is directly adopted to handle the unbalanced sample set in this paper, the model will tend to classify the input as a category with a larger sample size due to being dominated by the majority of classes, thereby resulting in a significant decline in the overall classification performance. To alleviate this issue, this paper introduces a weighted cross-entropy CE_{ω} to replace the original loss function, and its mathematical expression is as follows:

$$\mathrm{CE}_{\omega} = \sum_{k} \omega(k)\, p(k) \cdot \log \frac{1}{q(k)} \tag{7}$$

In the formula, $\omega(k)$ represents the weight of the KTH type of sample. Its calculation is:

$$\omega(k) = \left(\frac{1}{p(k)} \right) / \sum_{k} \left(\frac{1}{p(k)} \right) \tag{8}$$

In this way, by assigning higher weights to categories with smaller sample sizes while reducing the weights of the majority of class samples, the contribution of each category to the loss function can be effectively balanced, thereby significantly alleviating the structural bias of the model during the classification process.

(2) Kappa coefficient

The Kappa coefficient is a statistical indicator often used to evaluate the consistency between classification results and true labels. Its calculation method fully takes into account the impact of random consistency, and thus is widely used to measure the actual performance of classification models. The specific calculation formula for this coefficient is as follows:

$$\mathrm{kappa} = \frac{p_o - p_e}{1 - p_e} \tag{9}$$

In the above formula, p_0 represents the overall classification accuracy rate, whose value is equal to the ratio of the number of samples correctly classified by the model to the total number of samples.

Suppose the classification task contains m categories, and the actual number of samples for each category is a_1, a_2, …, a_m, the corresponding number of predicted samples is b_1, b_2, …, b_m, and the total sample size is n, then the following relationship can be established:

$$p_e = \frac{a_1 b_1 + a_2 b_2 + \ldots a_m b_m}{n^2} \tag{10}$$

The Kappa coefficient, by introducing a correction mechanism, can effectively reduce the evaluation bias of the model caused by the imbalance of sample distribution, and thus shows significant advantages when dealing with imbalanced datasets. The theoretical value range of this coefficient is [0, 1], and its value is positively correlated with the degree of consistency between the predicted and the actual results. Depending on the specific values, this consistency is usually classified into five grades as shown in Tab. 3.

Table 3 Consistency test table of kappa coefficient

Value of kappa coefficient	Level to which
0.0~0.20	Extremely low consistency (slight)
0.21~0.40	General consistency (fair)
0.41~0.60	moderate consistency
0.61~0.80	substantial consistency almost perfect
0.81~1	Level to which

4 SIMULATION VERIFICATION

In the practice of transformer fault detection and diagnosis, the analysis of dissolved gases in oil and infrared temperature measurement technology for high-voltage bushings have been realized for online monitoring, while

ultrasonic detection and partial discharge detection are usually involved as auxiliary diagnostic means after a fault occurs. This difference leads to different data collection cycles corresponding to various detection methods. To address the problem of integrating data from multiple time scales, based on the transformer fault cases and operation data provided by a maintenance company in a certain province in northwest China, this paper has formulated the following processing strategies: If the sample lacks oil chromatography or infrared detection data, it should be directly excluded; if the sample lacks ultrasonic or partial discharge detection data but there are valid data in the historical records, the most recent detection results should be used. If there is no historical data, the mean of other samples under the fault category to which this sample belongs shall be used for filling.

After the above data cleaning and filling processing, and by eliminating redundant and abnormal samples, 1019 valid samples were finally obtained. The specific distribution of the states of each transformer in the sample set is shown in Tab. 4. All samples were randomly divided in a ratio of 75% to 25%, with 764 samples serving as the training set and 255 samples as the test set.

Table 4 Transformer fault diagnosis dataset

Status	NF	LT	HT	LD	HD	PD
Quantity/piece	42	229	268	162	249	69

To verify that the accuracy of the model can be significantly improved after the fusion of multimodal feature information, this paper sets up three schemes for modeling analysis respectively, as shown in Tab. 5.

