frontiers | Frontiers in Mechanical Engineering

TYPE Original Research
PUBLISHED 03 November 2025
DOI 10.3389/fmech.2025.1688439

OPEN ACCESS

EDITED BY
Mohamed Arezki Mellal,
University of Boumerdés, Algeria

REVIEWED BY
Kenneth E. Okedu,
Melbourne Institute of Technology, Australia
Jinxin Wu,
Guangxi University, China

*CORRESPONDENCE
Zepeng Lv,
lvzep_lzp@outlook.com

RECEIVED 19 August 2025
ACCEPTED 29 September 2025
PUBLISHED 03 November 2025

CITATION
Lv Z, Yang D and Yu B (2025) Empirical verification of a transformer voiceprint fault diagnosis method based on convolutional neutral network-long-short term memory and Mel gammatone cepstral coefficient features. *Front. Mech. Eng.* 11:1688439.
doi: 10.3389/fmech.2025.1688439

Empirical verification of a transformer voiceprint fault diagnosis method based on convolutional neutral network-long-short term memory and Mel gammatone cepstral coefficient features

Zepeng Lv*, Dongping Yang and Bin Yu

State Grid Shanxi Electric Power Company Ultra High Voltage Substation Branch Ultra High Voltage Beiyue Station, Taiyuan, Shanxi, China

Introduction: Transformers are core equipment in power grids. Their malfunctions may cause widespread power outages or even grid paralysis. Accurate diagnosis is of vital importance.
Methods: Aiming at the problem of insufficient accuracy of traditional voiceprint diagnosis techniques under complex working conditions, this paper proposes a transformer voiceprint fault diagnosis method that integrates CNN and LSTM. Through the series fusion of MFCC and GFCC and Fisher criterion screening, the MGCC characteristic parameters that take into account both accuracy and noise resistance are constructed for model input. Empirical tests were carried out on the voiceprint signals of three types of working conditions: normal transformer, loose winding and loose core.
Results: The results show that the fault recognition rate of this method for normal working conditions is 88%, the recognition rate for loose winding working conditions is 93%, and the recognition rate for loose core working conditions is 98%.
Discussion: Studies show that the transformer voiceprint fault diagnosis method based on CNN-LSTM network has high diagnostic accuracy and can meet the requirements of practical applications.

KEYWORDS

transformer, voice print, fault diagnosis, deep learning model, maintenance work

1 Introduction

As a fundamental energy industry in China, the power industry plays a positive role in both ensuring people's livelihood and promoting industrial development. To meet the needs of social development, the scale of power grids is constantly expanding, which poses significant challenges to the safe operation of power grids. As an important core device for energy conversion and transmission in the power industry, the operation status of transformers directly affects the safe and stable operation of the entire power grid. Once a fault occurs, it may not only cause large-scale power outages but may even lead to the paralysis of the power network. Therefore, conducting real-time monitoring and

precise fault diagnosis of transformers has crucial practical significance (Liu et al., 2022; Chen L. et al., 2022; Chen T. et al., 2022; Liu, 2022). At present, various technical paths have been formed in the field of transformer fault diagnosis. Common methods include spectral diagnosis, acoustic diagnosis, and oil chromatography diagnosis. Oil chromatography diagnosis technology has been maturely applied in the identification of insulation deterioration type faults. However, it relies on the accumulation of characteristic gases in the oil and has an obvious lag, making it difficult to achieve real-time early warning. It has extremely low sensitivity to mechanical faults such as loose windings and core vibration. Although spectral diagnosis technology can directly reflect the state of insulating materials, it is limited by the high cost of equipment and the strict requirements for the on-site environment, and its application scenarios are relatively limited.

