ARCHIVES OF ELECTRICAL ENGINEERING VOL. 75(1), pp. 23–40 (2026)

DOI 10.24425/aee.2026.156800

Transformer fault diagnosis method based on multilevel acoustic information

XUAN LI ✉, PENG WU, JIAN SHAO, JIABI LIANG, QUN LI, YUNCAI LU, TONGLEI WANG

State Grid Jiangsu Electric Power Research Institute
Nanjing, 430000, China
e-mail: ✉ *xuanya183174@163.com*

(Received: 29.04.2025, revised: 23.01.2026)

Abstract: To more accurately obtain the feature information embedded in the acoustic pattern of transformers, a transformer fault diagnosis method is proposed based on multilevel acoustic information of 14 state types. In this method, a parallel dual-channel fault diagnosis model, CNN-BiLSTM-Transformer, is established. First, the modified Mel inversion coefficients and Mel spectrograms are extracted from the original acoustic pattern data. The modified Mel inversion coefficients and Mel spectrograms are then input into the parallel dual-channel model. In the first channel, a convolutional neural network model is used to extract the feature information of maps. In the second channel, a bidirectional long- and short-term memory network and a Transformer encoder are used to partially extract the temporal features in the MFCCs. Finally, the temporal features extracted from the two channels are fused through multimodal fusion for training. The experimental results show that the proposed diagnostic method can achieve an average accuracy of 99.5% in multiple fault diagnosis. Compared with current mainstream acoustic single-channel diagnostic models, the diagnostic rate of this model is improved by an average of 4.8%, exhibiting higher accuracy and robustness.

Key words: acoustic features, CNN-BiLSTM-Transformer, fault diagnosis, Mel spectrogram, MFCCs, multimodal fusion

1. Introduction

As an important connecting device in power systems, transformers play a crucial role in voltage transformation, current adjustment, power distribution and transmission of the power grid [1, 2]. With the continuous expansion of a grid scale and sharp increase in capacity, the requirements for the reliability and stability of transformers are becoming increasingly stringent. Real-time monitoring and charge detection of transformer operation conditions are of significant importance for timely identification of potential faults and ensuring safe, efficient and stable operation of the power grid [3, 4].

Advances in artificial intelligence technology have made transformer acoustic pattern recognition a hotspot in transformer fault diagnosis [5]. The transformer acoustic signal contains equipment information during operation and can be used as an important indicator of fault conditions [6, 7]. However, the complexity of the transformer operating environment and the diversity of acoustic signals pose significant challenges for extracting effective fault features from acoustic signals and applying them effectively to transformer fault diagnosis and classification. To date, the application of multimodal technology to transformer fault diagnosis remains limited. However, transformer acoustic pattern recognition technology has established a certain basis in the power industry. Wang *et al*. [8] used a weighted processing method and principal component analysis to extract the Mel inverse spectral coefficients of acoustic signals. These coefficients are then used to recognize different degrees of transformer core looseness using a vector quantization algorithm. Zhou *et al*. [9] used compressive sensing technology and a discriminative dictionary learning method to identify defects (including loose windings and iron cores in dry-type transformers), achieving an accuracy rate of greater than 90%. Zhang *et al*. [10] established an acoustic pattern recognition model based on a Mel time spectrum-convolutional neural network. This model can effectively recognize loose faults in iron cores and windings by leveraging the strengths of convolutional neural networks (CNNs) in t image recognition. Cui and Ma [11] put forward an acoustic pattern recognition model for loose faults in transformer cores, and used an enhanced MFCC and 3D-CNN approach to enhance classification accuracy. Liu *et al*. [12] proposed using a blind source separation algorithm, amplitude phase fluctuation method, and 50 HZ octave band cepstrum coefficient to extract acoustic signal features, remove interference signals from the original acoustic signal, and then introduce a gate-controlled recurrent neural network (GRU) to identify the DC bias state of the transformer. Although the above studies have achieved excellent performance in transformer fault diagnosis, most of them only focus on fault identification of transformer categories and do not fully consider the diversity of transformer faults in practice. In addition, compared with the single use of traditional networks such as the CNN [13] and recurrent neural networks [14], the advantages of different networks in their respective fields can be utilized for control.

To address the above issues, a methodology combining bidirectional long short-term memory (BiLSTM)-Transformer and CNN techniques is proposed for identifying faults in transformer acoustic pattern recognition. The proposed approach employs two types of audio features and Mel spectrograms to extract acoustic patterns. Combined with the ability of the CNN in feature extraction and efficient classification, the integration of the BiLSTM-Transformer in capturing complex temporal dependencies can comprehensively capture and understand sound patterns. In addition, it can effectively avoid the limitations of isolated models in practical applications. This innovative technological integration can improve the strengths of each model while addressing their weaknesses, thereby forming a more robust and comprehensive voiceprint recognition method.

2. Data preprocessing

Mel frequency cepstral coefficients (MFCCs) and Mel-Spectrograms are important in sound feature extraction, each with different characteristics of sound patterns [15, 16]. The MFCC converts time-domain signals into frequency-domain signals. Based on the auditory mechanism of the human ear, MFCCs reflect the static characteristics of sound signals such as timbre and intonation.

The calculation of the first-order MFCC is helpful for analyzing the spectrum of sound signals in the frequency domain. The calculation of the first-order and second-order differences in the MFCC allows for the extraction of dynamic characteristics such as speech speed and intonation changes. Mel spectrograms integrate information from time domain and frequency domain, and can clearly visualize the changes in sound frequency over time. It has been demonstrated that this spectrogram cannot only reflect the static characteristics of sound signals (such as pitch and timbre), but also capture the dynamic characteristics (including speech rate and intonation change). In contrast, Mel spectrograms can provide comprehensive information about the frequency distribution of sound signals and their temporal evolution. Each method has a distinct focus on sound feature extraction, and both are essential in sound processing and analysis.

2.1. MFCC feature extractions

The extraction process of MFCCs includes preprocessing, the fast Fourier transform (FFT), Mel filter bank, logarithmic operations, and discrete cosine transform [17, 18].