Table 5 Comparison schemes of transformer fault diagnosis models based on multimodal feature information fusion

Plan	Input information	Model structure
Option One	Only DGA data	DNN
Plan Two	DGA + IR data	DNN + CNN
Plan Three	DGA + IR + AE + UHF data	DNN + CNN

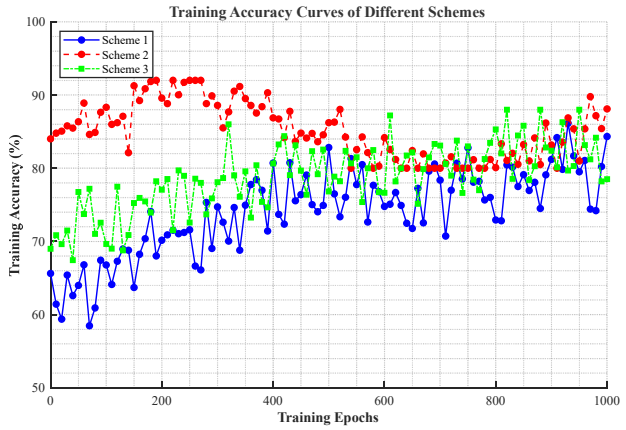

Figure 3 Training curves of the three schemes

After modeling the above three schemes respectively, the learning curves of their training processes and the final test accuracies are shown in Fig. 3 and Tab. 6 respectively. Analysis shows that the convergence process of Scheme Three (i.e., the multimodal feature fusion diagnosis model proposed in this paper) is the most rapid. After only 20 rounds of training, the accuracy has exceeded 90%, demonstrating the optimal training efficiency. On the test set, Scheme Three achieved an accuracy rate of 96.47%, which was a significant improvement over both Scheme One and Scheme Two. The performance comparison of the three schemes clearly indicates that the richer the information sources input by the model, the better its diagnostic performance. This verifies the effectiveness of the multimodal feature fusion strategy in improving the accuracy of transformer fault classification.

Table 6 Comparison of the accuracy of transformer fault diagnosis models based on multimodal feature information fusion

Plan	Plan One	Plan Two	Plan Three
Final test accuracy / %	74.92	84.32	96.48

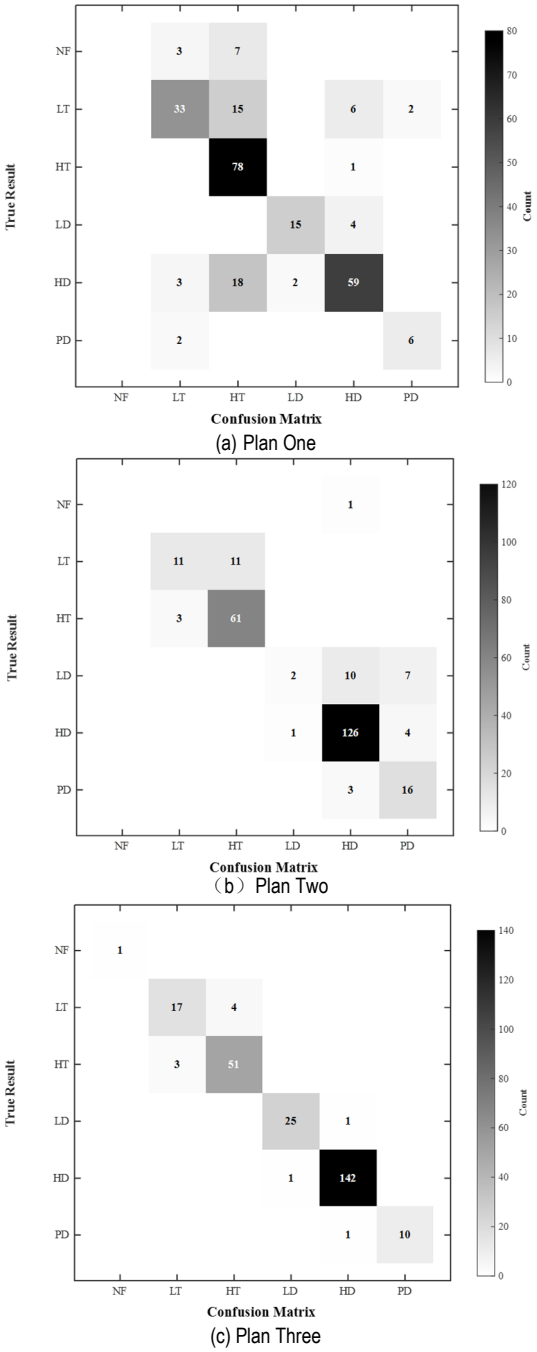

Figure 4 Confusion matrix of the prediction results of the three schemes

For multi-classification tasks, the confusion matrix can visually present the classification performance of the

model. The columns of this matrix represent predicted categories, the rows represent true categories, and the values on the main diagonal correspond to the number of samples that have been correctly classified.