In contrast, acoustic diagnostic technology, with its non-invasive detection advantage, can directly capture the continuous noise generated by devices such as winding vibration, core vibration, and cooling fans during the operation of transformers. These noises contain extremely rich information about the equipment status and will radiate to the surroundings through the internal structure of the transformer and the air. To a large extent, they can reflect the actual operating status of the transformer and thus show unique potential in mechanical fault diagnosis (Shao et al., 2024; Song et al., 2023). However, the conventional voiceprint diagnosis mode reveals obvious deficiencies in complex working conditions: On the one hand, environmental interferences at the substation site (such as noise from other equipment operations and external traffic noise) can pollute the voiceprint signal. Although traditional feature extraction parameters, such as a single Mel cepstral coefficient (MFCC), are mature and widely applied, they are affected by the masking effect when the frequency of pure sound is close to that of mixed noise, and their ability to represent sound signals decreases significantly, resulting in a marked reduction in classification accuracy. On the other hand, the existing diagnostic schemes based on deep learning also have limitations. A single convolutional neural network (CNN) is good at extracting local spatial features and is widely used in fields such as image classification and fault diagnosis, but it cannot effectively mine the temporal correlation of fault development in voiceprint signals. A single long-short-term memory network (LSTM) can handle time series data and solve the long-term dependency problem of recurrent neural networks (RNNS) but is insufficient in extracting local subtle features (such as the weak vibration differences of early faults). Even mainstream hybrid models such as CNN-GRU, which extract spatial features through convolutional networks and introduce gated recurrent units (GRUs) to model temporal dependencies, have limitations in their memory ability for long sequence voiceprint signals. The transformer model performs well in global time series modeling but lacks sensitivity to local features and has a high computational complexity, making it difficult to adapt to edge computing scenarios in the actual operation and maintenance of substations (Chen et al., 2025; Zhu et al., 2025).

In view of the problems of low diagnostic accuracy, weak anti-interference ability, and difficulty in balancing temporal and spatial features of the existing technologies under complex working conditions, this article takes the optimization and upgrading of acoustic diagnosis technology as the core, focuses on the complementary advantages of CNN and LSTM, and proposes a transformer voiceprint fault diagnosis technology based on a CNN-LSTM network. This technology integrates the accuracy of MFCC with the noise immunity of gammatone frequency cepstral coefficients (GFCCs) to construct the fusion feature parameter mel gammatone cepstral coefficient (MGCC) to enhance feature robustness. Then, it uses a CNN to extract the local spatial features of the voiceprint signal (such as high-frequency vibration components related to faults). By using LSTM to capture the long-term dynamic patterns of fault development, a complete diagnostic chain of "feature fusion—spatial extraction—time series modeling" is formed. This article will empirically test the feasibility of the CNN-LSTM diagnostic technology in transformer fault diagnosis, aiming to address the limitations of existing technologies, further improve the monitoring effect of transformer faults, and provide reliable technical support for the engineering application of transformer fault diagnosis.

2 Basic concepts

2.1 Transformer fault diagnosis technology based on voice print recognition

Voice print recognition is the process of extracting the speaker's speech signal, comparing it with a previously trained speech set, and confirming the speaker's identity. Because the sound emitted by the transformer is similar to the voice of the speaker, both of which contain a large amount of state information, the voice print recognition technology is applied to the field of transformer fault diagnosis, as shown in Figure 1.

As can be seen from Figure 1, this diagnostic model consists of three modules: sound acquisition and processing, feature extraction, and complex type recognition (Feng, 2019). The basis of voiceprint recognition technology is the establishment of a voiceprint library. The sound signals during the operation of the transformer are collected by means of the microphone sensor, and then the sound signals are amplified by the preamplifier and imported into the data acquisition card for digital-to-analog conversion. The converted data are then transmitted to the computer for analysis. At present, there are two commonly used test methods for sound signals: the sound pressure method and the sound strength method. In the sound signal feature extraction method, according to the audible sound frequency range and sampling theorem, the sampling frequency must be above 40 kHz to ensure that the sound information is not lost, but this means that the 1s sound signal needs to save 4 plays of data. Therefore, in order to speed up the calculation speed and improve the calculation performance, it is necessary to extract the transformer sound signal for special diagnosis. Commonly used sound signal feature quantities include mean value, method, standard deviation, root mean square, MFCC, and GFCC. Among them, MFCC and GFCC are the mainstream feature extraction parameters at present (Li et al., 2022; Zhang et al., 2022). The Mel cepstrum feature parameter is a mature and widely used acoustic feature parameter in voice print recognition. Some researchers have pointed out that there is a nonlinear correspondence between the frequency of the

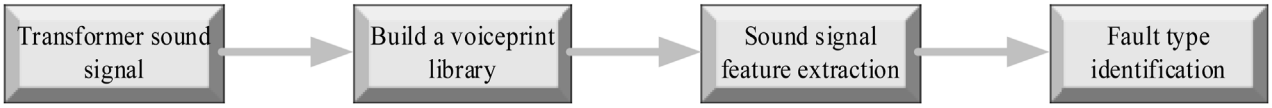

FIGURE 1
Fault diagnosis technology based on voice print recognition.