The preprocessing stage typically consists of three main steps: pre-emphasis, frame-splitting, and windowing. The Fourier transform of the time-domain signal of each frame generates a linear spectrum, $X(k)$, expressed as follows:

$$X(k) = \sum_{n=0}^{N-1} y(n) e^{-\frac{j2\pi nk}{N}} \leq nk \leq N-1, \tag{1}$$

where $y(n)$ is the preprocessed time-domain signal; n is a sampling point in the time-domain; k is a discrete frequency point in the frequency-domain; and N is the length of the first frame of the signal.

The linear spectrum $X(k)$ is processed through the Mel filter bank to generate the Mel frequency, expressed as follows:

$$s(m) = \ln\left(\sum_{k=0}^{N-1} |X(k)|^2 H_m(k)\right), \tag{2}$$

where $s(m)$ is the output of the m-th Mel filter, and $H_m(k)$ is the m-th filter parameter.

Subsequently, the discrete cosine transform (DCT) is performed to obtain MFCCs by taking the logarithm of the Mel frequency, expressed as follows:

$$c(r) = \sum_{b=0}^{P-1} s(m) \cos\left(\frac{\pi r\,(b-0.5)}{M}\right), \quad 1 \leq r \leq L, \tag{3}$$

where b is the frequency channel index; M is the number of filters in the Mel filter bank; P is the order of the MFCC; $c(r)$ is the value of the r dimensional inverse spectral coefficient.

The standard MFCC parameters only reflect the static characteristics of sound parameters. Conversely, the first- and second-order difference of MFCCs can highly reflect the dynamic characteristics of sound [19, 20]. The combination of dynamic and static features can effectively

improve the recognition of the system. The first-order difference feature is calculated as follows:

$$d_t = \begin{cases} C_{t+1} - C_t, & t < K \\ \dfrac{\sum_{k=1}^{K} k(C_{t+k} - C_{t-k})}{\sqrt{2\sum_{k=1}^{K} k^2}}, & \text{else} \\ C_t - C_{t-1}, & t \geq Q - K \end{cases}, \tag{4}$$

where d_t denotes the t-th first-order difference; C_t denotes the t-th cepstrum coefficient; Q denotes the order of the cepstrum coefficient, and K denotes the time difference of the first-order derivatives, taken as 1 or 2.

The order of the discrete cosine transform is 13, and after first-order and second-order differencing, the differencing results are combined into MFCC parameters. This process will result in the collection of a 39-dimensional feature vector, assigning 39 features to each frame of sound data.

2.2. Mel spectrogram preprocessing

Mel spectrograms are a spectral representation method based on human auditory properties [21, 22]. They convert audio signals into spectrograms that are more in line with human auditory habits by simulating human auditory sensitivity to sounds of different frequencies. Typically, Mel spectrograms use a nonlinear Mel scale to represent frequencies, allowing for more accurate display of low-frequency components. However, high-frequency components are relatively compressed to align with the human ear's capacity to perceive sound frequencies. The generation of a Mel spectrogram requires transforming the audio signal from time-domain Fourier transformation to frequency-domain. Subsequently, a filtering operation is performed on the time-frequency domain signal through a triangular filter bank, with each filter corresponding to a specific Mel frequency interval. The Mel filter bank is shown in Fig. 1, and the transfer function of the filter bank is expressed as follows:

$$H_m(f) = \begin{cases} 0, & f < x(m-1) \\ \dfrac{f - x(m-1)}{x(m) - x(m-1)}, & x(m-1) \leq f \leq x(m) \\ \dfrac{x(m+1) - f}{x(m+1) - x(m)}, & x(m-1) \leq f \leq x(m+1) \\ 0, & f < x(m+1) \end{cases}, \tag{5}$$

where m is the filter bank number, taken as 40; $x(m)$ is the center frequency of the triangular filter bank; $H_m(f)$ is the transfer function of the m-th filter in the Mel filter bank with respect to frequency f, expressed as follows:

$$x(m) = \left(\frac{Q}{f_s}\right) \mathrm{Mel}^{-1}\left(\mathrm{Mel}(f_{\min}) + m\frac{\mathrm{Mel}(f_{\max}) - \mathrm{Mel}(f_{\min})}{M+1}\right), \tag{6}$$

where $f_{\max}$ and $f_{\min}$ are the maximum and minimum values of the filter range frequency, respectively; f_s is the sampling frequency of the acoustic pattern, and Q is the frame length of the discrete Fourier transform.

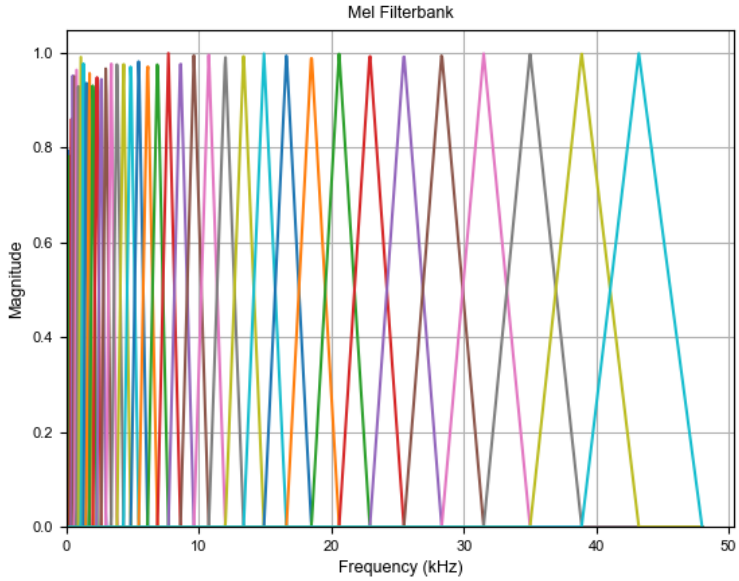

Fig. 1. Mel filter bank

The power spectra of filtered signals are calculated and logarithmic compression is applied to model the non-linear response of the human ear to sound intensity. Finally, these power spectral values are converted to the Mel scale to obtain the Mel spectrogram of the power transformer. As shown in Fig. 2, the features of the Mel spectrogram are compared when different faults occur, which can provide a good training process for CNN feature extraction.