The confusion matrices corresponding to the three diagnostic schemes are shown in Fig. 4. The number in each square in the figure represents the number of samples in the corresponding category. Taking the "33" in the second row and second column of Fig. 4a as an example, it indicates the number of samples where both the actual category and the predicted category are LT. The depth of the color on the right ribbon reflects the number of samples. The darker the color, the more samples there are.

It can be seen from the confusion matrix that Scheme One only uses DGA numerical data but lacks image information, and the classification effect of all categories is not ideal. In the absence of AE and UHF data, the discrimination ability of Scheme Two for three types of discharge faults (LD, HD, PD) is significantly insufficient. Scheme Three integrates all multi-source information and demonstrates high classification accuracy in all categories, further verifying the improvement effect of multimodal data fusion on model performance. It is worth noting that the misclassified samples in Scheme Three are mainly concentrated in LT and HT categories. Specifically, four LT samples were misjudged as HT, and three HT samples were misjudged as LT. Similar phenomena also exist in Scheme One and Scheme Two. This indicates that the feature extraction ability of CNN for IR images may be weaker than that for AE and UHF images. As for whether the reason lies in the insufficient discrimination of IR images themselves or the limitations of the network structure design, further exploration is still needed.

In addition, the Kappa coefficients of the three schemes and their corresponding consistency grades are shown in Tab. 7. It can be seen that the Kappa coefficient of Scheme Three is significantly higher than that of Scheme One and Scheme Two, and its evaluation grade is "almost exactly the same", which once again confirms the superiority of this scheme.

Table 7 Comparison of kappa coefficients and consistency test results of the three schemes

Plan	Value of kappa coefficient	Level to which
Option One	0.654	High consistency
Plan Two	0.753	High consistency
Plan Three	0.943	Almost exactly the same

Aiming at the common problem of sample imbalance in transformer fault diagnosis, this study proposes to adopt the weighted cross-entropy loss function to replace the traditional cross-entropy. To verify its effectiveness, a control experiment was constructed: diagnostic models were established using the improved loss function and the traditional loss function respectively, and their performance was compared. The comparison experiment results of the two models are shown in Fig. 5, where "Accuracy" represents the overall classification accuracy of the models.

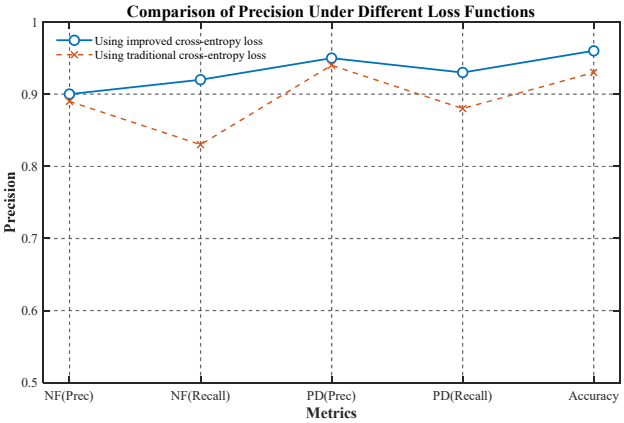

Figure 5 Influence of the improved cross-entropy loss function on the model

As can be seen from Fig. 5, after introducing the improved weighted cross-entropy loss function, the model's recognition performance for fault categories with fewer training samples such as NF and PD has been significantly improved. Meanwhile, the overall classification accuracy of the model has also improved simultaneously, and the results are in line with the research expectations.