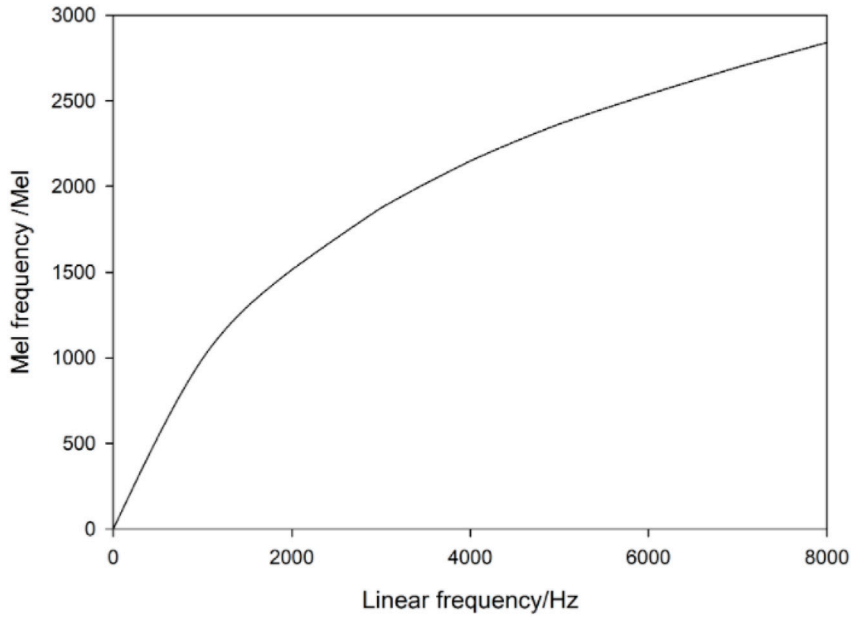

FIGURE 2
Relationship to actual frequency.

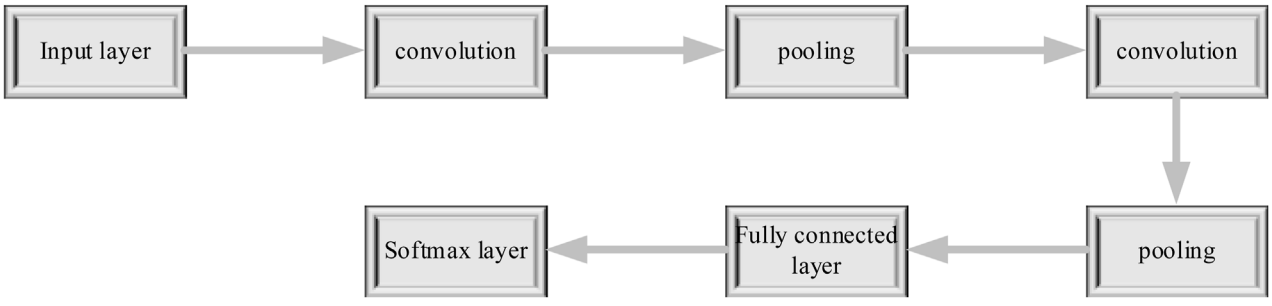

FIGURE 3
Convolutional neural network structure.

sound signal heard by the human ear and its actual frequency. If the frequency of the sound signal is lower than 1000 Hz, the auditory perception ability of the human ear has a linear relationship with the actual frequency, and the counter-reaction has a logarithmic relationship. This nonlinear logarithmic relationship is called the Mel scale. The relationship with frequency is shown in Formula 1.

$$\mathrm{Mel}(f) = 2595 \lg(1 + f/700) \quad (1)$$

Where f represents the frequency and Mel (f) represents the Mel frequency. The relationship between the actual frequency is shown in Figure 2.

When the frequencies of pure sound and mixed noise are similar, the human ear cannot distinguish between the two signals due to the masking

effect. The critical bandwidth that enables people to hear pure sound signals simply by reducing the bandwidth is called the critical bandwidth, which is calculated as shown in Formula 2.

$$BW_c = 25 + 75 \times \left[1 + 1.4 \times \frac{f_c}{1000}\right]^{0.69} \quad (2)$$

Where f_c is the central frequency. A set of Mel filter banks can be constructed from the center frequency to simulate the bandwidth frequency. In practice, triangular filters are usually used to approximate the equivalence of this set of Mel filters. The transfer function for each triangular filter is defined as shown in Formula 3.

$$H_m(k) = \begin{cases} 0 & k < f_{m-1} \\ \dfrac{k - f_{m-1}}{f_m - f_{m-1}} & f_{m-1} < k < f_m \\ \dfrac{f_{m+1} - k}{f_{m+1} - f_m} & f_m < k < f_{m+1} \\ 0 & k > f_{m+1} \end{cases} \quad m = 0, 1, 2 \cdots, M-1 \quad (3)$$