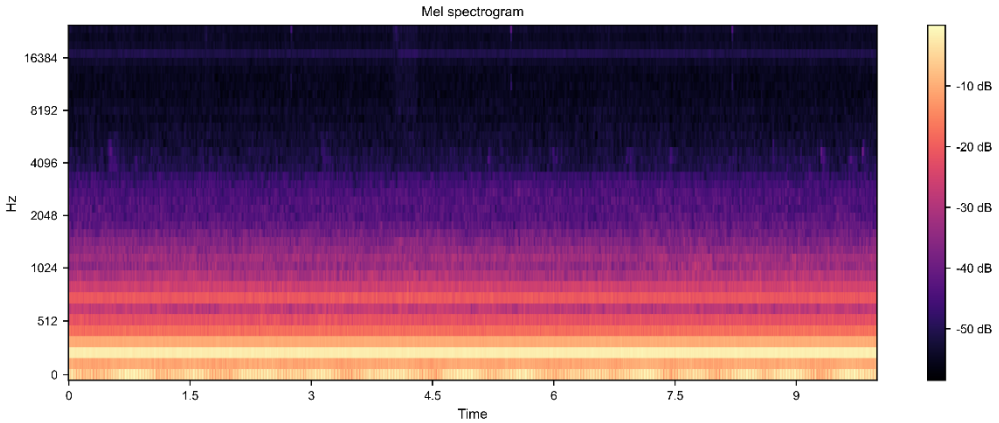

Fig. 2. Mel spectrum

3. Dual-channel voiceprint recognition model construction

3.1. CNN

The CNN has become the preferred method in image recognition due to its high performance in image feature extraction [23]. The unique configuration of convolutional and pooling layers can effectively capture local spatial hierarchical information within an image, laying a foundation for image analysis tasks. This hierarchical feature extraction mechanism enables the CNN to excel in processing high-dimensional data, especially in image recognition, where it can automatically learn key features such as edges, texture, and shape to achieve efficient classification and recognition [24].

Given the inherent advantages of convolutional neural networks in feature extraction and their synergistic effect with Mel spectrograms, using convolutional neural networks for Mel spectrogram processing has significant advantages.

The Mel spectrogram provides a two-dimensional time-frequency representation of audio signals, where the time axis corresponds to temporal sequences and the frequency axis encodes frequency information after the Mel-scale transformation. The convolutional and pooling operations in CNNs significantly enhance the model's capacity to automatically extract critical time-frequency features. Specifically, the convolutional layers capture local time-frequency patterns such as pitch and rhythm, while the pooling layers reduce redundancy by retaining essential information through dimensionality reduction.

This methodology cannot only diminish the complexity associated with manual feature design but also elevate the model's proficiency in characterizing audio signals. In addition, the hierarchical architecture of CNNs facilitates the progressive construction of complex feature representations, moving from low-level to high-level abstractions, thereby further refining the accuracy of audio recognition and classification tasks.

3.2. BiLSTM

An LSTM network is a variant of a recurrent neural network (RNN) that can effectively handle long-time sequences [25]. An LSTM consists of a forgetting layer, an input layer, and an output layer. Considering the inconsistent size and specifications of the collected images, it is necessary to preprocess the images before putting them into the model training.

In the context of non-linear data such as MFCCs, it has been found that traditional LSTM networks are insufficient in extracting data features. In contrast, the use of BiLSTM networks has been proven to be more effective in analyzing speech data in a wide range of information – Fig. 3. The processing of sequential data is achieved through concurrent operations of two LSTM networks. One traverses from front to back, and the other from back to front, to integrate information from both directions [26, 27].

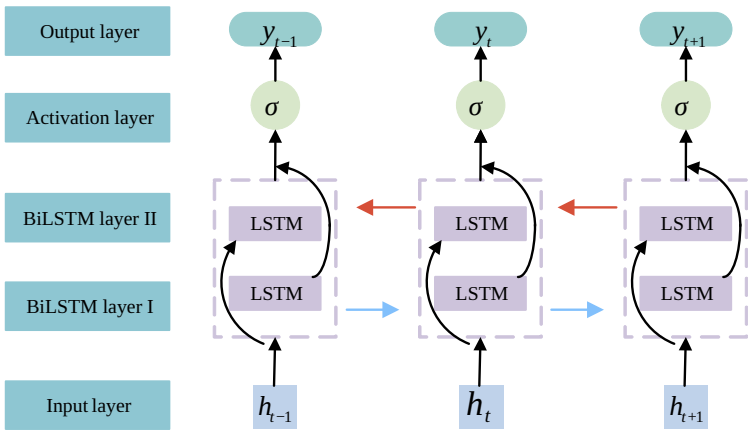

Fig. 3. BiLSTM module structure

3.3. Transformer

The Transformer model is built on the self-attention mechanism, which is a complex data processing framework that facilitates the capture of long-distance dependencies in sequence data. One major strength of the Transformer model is its ability to process all elements simultaneously, thereby significantly improving the efficiency of model training. However, in fault diagnosis, data sources mainly include sensor signals, which largely depend on local features or periodic alterations in the time series compared to the global background of the sequence. Consequently, obtaining optimal classification results in fault diagnosis by directly applying the Transformer may be a challenge. Considering the advantages of the BiLSTM in capturing temporal features of audio data, the combination of the Transformer encoder and BiLSTM can be an effective strategy for transformer fault diagnosis. Taking into account the temporal information of audio signals through its bidirectional structure, the BiLSTM can accurately capture temporal features in the audio. Conversely, the transformer has demonstrated a high capability in analyzing complex hierarchical structures and dependencies of temporal points in audio data, due to its advanced multi-attention mechanism and parallel processing capacity.

The BiLSTM-Transformer encoder is proposed as the network architecture to extract features of MFCCs. The encoder maps the input sequences to a high-dimensional representation and replaces the decoder with a fully-connected layer. The sub-module structure of the Transformer encoder is shown in Fig. 4. The sub-module mainly consists of multi-head attention (multi-head attention) and a feed-forward network layer, and introduces residual connection and layer normalization to prevent gradient degradation and accelerate algorithm convergence.