Under the same network parameter Settings, the adaptive learning rate method proposed in this paper and the traditional fixed learning rate strategy are respectively adopted to train the model. Fig. 6 shows the variation of the root mean square error between the output of the Softmax layer and the real label during the training process of the two methods. It can be seen from the results that as the number of iterations increases, the errors of both methods decrease rapidly. However, after 20 iterations, the reconstruction error of the method proposed in this paper is significantly lower than that of the fixed learning rate method, and the error decline rate is faster. When the iteration reaches approximately 100 times, the root mean square error of the method proposed in this paper has dropped below 0.08, while the traditional method requires approximately 400 iterations to reach the same level. The above experiments show that the adaptive learning rate adjustment mechanism proposed in this paper can effectively accelerate the model convergence process and improve the reconstruction accuracy, thereby achieving efficient adaptive construction of deep learning models.

The diagnostic results of the transformer fault diagnosis methods using the method proposed in this paper and those based on a fixed learning rate deep learning model were compared respectively. Fig. 7 shows the change curves of the diagnostic accuracy rates of the two methods as the number of iterations increases. As can be seen from the results in Fig. 7, with the increase in the number of iterations, the reconstruction errors of the deep learning models constructed by the two methods keep decreasing, and the diagnostic accuracy of the two methods also keeps improving. Since the method proposed in this paper can adaptively adjust the learning according to the iterative process and then adaptively construct the deep learning model, the correct rate of transformer fault diagnosis of the method proposed in this paper is also relatively high under each iteration number. When the number of iterations reaches 250, the fault diagnosis accuracy rate of the method proposed in this paper is 93%.

Subsequently, due to the relatively high reconstruction accuracy of the deep learning model, the fault diagnosis accuracy rate has basically remained stable as the number of iterations increases. When the traditional fixed learning rate method is adopted and the number of iterations is 400 times, the accuracy rate of fault diagnosis results is only 87%. The method proposed in this paper is significantly more effective than the traditional ones.

The spectrum of the vibration signal of the single-sided measurement point of the 8th group of data of a certain transformer as a pilot is shown in Fig. 8, and the spectra of the two measurement points of the first group of data of the faulty transformer obtained are shown in Fig. 9. As can be seen from Fig. 8 and Fig. 9, regardless of whether the transformer is normal or faulty, the amplitudes corresponding to frequency points after 1 kHz are very small. It can be considered that the components corresponding to these frequency points are independent of the state of the transformer.

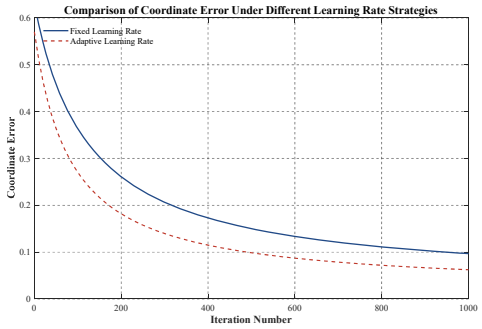

Figure 6 Reconstructs the error variation curve

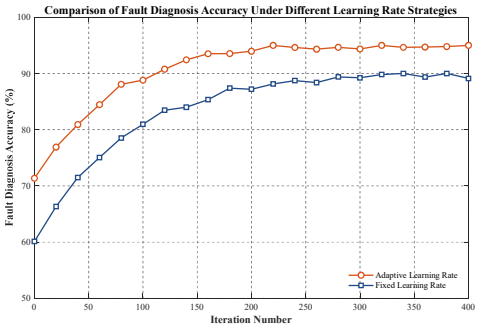

Figure 7 Influence of the number of iterations on the diagnostic accuracy

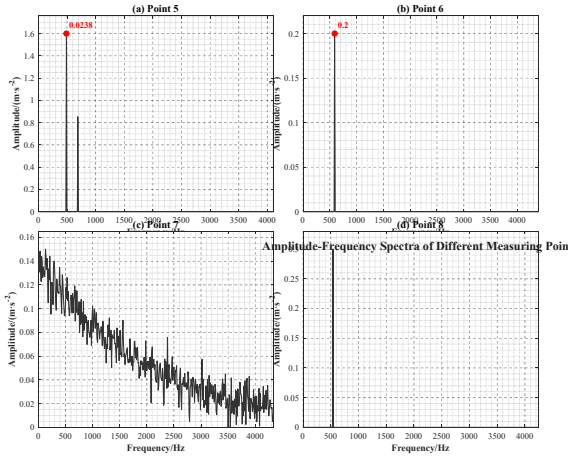

Figure 8 Vibration signal spectra of some measurement points (pilot transformer)

The frequency points of the spectrum obtained by FFT transformation are spaced at 50 Hz, with a total of 21 points ranging from 0 to 1000 Hz. The first 21 points of the spectrum were extracted using MATLAB, and the spectrum data were combined into a sample set in the form. Finally, 112 normal transformer samples and 3 fault samples were obtained, and the dimension of each sample was 168.