Where m is the order of the Mayer filter, determined by the cutoff frequency of the sound signal, and M = 24, f_m is the center frequency of the Mayer filter after comprehensive analysis. Due to complex and changeable environmental factors, MFCCs cannot properly characterize sound signals, resulting in a significant decline in the classification accuracy. Therefore, the GFCC characteristic parameters were developed to track the trend, mimicking the function of the human ear through simulation. The expression of the J-filter is shown in Formula 4.

$$G_j(t) = At^{a-1}e^{-2\pi b_j t}\cos\left(2\pi f t_j + \varphi_j\right), \quad t \geq 0, 1 \leq j \leq N \quad (4)$$

In the formula, A and f represent the gain and gravity center frequency of the gammatone filter, respectively. φ, a, and N represent phase, order, and number, respectively. E represents the attenuation factor, which is related to the length of the control filter, the specific expressions are shown in Formulas 5 and 6:

$$b_j = 1.019ERB\left(f_j\right) \quad (5)$$

$$\mathrm{ERB}\left(f_j\right) = 24.7\left(\frac{4.37 f_j}{1000} + 1\right) \quad (6)$$

Where $\mathrm{ERB}(f_j)$ represents the equivalent rectangular bandwidth. The center frequency of each filter is equally divided on the ERB scale and then mapped to the linear scale.

2.2 Convolutional neural network architecture

A convolutional neural network (CNN), one of the most commonly used mainstream deep learning algorithms, supports multi-layer neural networks, which can simultaneously learn feature parameters and classifiers. CNNs are widely used in image classification, face recognition, and fault diagnosis (Shi et al., 2022). The structure of a CNN is similar to that of a multi-layer perceptron, and its basic network structure consists of five parts: first, the input layer; second, a convolutional layer; third, the pooling layer; fourth, the fully connected layer; fifth, the output layer, the structure of the convolutional neural network is shown in Figure 3:

3 Basic framework of a CNN-LSTM network

3.1 Long-short term memory network architecture

A special case of recurrent neural network (RNN), a long-short term memory network (LSTM) has obvious advantages in processing time series by solving the long-term dependence problem of RNNs through memory units (Cao et al., 2023; Liu et al., 2019). The framework is shown in Figure 4.

It can be seen that an LSTM is designed based on a gate circuit. Compared with an RNN, an LSTM adds three architectures: input gate, output gate, and forgetting gate. These three parts control the flow direction of short-duration information and effectively screen and update information. Through the forgetting gate, it can be known whether the state C_{t-1} of the previous moment exists in the current moment state C_t. If the output of the forgetting function f_t is 0, then the information in C_{t-1} is completely forgotten; if the output of f_t is 1, then the information in C_{t-1} is completely passed. The corresponding forgetting formula is shown in Formula 7.

$$f_t = \sigma\left(A_f\left[h_{t-1}, x_t\right] + b_f\right) \quad (7)$$

It can be known through the input gate whether x_t is saved to the current moment state C_t. Under the processing of the tanh function, the hidden states h_{t-1} and x_t at the previous moment will be transformed into the candidate state C_t^*. i_t is the output of the σ function that takes h_{t-1} and x_t as input values, and the value of i_t is selected from 0 to 1. The expression of the input function is shown in Formulas 8, 9:

$$i_t = \sigma\left(A_i\left[h_{t-1}, x_t\right] + b_i\right) \quad (8)$$

$$C_t^* = tan\,h\left(A_c\left[h_{t-1}, x_t\right] + b_c\right) \quad (9)$$

Then, after updating the calculation formula from the state C_{t-1} of the previous moment to the state C_t of the current moment, it is shown in Formula 10.

$$C_t = f_t \times C_{t-1} + i_t \times C_t^* \quad (10)$$

The proportion of C_t in the hidden state h_t at the current moment can be known through the output gate. Output h_{t-1} and x_t through σ to obtain O_t, and multiply C_t and O_t processed by tanh to get h_t. The corresponding calculation formulas are shown in Formulas 11, 12:

$$O_t = \sigma\left(A_o\left[h_{t-1}, x_t\right] + b_o\right) \quad (11)$$

$$h_t = O_t \times tan\,h\left(C_t\right) \quad (12)$$

In the LSTM neural network, the weights are represented by A, and the bias parameters are represented by b.