When the time series are input to the multi-head self-attention layer, it is necessary to add position encoding (also known as position embedding) to describe the relative positional relationship between the time series. The position encoding is expressed as follows:

$$\begin{cases} PE(k,2i) = \sin\left(k/10\,000^{\frac{2i}{d}}\right) \\ PE(k,2i+1) = \cos\left(k/10\,000^{2i/d}\right) \end{cases}, \tag{7}$$

where k is the sequence length; i is the dimension of the feature, and d is the feature length.

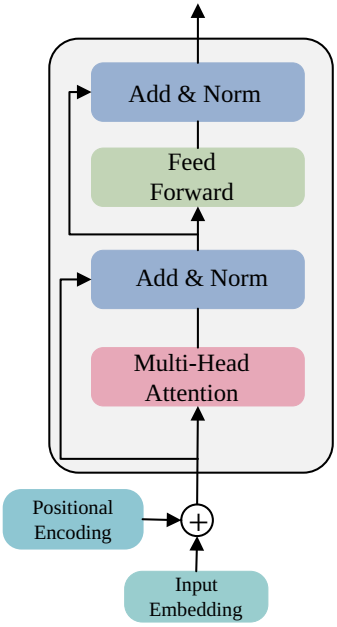

Fig. 4. Transformer encoder structure

The transformer introduces multi-head attention to further enhance the expressive capability of the model [28]. Through computing the attention scores of multiple 'heads' in parallel, each head captures the features of input data from different dimensions or perspectives, greatly enhancing the model's ability to capture diverse information and complex dependencies. The operations are as follows:

$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{8}$$

$$\text{head}_i = \text{Attention}(Q_i, K_i, V_i), \tag{9}$$

$$\text{Multihead}(Q, K, V) = \text{Concat}\left[\text{head}_1, \text{head}_2, \ldots, \text{head}_i\right] W^o, \tag{10}$$

where Attention denotes the attention mechanism; Q, K, and V represent the query, key and value matrices, respectively; d_k denotes the columns of key matrices; head is an independent attention calculation unit in the multi-head attention mechanism; Multihead is a mechanism that computes the attention scores of multiple "heads" in parallel; Concat is short for concatenation; i is the number of heads, and W^o is the weight matrix of training.

The feed-forward neural network layer (feed-forward network) consists of two layers of fully connected networks, each mapping linearly to an input vector. The middle-hidden layer is activated using the ReLU function to enhance the expressive power of the model, which further extracts features based on the multi-head attention mechanism [29]. The feedforward neural network is expressed as follows:

$$\text{FFN}(\alpha) = \text{Re}LU(\alpha W_1 + b_1)W_2 + b_2, \tag{11}$$

where FFN denotes the feed-forward neural network; α denotes the normalized output vector; W_1, W_2 denote the weight matrices; b_1 and b_2 denote the bias terms.

The transformer adopts a multi-head self-attention mechanism to globally model the features extracted from the BiLSTM network, followed by feature transformation via a feed-forward network (FFN) to further refine the representation. This fusion model aims to combine the strengths of both modules, enabling a more comprehensive understanding of temporal dynamics in audio while capturing various patterns and features in audio.

3.4. Multimodal fusion

Compared with the unimodal model, the multimodal model is more powerful in fault diagnosis and has achieved significant improvements in accuracy. A central challenge in multimodal fusion lies in effectively integrating multiple unimodal data sources into a single multimodal dataset, while preserving maximum information content [30, 31]. Model-independent fusion strategies are commonly categorized into three paradigms: early fusion, late fusion, and hybrid fusion, each with distinct advantages and limitations that must be carefully evaluated according to specific application requirements. Based on a comparative analysis of these fusion strategies, early fusion is determined as the most suitable approach for this acoustic feature-based framework. This selection was driven by several key factors: first, the extracted Mel-spectrograms and MFCC features are inherently complementary, as they represent different perspectives of the same acoustic source; then early fusion facilitates the learning of complex cross-modal correlations at the feature level, enabling the model to develop richer representations than single-modal approaches; finally, this approach helps minimize information loss before the classification stage while maintaining reasonable computational requirements. In our implementation, early fusion is applied at the feature level by integrating representations from CNN and BiLSTM-Transformer branches. Specifically, the CNN-derived feature maps from Mel-spectrogram processing are combined with the temporal features extracted by the BiLSTM-Transformer branch from MFCCs through concatenation, followed by further integration via fully connected layers. This strategy effectively addresses the structural differences between spectral and temporal feature representations while preserving their complementary characteristics. The experimental results (Table 1) validate the effectiveness of this approach. Compared with late fusion (78.92% accuracy, 78.87% F1-score) and hybrid fusion (85.35% accuracy, 84.30% F1-score) alternatives, the early fusion strategy achieves superiority, providing high performance (99.61% accuracy, 99.58% F1-score). The demonstrated performance advantages confirm that early fusion provides an optimal balance between information preservation and model complexity for transformer fault diagnosis applications, where high diagnostic accuracy is crucial.

Table 1. Performance of different fusion strategies

Fusion strategy	**Acc (%)**	**F1-score (%)**
Early fusion	99.61	99.58
Late fusion	78.92	78.87
Hybrid fusion	85.35	84.30

3.5. Application of the transformer acoustic diagnostic model

To achieve the classification of transformer fault categories, a feature fusion parallel optimization model based on deep learning is designed. The model uses the CNN and BiLSTM-Transformer in parallel to achieve feature fusion of Mel spectrograms and MFCCs extracted from raw acoustic data to optimize the original CNN and BiLSTM-Transformer, respectively. It can extract features more efficiently and improve the classification performance of the model.

The overall structure of the model is shown in Fig. 5, which mainly includes the CNN module, the BiLSTM-Transformer module, and the feature merging module.