In the comparison of feature extraction effects, the feature distributions obtained by multimodal fusion and deep learning analysis methods show significant overlap in the features of three types of faults: winding nesting, loosening, and deformation. Overall, they present the characteristics of a wide distribution range and weak aggregation. There is a lack of sufficient separation between various categories, which is not conducive to the subsequent diagnostic tasks. In contrast, the features extracted by the stack autoencoder (SAE) method adopted in this paper have highly clustered fault samples of the same type, with clear boundaries between different categories, demonstrating excellent intra-class compactness and inter-class separability. This provides a more discriminative feature representation for fault identification, and the comparison of its effects is shown in Fig. 10.

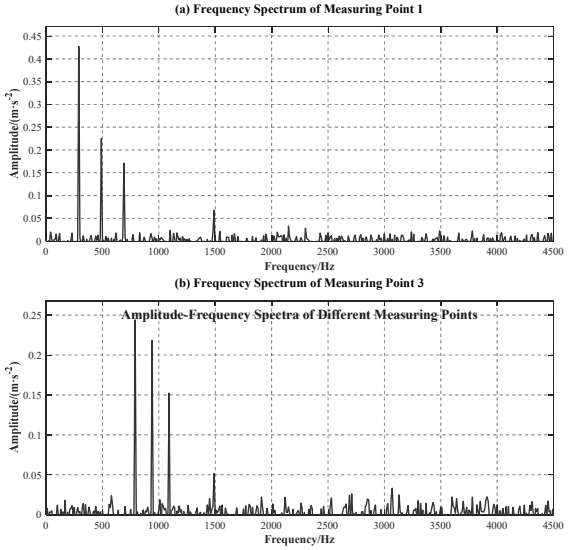

Figure 9 Spectral analysis of some measurement points (fault transformer)

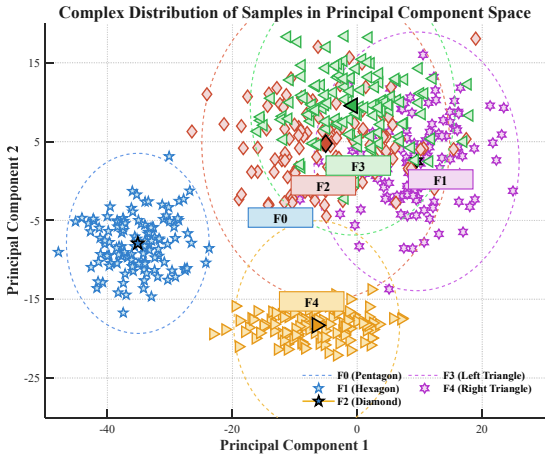

Figure 10 Features extracted by multimodal fusion and deep learning methods

Based on deep learning models, the ability to extract and analyze features is strong. Meanwhile, the adaptive construction of deep learning models can be achieved through the adjustment coefficient of the learning rate, and the reconstruction accuracy is relatively high. Therefore, the transformer fault diagnosis method based on this paper has a significantly higher diagnostic accuracy rate than other methods.

5 CONCLUSION

This paper applies deep learning models to transformer fault diagnosis. At the same time, by combining the shortcomings of traditional deep learning models based on DBN and making adaptive improvements, an adaptive deep learning model-based transformer fault diagnosis method is proposed. The number of hidden layers in an adaptive deep learning model has a significant impact on the accuracy of transformer fault diagnosis. Considering both the model convergence time and the diagnostic accuracy, the number of hidden layers can be set to 4. An adaptive deep learning model was constructed based on the adaptive adjustment coefficient, which significantly improved the reconstruction accuracy and shortened the convergence time. The hybrid mode can more comprehensively and effectively utilize different measurement data to extract the modal shape. This method combines the high sensitivity of the acceleration response in global monitoring with the advantage of precise local damage location in strain response, thereby achieving more comprehensive and reliable damage identification. The feature extraction and analysis capabilities are relatively strong. Meanwhile, the diagnostic model construction accuracy is high and the convergence time is short, significantly improving the accuracy of the diagnostic results. The improved algorithm proposed in this paper has only shown its effectiveness to a limited extent. Any algorithm has its own local adaptability. According to the "No Free Lunch Theorem", algorithms have disadvantages in other aspects and further research and discussion are still needed.