3.2 Transformer diagnostic model architecture under a CNN-LSTM network

In the current research, the hybrid model has received extensive attention due to its multimodal feature extraction ability. Mainstream methods, such as CNN-GRU, extract spatial features through convolutional networks and introduce gated recurrent units (GRUs) to model temporal dependencies, but their memory ability

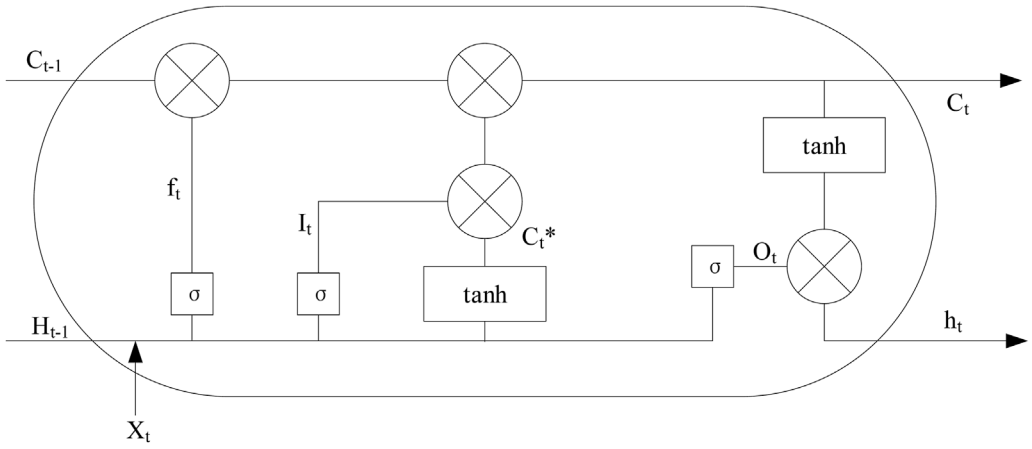

FIGURE 4
LSTM architecture.

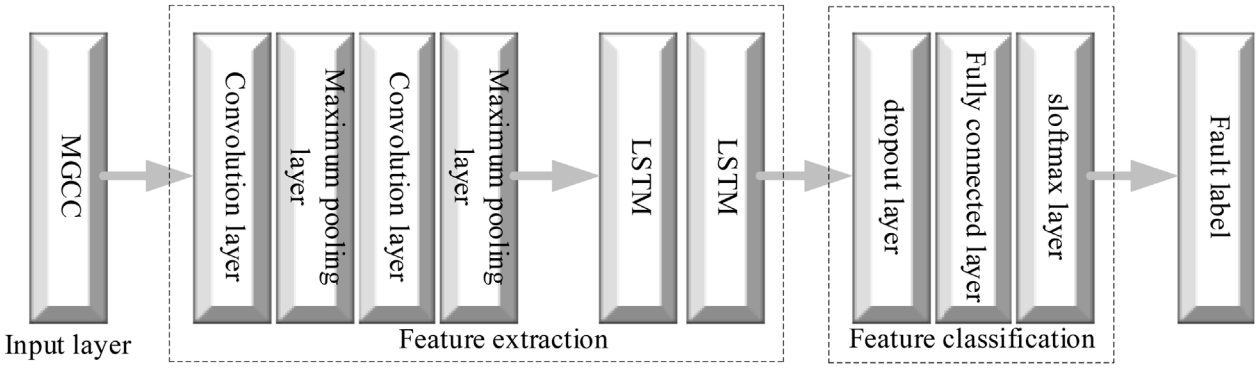

FIGURE 5
CNN-LSTM network model architecture.

for long sequences is limited. Although the transformer performs well in global time series modeling, it is insufficiently sensitive to local features and has a high computational complexity. In contrast, the combination of CNN-LSTM has unique advantages: it has the complementarity of spatial–temporal features. CNNs and LSTMs are both mainstream algorithms in deep learning, but they have different advantages. The advantage of a CNN lies in the extraction of local features, which can realize the extraction of spatial features of the dry substation noise signal. An LSTM is more commonly used to process text sequences of varying lengths and can complete the extraction of time characteristics of dry transformer noise signals. This work combined the advantages of these two algorithms to make up for each other, cooperate to complete the transformer fault diagnosis, and improve the diagnosis efficiency and accuracy. The specific combination methods are as follows: First, the gammatone cepstral coefficient (MGCC) feature parameters of the transformer noise signal calculated in the first part are input into the CNN, and the CNN carries out feature extraction. Second, the output of the CNN is taken as the input of the LSTM, and the second time sequence signal is extracted for the features. Third, after the transformation of the flatten layer and the smooth integration of the full connection layer, it is sent to the softmax layer to complete the sub-classification of transformer faults in different working conditions. The specific architecture is shown in Figure 5.