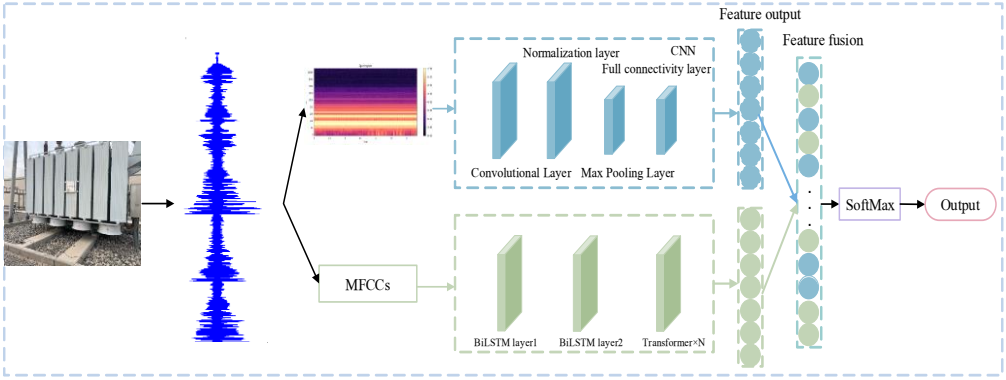

Fig. 5. Structure of the CNN-BiLSTM-Transformer model

After inputting the original voiceprint data into the model, the corresponding Mel spectrograms and MFCCs are obtained. Then, the obtained Mel spectrograms and MFCCs are input into the model. The CNN and BiLSTM-Transformers are used to extract and fuse features from the input. Finally, the SoftMax classifier is used to classify and output the final results.

4. Experiments and analysis of results

4.1. Collection of voiceprint data samples

To evaluate the accuracy of the method, a dataset consisting of 5 000 acoustic samples of 14 state categories was compiled. The data were collected from a pool of over 50 power transformers in operational substations and from specialized test platforms. Figure 6 shows the diagram of the sound pattern data monitoring device and field collection.

Importantly, each audio sample corresponds to a single state (normal or one specific fault) from a single transformer. The dataset is designed for state identification, meaning each sample is assigned a single label based on its dominant acoustic characteristic. This approach allows the model to learn distinct acoustic patterns for each condition, avoiding the complexity of multiple co-occurring faults.

The sampling rate of the sound data is 48 kHz. We collected 3 600 pieces of sound data during normal operation of transformers with 4 different voltage levels and 1 400 pieces of abnormal

sample data from 10 transformer defects. The sample data is randomly disturbed and divided into a training set and a test set in a ratio of 8:2. The distribution of the sample data and the corresponding labels are shown in Table 2.

Fig. 6. Voiceprint monitoring device and field deployment

Table 2. Sample distribution of the voiceprint

State types	Quantities	Labels	State	Quantities	Labels
110 kV Power transformer	900	0	Corona discharge	150	7
220 kV Power transformer	900	1	Sustained discharge	50	8
500 kV Power transformer	900	2	Loose clamps	100	9
±800 kV Power transformer	900	3	Motor shaft noise	150	10
Short-circuit shock	60	4	Fan noise	150	11
Surface discharge	50	5	Heavy overload	300	12
Intermittent discharge	40	6	DC polarisation	150	13

4.2. Experimental training process

The obtained raw acoustic signals are preprocessed to extract the Mel-spectrogram and MFCC features (in Section 1). To establish the dataset, all samples are randomly split into training and testing sets while strictly preserving the original class distribution (such as stratified sampling) with a ratio of 8:2. This ensures a fair representation of majority and minority classes in both sets.

The training dataset is fed into the established parallel two-channel model. To directly address the significant class imbalance (as shown in Table 1), a weighted cross-entropy loss function is adopted during training. The class weights are set to be inversely proportional to their frequencies in the training set, which increases the penalty for misclassifying samples from underrepresented fault categories. Iterative training is performed to adjust hyperparameters (e.g., learning rate, dropout rate). The training process continues until the performance of the model converges on a retained validation set and meets the expected diagnostic accuracy requirements.

The trained model is evaluated on the independent test set. Comprehensive metrics, including accuracy, precision, recall, and F1-score, are calculated to assess the performance of the model for all states, especially on the minority classes. The final diagnosis results are obtained and analyzed.

The multimodal fusion diagnostic model of the proposed parallel two-channel networks is shown in Fig. 7. The model is trained using the Adam optimizer, which has strong adaptive learning capability and low memory requirements. To avoid overfitting, a regularization method is

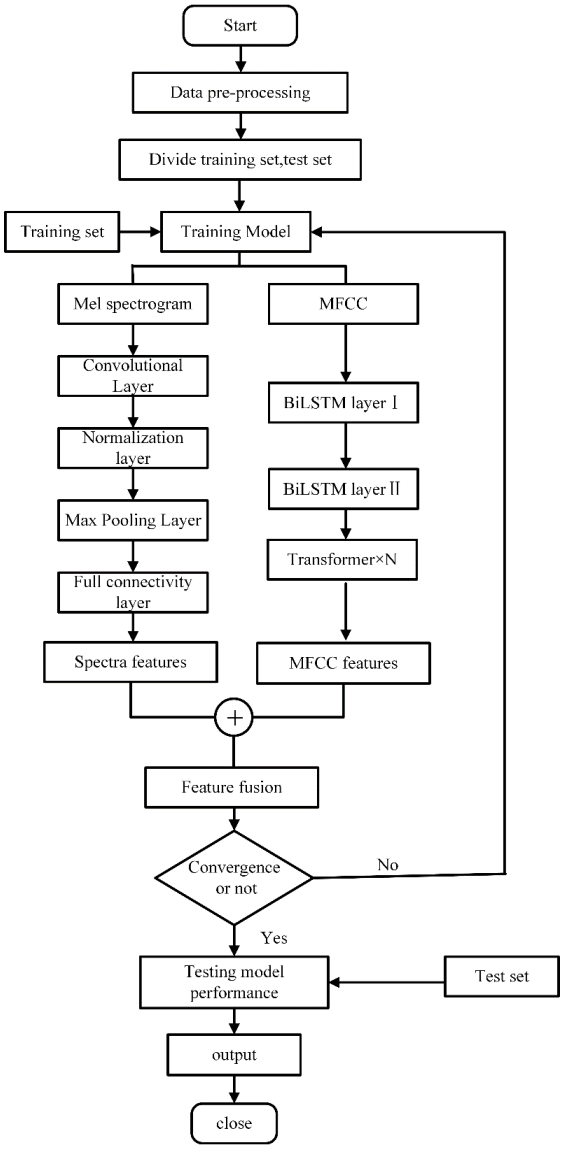

Fig. 7. Flow chart of fault diagnosis

introduced during the training. The dropout is set to 0.4; the cross-entropy is selected as the loss function; the initial learning rate is defined to be 0.001, and the number of training iterations is set to be 50. The specific parameters of the model are shown in Table 3.