6 REFERENCES

[1] Gao, W., Wang, Y., & Liu, Y. (2024). Research on Power Transformer Fault Diagnosis Based on Experience and Data. *2024 Second International Conference on Cyber-Energy Systems and Intelligent Energy (ICCSIE)*, 1-6. https://doi.org/10.1109/ICCSIE61360.2024.10698734

[2] Tang, P., Zhang, Z., & Tong, J. (2024). Research on transformer fault diagnosis based on active learning with imbalanced data of dissolved gas in oil. *Review of Scientific Instruments*, *95*(5), 12-27. https://doi.org/10.1063/5.0200813

[3] Nanfak, A., Eke, S., & Kom, C, H. (2025). A dissolved Gases Analysis Method for Power Transformer Faults Diagnosis Based on the Observation of Subsets of Labelled Fault Data. *Journal of Electrical Engineering & Technology*, *20*(4), 2019-2028. https://doi.org/10.1007/s42835-025-02144-2

[4] Yan, P., Chen, F., & Kan, X. (2023). Research on transformer fault diagnosis based on an IWHO optimized MS1DCNN algorithm and LIF spectrum. *Analytical Methods*, *15*(29), 15-32. https://doi.org/10.1039/D3AY00713H

[5] Lin, C. & Fang, R. (2024) Research on Fault Diagnosis Method of Steam Turbine Generator Rotor Abnormal Vibration Based on Probabilistic Neural Networks. *Springer Proceedings in Physics*, 115-124. https://doi.org/10.1007/978-981-97-3686-7_10

[6] Xu, P., Jia, Z., & Yao, G. (2024). Research on Power Transmission Line Fault Diagnosis Based on Transformer-BiGRU. *2024 The 9th International Conference on Power and Renewable Energy (ICPRE)*, 836-841. https://doi.org/10.1109/ICPRE62586.2024.10768414

[7] Sui, X., Li, J., & Wang, Z. (2023). Research on Power Transformer Fault Diagnosis Based on Improved Wavelet Packet Energy and Hidden Markov Model. *2023 IEEE 6th International Electrical and Energy Conference (CIEEC)*, 3167-3172. https://doi.org/10.1109/CIEEC58067.2023.10166837

[8] Wang, Z., Yang, S., & Xiong, Y. (2024). Research on Fault Diagnosis Method for Power Transformers Based on Unbalanced Data Driving. *2024 5th International Conference on Mechatronics Technology and Intelligent Manufacturing (ICMTIM)*, 421-424. https://doi.org/10.1109/ICMTIM62047.2024.10629496

[9] Lu, J., Ji, W., & Yu, J. (2025). Data driven deep learning fault diagnosis method based on vision transformer and multi-head attention for different working condition. *Engineering Research Express*, *7*(1), 15205-15218. https://doi.org/10.1088/2631-8695/ada3b2

[10] Hou, P., Zhang, J., & Jiang, Z. (2023). A Bearing Fault Diagnosis Method Based on Dilated Convolution and Multi-Head Self-Attention Mechanism. *Applied Sciences-Basel*, *13*(23), 17-29. https://doi.org/10.3390/app132312770

[11] Hong-Wei, F., Ning-Ge, M. et al. (2024). New intelligent fault diagnosis approach of rolling bearing based on improved vibration gray texture image and vision transformer. *Part C: Journal of Mechanical Engineering Science*, *238*(13), 14-29. https://doi.org/10.1177/09544062221085871

[12] Khalladi, S. A., Ouessai, A., & Benamara, N. K. (2024). Efficient Road Traffic Video Congestion Classification Based on the Multi-Head Self-Attention Vision Transformer Model. *Transport and Telecommunication Journal*, *25*(1), 20-30. https://doi.org/10.2478/ttj-2024-0003

[13] Chen, Y. & Zhang, R. (2025). Deep Multiscale Convolutional Model With Multihead Self-Attention for Industrial Process Fault Diagnosis. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 1-10.