To better test the inspection of the substation's voice print fault under the CNN-LSTM network, the sound signals under three different working conditions of transformer A: normal, winding fault, and iron core fault are intercepted, the corresponding parameters are set, and the sample data of transformer fusion feature parameter MGCC are transmitted to LSTM for training and identification. Suppose the LSTM has only one layer, then add the fully connected layer and the flatten layer. After completing the parameter settings and conducting experiments, the model confusion matrix can be obtained, as shown in Figure 6.

It is not difficult to find that the recognition rate of the transformer under normal operating conditions is 72%. The recognition rate when the winding loosening problem occurs is 82%. The recognition rate when the core was loose is 82%, and the average overall recognition rate is 78.6%.

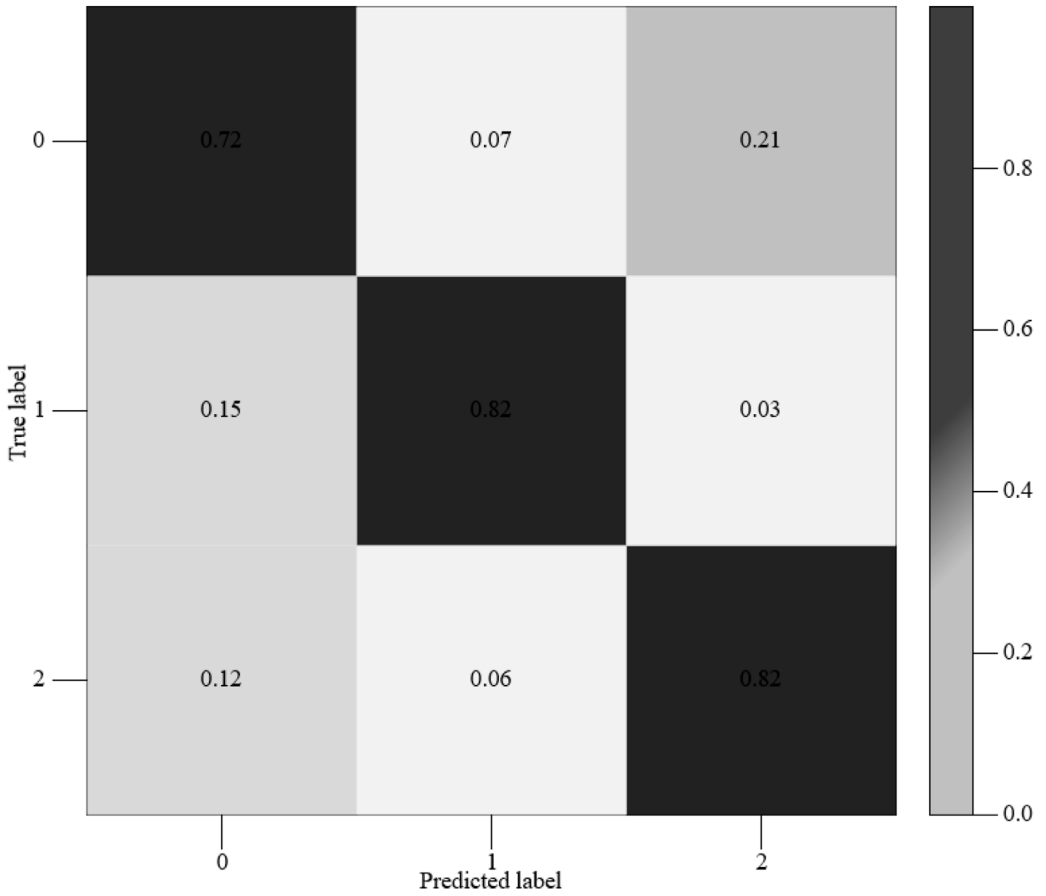

FIGURE 6
Confusion matrix diagram of the LSTM network model.

3.3 Inspection results of the substation voice print fault under the CNN-LSTM network

3.3.1 Parameter setting

The frequency used is 51,200 Hz, and the transformer sound signal sample is pre-weighted, frame divided, and window added. The pre-weighted system is 0.9375, the frame length is 10 ms, the frame shift is 5 ms, and the window function is a Hamming window. The pre-processed transformer sound signal is passed through the 24-dimensional Mel filter and the 64-dimensional Gammatone filter bank, respectively, and then the DCT transformation is performed to obtain the 12-dimensional 2 MFCC feature parameters and the 30-dimensional GFCC feature parameters. Finally, the two feature parameters are fused in series, and the contribution degree of each dimension of the fusion feature parameters is calculated using the Fisher criterion. The first 12 dimensions with the greatest contribution degree are selected to form the fusion feature parameter MGCC after dimensionality reduction.