Table 3. Model parameters

Modules	Parameters
Hidden layer dimension	64
Number of heads of multi-attention mechanisms	8
Head dimension	128
Feedforward neural network layer dimension	2 048
Encoder layer	3
Batch size	64

4.3. Experimental environment and configuration

The proposed model was implemented using Python 3.9 and PyTorch 2.0.1 frameworks. The experiments were conducted on a server equipped with an Intel Xeon Gold 6248R CPU, 256 GB RAM, and four NVIDIA GeForce RTX 3090 GPUs.

The acoustic data were acquired using a GRAS 46AE ¼ CCP Free-field Microphone Set (frequency response: 4 Hz–70 kHz; sensitivity: 50 mV/Pa) connected to a National Instruments NI-9234 sound and vibration input module.

The acoustic signals were collected from various power transformers, including SZ11-50 000/110 (110 kV, 50 MVA), SSZ11-180000/220 (220 kV, 180 MVA), OFPSZ-250000/500 (500 kV, 250 MVA), and ±800 kV UHVDC Converter Transformers (rated capacity: 400 MVA per valve side), during their normal energized operation in substations.

4.4. Experimental results and analysis

The mainstream acoustic algorithms first extract the MFCC and MEL spectrograms of the acoustic samples, and then use deep learning networks to identify the extracted features. To comprehensively evaluate this method, additional experiments were carried out; we performed 5-fold cross-validation to verify the stability of this model. The results show that the performance of all folds is consistent, with an average accuracy of 99.3% (±0.2%), demonstrating the robustness of this method.

To evaluate the practical applicability of this method, its performance under different noises was tested by adding Gaussian white noise to the test signals. Even at a low signal-to-noise ratio (SNR) of 10 dB, this model maintains high accuracy (> 98%), demonstrating its excellent noise robustness in real-world deployment scenarios.

The confusion matrix illustrating the performance of various fault detection algorithms on the test set is shown in Fig. 8. Several transformer acoustic pattern algorithms are selected for comparison. Figure 9 shows the relationship between the accuracy of different algorithm models on the test set and the number of training rounds. This model achieves optimal performance, exhibiting convergence speed and higher accuracy as the training rounds increase.

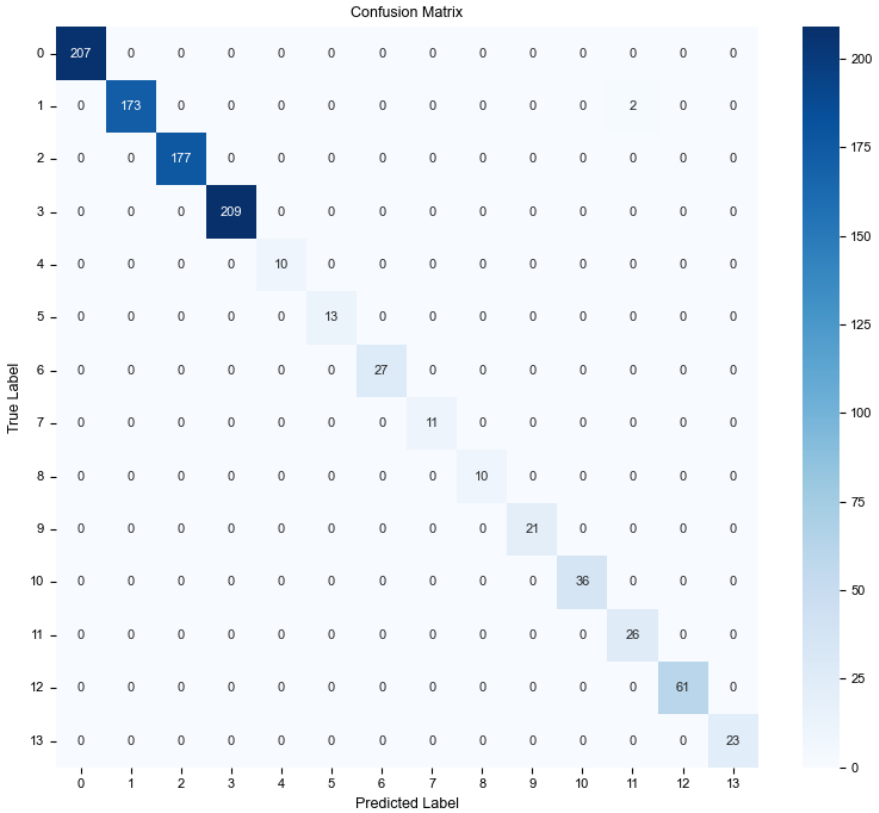

Fig. 8. Confusion matrix of the test set

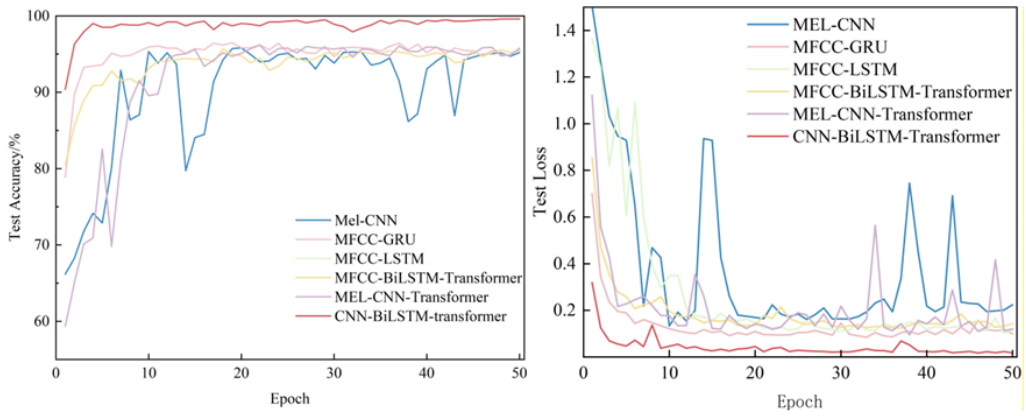

Fig. 9. Accuracy and loss value change curve of different algorithms

Table 4 shows different performance metrics of this algorithm model and other algorithms. The accuracy of this algorithm increases by using the MEL-CNN (4.77%), MFCC-GRU (5.56%), MFCC-LSTM (6.16%), MFCC-BiLSTM-Transformer (3.8%), and MEL-CNN-Transformer (3.9%), respectively. The other comparison metrics are also improved in different magnitudes.