[14] Ai, Z., Cao, H., & Wang, J. (2023). Research Method for Ship Engine Fault Diagnosis Based on Multi-Head Graph Attention Feature Fusion. *Applied Sciences-Basel*, *13*(22):20-37. https://doi.org/10.3390/app132212421

[15] Peng, C., Li, H., & Gui, W. (2025). Fault Diagnosis Method for Rotating Machinery Based on MSCNN-MGAT. *IEEE Transactions on Instrumentation and Measurement*, *74*(5), 1-11. https://doi.org/10.1109/TIM.2025.3587368

[16] Wang, H., Ju, X., & Zhu, H. (2025). SEFormer: A Lightweight CNN-Transformer Based on Separable Multiscale Depthwise Convolution and Efficient Self-Attention for Rotating Machinery Fault Diagnosis. *Computers, Materials & Continua*, *82*(1), 95-103. https://doi.org/10.32604/cmc.2024.058785

[17] Wang, H., Wu, Z., & Li, Q. (2024). A fault diagnosis method for variable speed planetary gearbox based on ADGADF and Swin Transformer. *Insight: Non-Destructive Testing & Condition Monitoring*, *66*(4), 232-254. https://doi.org/10.1784/insi.2024.66.4.232

[18] Zhu, S., Chen, J., & Xie, R. (2024). Research on Intelligent Fault Diagnosis Method of Planetary Gearbox Based on ITD and CNN. *2024 Global Reliability and Prognostics and Health Management Conference (PHM-Beijing)*, 1-7. https://doi.org/10.1109/PHM-Beijing63284.2024.10874647

[19] Sun, X., Ding, H., Li, N., Dong, X., Sun, J., & Zheng, G. (2025). Intelligent Fault Diagnosis Method for Shearer

Rocker Gear Based on Swin Transformer and Multiscale Convolution Parallel Integration. *IEEE Transactions on Instrumentation and Measurement*, *74*, 1-16. https://doi.org/10.1109/TIM.2025.3551002

[20] Yang, S., Gao, G., & Wang, Z. (2025). Intelligent Fault Diagnosis System for Running Gear of High-Speed Trains. *Sensors*, *25*(17), 5269. https://doi.org/10.3390/s25175269

[21] Yang, N., Liu, J., & Zhao, W, Q. (2025). Research on Quantitative Diagnosis of Helical Gear Pitting Fault Severity Based on CycleGAN and Improved DANN. *IEEE Transactions on Instrumentation and Measurement*, *74*(2), 1-15. https://doi.org/10.1109/TIM.2025.3569004

[22] Sun, H., Hao, R., & Zheng, X. (2024). Gear Box Fault Diagnosis Based on Neural Network with Multi-Scale Weighted Fusion. *2024 IEEE International Conference on Sensing, Diagnostics, Prognostics, and Control (SDPC)*, *43*, 301-307. https://doi.org/10.1109/SDPC62810.2024.10707724

[23] Du, X., Jia, L., & Ul Haq, I. (2022). Fault diagnosis based on SPBO-SDAE and transformer neural network for rotating machinery. *Measurement*, *188*, 110545. https://doi.org/10.1016/j.measurement.2021.110545

[24] Zhu, C., Ma, L., & Zhou, Y. (2025). Abnormal detection method of hydro-turbine generator based on vision transformer. *Intelligent Decision Technologies-Netherlands*, *19*(5), 3055-3071. https://doi.org/10.1177/18724981251356644

[25] Fei, X. & Haijun, W. (2024). Research on Ferrographic Image Fault Diagnosis Based on Channel Overlapping Technique and Information Fusion Mechanism. *Journal of Tribology*, *146*(7), 828-846. https://doi.org/10.1115/1.4064858

[26] He, D., Xu, Y., & Sun, H. (2025). Self-supervised learning for vehicle bearing fault diagnosis based on time-frequency dual-domain contrast and fusion. *Nonlinear Dynamics*, *113*(14), 17385-17412. https://doi.org/10.1007/s11071-025-11101-7

Contact information:

Cunjian TIAN
Extra-high Voltage Branch of State Grid Fujian Electric Power Co., Ltd.,
Fuzhou Fujian, 350011, China
E-mail: tiancunjian_fj@163.com