To clarify the reliability and traceability of the experimental data, the core information of the voiceprint dataset used in this experiment is summarized in the following table. Supplement the key details such as the data source, collection environment and sample size distribution, as shown in Table 1.

Conduct experiments based on the above dataset. To meet the needs of different working conditions, 450 MGCC samples with fused characteristic parameters were selected. Among them, 70% of the samples (315) were selected for the training of the CNN network model, and the remaining 30% of the samples (135) were used for accuracy testing. The relevant parameters were set for the CNN network, such as the batch size being 200, the learning rate being 0.0001, the number of iterations being 60, and the number of LSTM network layers being 1. The Adam optimization algorithm was used for verification. The influence of different numbers of neurons on the recognition rate was discussed, and the obtained results are shown in Table 2.

It can be seen that when other parameters remain unchanged, the recognition rate of the CNN-LSTM network increases with the increased number of LSTM neurons. When the recognition rate reaches 92.65%, it is meaningless to continue to increase the number of neurons, and doing so may even lead to overfitting problems. Therefore, it is most appropriate to set the number of LSTM neurons as 5, and the input layer dimension of the transformer noise signal feature parameter MGCC is 12 × 19. After passing through two convolutional pooling layers and one LSTM layer, respectively, the extracted feature information is sent to the fully connected layer for

TABLE 1 Basic information table of the transformer voiceprint dataset.

Data attribute	Specific content
Data source	On-site collection of current A model transformers at 110 kV outdoor substations in X area + laboratory simulation fault verification (ensuring data coverage of actual operation and maintenance scenarios and standard fault conditions)
Collection object	Model A oil-immersed distribution transformer (rated capacity 500 kVA, rated voltage 10 kV/0.4 kV, operating life 3 years, no historical major fault records, meeting the common specifications of power grid operation and maintenance equipment)
Collection environment	1. On-site collection environment: A windproof and soundproof cover is installed in the designated area of the outdoor substation, 2 m away from the transformer body (to avoid near-field vibration interference), to reduce the influence of environmental wind noise and noise from the cooling fans of adjacent equipment. The background noise value of the environment is 35–45 dB (measured by a sound level meter). 2. Laboratory simulation environment: Shielded acoustic laboratory, background noise ≤25 dB, simulating typical loads of a substation (30%, 50%, and 80% rated load)
Collection equipment	1. Microphone: 48 kHz sampling rate capacitive acoustic sensor (model: CM-100), sampling accuracy 24 bits, frequency response 20 Hz–20 kHz (covering the effective frequency band of transformer voiceprint) 2. Data acquisition card: USB-6363 (16-bit analog input accuracy, supporting multi-channel synchronous acquisition) 3. Auxiliary equipment: Sound level meter (TES-1353), wind shield, tripod
Collect parameters	The sampling frequency is 51,200 Hz (in accordance with the requirement of Section 2.1 that "the sampling frequency should be ≥ 40 kHz to avoid loss of sound information"), the collection duration for each sample is 5 s (each sample contains 500 frames of voiceprint data, with a frame length of 10 ms), and the collection is repeated 30 times per working condition (to ensure sample diversity)
Operating condition category	1. Normal operating conditions (no abnormal vibration of the transformer, and the temperatures of the insulating oil and windings are both within the rated range) 2. Winding loose condition (simulating a 2-mm loose winding wire, achieving the standard fault state through mechanical debugging) 3. Core loosening condition (simulating a 1 mm loosening of core silicon steel sheets, which conforms to common core failure modes)
Sample size distribution	The total sample size is 450 (MGCC fusion feature samples), with a balanced distribution of 150 samples in each of the three working conditions (to avoid sample bias affecting the model training effect) 1. Normal operating conditions: 150 (including 50 on-site collected samples +100 laboratory verification samples) 2. Winding loosening condition: 150 (including 50 on-site simulated fault samples +100 laboratory standard fault samples) 3. Core loosening condition: 150 (including 50 on-site simulated fault samples +100 laboratory standard fault samples)
Data preprocessing	Pre-emphasis (coefficient 0.9375), framing (frame length 10 ms and frame shift of 5 ms), windowing (Hamming window), FFT (512 points), MFCC extraction (24-dimensional Mel filter → 12-dimensional DCT), GFCC extraction (64-dimensional gammatone filter → 30-dimensional DCT), MGCC fusion (series + Fisher criterion screening to 12-dimensional)
Data storage format	WAV format (original voiceprint signal), CSV format (MFCC/GFCC/MGCC characteristic parameters), storage path associated with working condition labels (such as "Normal_20231001_0800.wav," "Winding_Loose_MGCC_001.csv") facilitate data traceability and model invocation

TABLE 2 Influence of different numbers of neurons on recognition results.