Table 4. Comparison with classical algorithms

Networks	Acc/%	Recall/%	F1/%
Our model	99.61	99.56	99.58
CNN	94.84	95.32	95.80
GRU	94.05	94.70	94.80
LSTM	93.45	94.70	94.90
BiLSTM-Transformer	95.75	94.20	93.86
CNN-Transformer	95.73	95.56	95.67

5. Conclusion

A new fault diagnosis method is proposed, which involves feature extraction from Mel spectrograms using the CNN and MFCCs, applying the BiLSTM-Transformer. These features are then fused to achieve the diagnosis of potential faults. The findings demonstrate that:

1. The fusion of Mel spectrograms and MFCCs features can facilitate more comprehensive data characterization, leveraging information from different modes. By validating the datasets, the model can achieve an accuracy of greater than 99.5% in identifying normal operation and defective samples of level transformers;
2. The accuracy and F1 score of the classification model are 4% higher than the mainstream network models. Comparing the dual-channel fusion model with other single-feature extraction methods, we see that the model can combine the advantages of the CNN and BiLSTM-Transformer and shows high performance in fault diagnosis.

The proposed method contributes to transformer fault diagnosis by incorporating multi-level acoustic information. However, the recognition of fault types remains limited due to the constraints of the dataset size. Therefore, future efforts should focus on expanding the fault sample library and integrating acoustic features from more hierarchical levels to enable deeper analysis and more comprehensive research.

This study is subject to certain limitations. First, the model is trained and validated on a dataset with a fixed set of faults. Its ability to generalize novel or rare fault signatures not encountered during training remains to be further investigated; second, the non-trivial computational complexity of the deep learning model may hinder its deployment in real-time on resource-constrained edge devices within power grid systems.

Future research will focus on:

1. expanding the acoustic fault dataset to include a wider variety of fault conditions and operational environments;
2. exploring model compression and knowledge distillation techniques to develop lighter-weight versions of the diagnostic model for practical applications; and

3. investigating the model's capability in diagnosing multiple co-occurring faults within a single transformer unit.

Acknowledgements

The work was supported by the National Key Research and Development Program of China (No. 2022YFF0708400).

References

[1] Liao C.B., Yang J.X., Qiu Z.B., Hu X., Jiang Z.H., Li X., *Fault diagnosis of oil-immersed transformers based on missing data imputation*, High Voltage Engineering, vol. 50, no. 9, pp. 4091–4100 (2024), DOI: 10.13336/j.1003-6520.hve.20231532.

[2] Kang J.Y., Zhang S.X., Zhang Q.P., Gao B., Yan Z.H., Chen H.Z., *Fault diagnosis method of transformer based on ANOVA and BO-SVM*, High Voltage Engineering, vol. 49, no. 5, pp. 1882–1891 (2023), DOI: 10.13336/j.1003-6520.hve.20220630.

[3] Wang J.P., Xu G.L., Yan F.J., Wang J.J., Wang Z.S., *Defect transformer: An efficient hybrid transformer architecture for surface defect detection*, Measurement, vol. 211, 112614 (2023), DOI: 10.1016/j.measurement.2023.112614.

[4] Niu B., Wei Y., Zhang K., Yu Z., *An abnormal audio generation meth-od for fault diagnosis of power transformers*, 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, IEEE, pp. 1–5 (2025), DOI: 10.1109/ICASSP49660.2025.10887959.

[5] An K., Zhang Y., *LPViT: A transformer based model for PCB image classification and defect detection*, IEEE Access, vol. 10, pp. 42542–42553 (2022), DOI: 10.1109/ACCESS.2022.3168861.

[6] Zhou X., Yi K., Li G., Tian T., Yang X., *A transformer DGA fault diagnosis approach based on neighborhood rough set and AMPOS-ELM*, Journal of Electric Power Science and Technology, vol. 37, no. 3, pp. 157–164 (2022), DOI: 10.19781/j.issn.1673-9140.2022.03.019.

[7] Secic A., Aizpurua J.I., Garro U., Muxika E., Kuzle I., *Transformer OLTC operation monitoring framework through acoustic signal processing and convolutional neural networks*, IEEE Transactions on Instrumentation and Measurement (2025), DOI: 10.1109/TIM.2025.3550221.

[8] Wang F.H., Wang S.J., Chen S., Yuan G.G., Zhang J., *Transformer voiceprint recognition model based on improved MFCC and VQ*, Proceedings of the CSEE, vol. 37, no. 5, pp. 1535–1543 (2017), DOI: 10.13334/j.0258-8013.pcsee.152581.

[9] Zhou D.X., Wang F.H., Dang X.J., Zhang X., Liu S.G., *Dry Type Transformer Voiceprint Recognition Based on Compressed Observation and Discrimination Dictionary Learning*, Proceedings of the CSEE, vol. 40, no. 19, pp. 6380–6390 (2020), DOI: 10.13334/j.0258-8013.pcsee.191577.

[10] Zhang C.Y., Luo S.H., Yue H.T., Wang B.W., Liu Y.P., *Pattern Recognition of Acoustic Signals of Transformer Core Based on Mel-spectrum and CNN*, High Voltage Engineering, vol. 46, no. 2, pp. 413–423 (2020), DOI: 10.13336/j.1003-6520.hve.20200131005.

[11] Cui J.J., Ma H.Z., *Voiceprint recognition model of transformer core looseness fault based on improved MFCC and 3D-CNN*, Electric Machines and Control, vol. 26, no. 12, pp. 150–160 (2022), DOI: 10.15938/j.emc.2022.12.015.

[12] Liu Y.P., Wang B.W., Yue H.T., Gao F., Han S., Luo S.H., Zhang C.C., *Identification of Transformer Bias Voiceprint Based on 50Hz Frequency Multiplication Cepstrum Coefficients and Gated Recurrent Unit*, Proceedings of the CSEE, vol. 40, no. 14, pp. 4681–4694+4746 (2020), DOI: 10.13334/j.0258-8013.pcsee.191922.

[13] Nethala S., Chopra P., Kamaluddin K., Alam S., Alharbi S., Alsaffar M., *A deep learning-based ensemble framework for robust Android malware detection*, IEEE Access (2025), DOI: 10.1109/ACCESS.2025.3551152.

[14] Saarika K., Varsha V., Harsha P.S., Sheikh A.N., *Deep learning for automated image captioning: A CNN and transformer model analysis*, 2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), Bengaluru, India, IEEE, pp. 2160–2165 (2025), DOI: 10.1109/IDCIOT64235.2025.10914740.

[15] Zhang T., Feng G., Liang J., An T., *Acoustic scene classification based on Mel spectrogram decomposition and model merging*, Applied Acoustics, vol. 182, 108258 (2021), DOI: 10.1016/j.apacoust.2021.108258.

[16] Zhang S., Su F., Wang Y., Mai S., Pun K.P., Tang X., *A Low-Power Keyword Spotting System with High-Order Passive Switched-Capacitor Bandpass Filters for Analog-MFCC Feature Extraction*, IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 70, no. 11, pp. 4235–4248 (2023), DOI: 10.1109/TCSI.2023.3299855.

[17] Wang Y., Wang H., Li Z., Zhang H., Yang L., Li J., Wang Q., *Sound as a bell: A deep learning approach for health status classification through speech acoustic biomarkers*, Chinese Medicine, vol. 19, no. 1, 101 (2024), DOI: 10.1186/s13020-024-00973-3.

[18] Joysingh S.J., Vijayalakshmi P., Nagarajan T., *Significance of chirp MFCC as a feature in speech and audio applications*, Computer Speech & Language, vol. 89, 101713 (2025), DOI: 10.1016/j.csl.2024.101713.

[19] Yan Y., Simons S.O., van Bemmel L., Reinders L.G., Franssen F.M., Urovi V., *Optimizing MFCC parameters for the automatic detection of respiratory diseases*, Applied Acoustics, vol. 228, 110299 (2025), DOI: 10.1016/j.apacoust.2024.110299.

[20] Fahad M.S., Deepak A., Pradhan G., Yadav J., *DNN-HMM-based speaker-adaptive emotion recognition using MFCC and epoch-based features*, Circuits, Systems, and Signal Processing, vol. 40, pp. 466–489 (2021), DOI: 10.1007/s00034-020-01486-8.

[21] Ma H.Z., Wang J., Yang Q.F., Ni Y.M., *SMA-optimized SVM transformer state identification method based on acoustic vibration feature differentiation*, Electric Machines and Control, vol. 27, no. 10, pp. 42–53 (2023), DOI: 10.15938/j.emc.2023.10.005.

[22] Yu D., Zhang W., Wang H., *Abnormal voiceprint diagnosis method of oil-immersed transformer based on LSTM neural network*, Smart Power, vol. 51, no. 2, pp. 45–52 (2023).

[23] Yang J., Zhao J.M., Meng R.Q., Zhang D.X., Li B.Y., Wu Y.X., *Power system operation state identification based on particle swarm optimization and convolutional neural network*, Power System Technology, vol. 48, no. 1, pp. 315–324 (2024), DOI: 10.13335/j.1000-3673.pst.2022.2257.

[24] Feng S., Peng X.J., Chen J.N., Lu Y.W., Chen L., Hong X., Lei J.X., Tang Y., *Forced Oscillation Location and Propagation Prediction Based on Temporal Graph Convolutional Network*, Proceedings of the CSEE, vol. 44, no. 4, pp. 1298–1310 (2024), DOI: 10.13334/j.0258-8013.pcsee.222657.

[25] Messaoudi M., Kameli S.M., Refaat S.S., Abu-Rub H., Trabelsi, M., *Deep learning based corona discharge severity classification for high voltage equipment*, IECON 2024 - 50th Annual Conference of the IEEE Industrial Electronics Society, Chicago, IL, USA, IEEE, pp. 1–5 (2024), DOI: 10.1109/IECON55916.2024.10905579.

[26] Wang Y., Huang M., Zhu X., Zhao L., *Attention-based LSTM for aspect-level sentiment classification*, Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 606–615 (2016).

[27] Zeng J., Ma X., Zhou K., *Enhancing attention-based LSTM with position context for aspect-level sentiment classification*, IEEE Access, vol. 7, pp. 20462–20471 (2019), DOI: 10.1109/ACCESS.2019.2893806.

[28] Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Houlsby N., *An image is worth 16x16 words: Transformers for image recognition at scale*, arXiv preprint arXiv:2010.11929 (2020).

[29] Krueangsai A., Supratid S., *Effects of shortcut-level amount in lightwei-ght ResNet of ResNet on object recognition with distinct number of categories*, 2022 International Electrical Engineering Congress (iEECON), Khon Kaen, Thailand, IEEE, pp. 1–4 (2022), DOI: 10.1109/iEECON53204.2022.9741665.

[30] Wang H.X., Wang B., Dong X.Z., Yao L.Z., Zhang J.X., Ma H.R., *Semantic Difference and Performance Difference Analysis Method for Power Multimodal Data fusion*, High Voltage Engineering, vol. 50, no. 9, pp. 4037–4047 (2024), DOI: 10.13336/j.1003-6520.hve.20230490.

[31] Long K., Ma L., Gao Y., Yu G., *Feature stacking fusion in multimodal neural architecture search*, 2024 6th International Conference on Data-driven Optimization of Complex Systems (DOCS), Hangzhou, China, IEEE, pp. 414–419 (2024), DOI: 10.1109/DOCS63458.2024.10704481.