Number of neurons	3	4	5	6	7	8
Recognition rate	91.78	92.23	92.65	89.47	88.35	88.41

classification and recognition. The full connection layer is also only 3, which means that the number of types of working conditions in the transformer noise signal is 3, and the total training parameter is 30,799, which is 50% higher than the traditional CNN network structure parameter.

3.3.2 Convolutional neutral network-long-short term memory network transformer voice print fault diagnosis test results

Based on the setting of relevant parameters, the transformer operating state is identified here, and the results are shown in Figure 7.

Figure 7 is the confusion matrix diagram after the CNN-LSTM classification model trained the mixed characteristic parameters MGCC of the transformer under three working conditions. The states 0, 1, and 2 in the figure represent the three mechanical states of the transformer: normal, loose winding, and loose iron core, respectively. Among them, the recognition rate in the 0 state is 89%; The recognition rate in state 1 is 94%. The recognition rate in

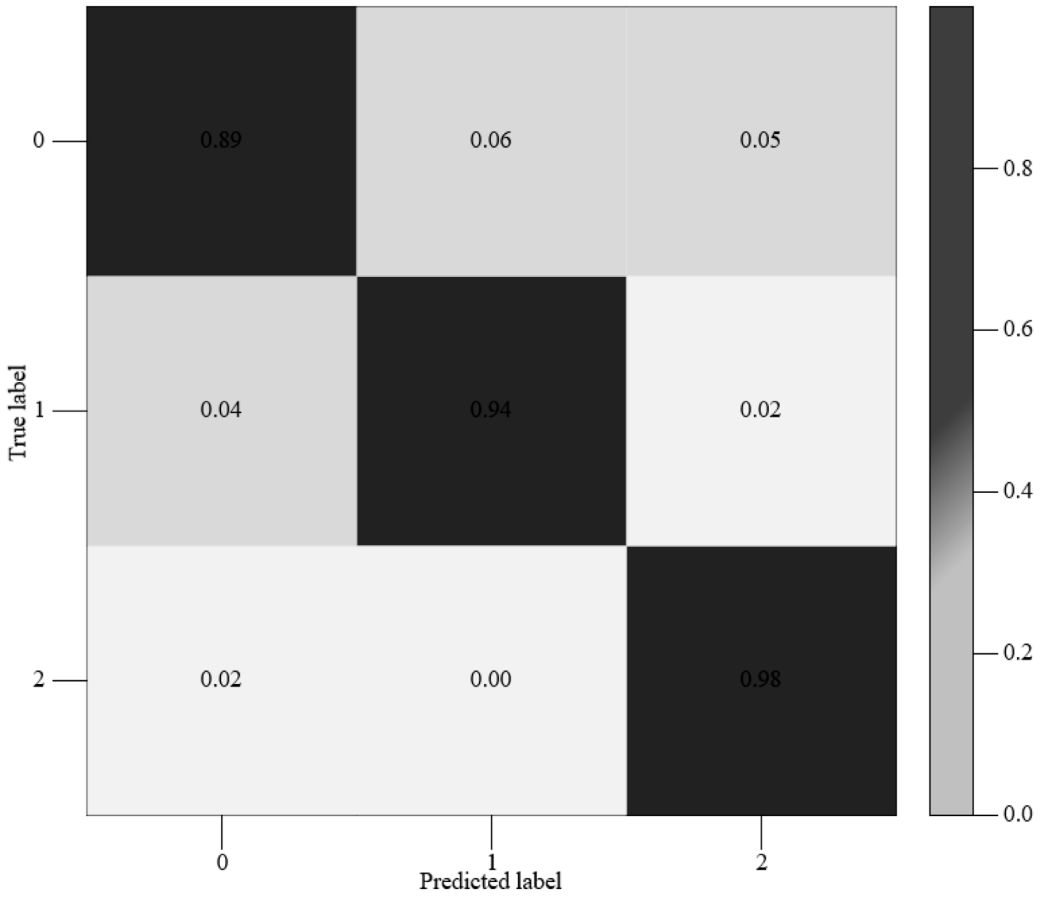

FIGURE 7
Confusion matrix diagram of CNN-L.

state 2 is 98%. It can be seen that the recognition rate of transformer iron core loosening is the highest, followed by winding loosening, and the recognition rate under normal working conditions is the lowest; that is, samples under normal working conditions are easily confused with winding loosening or iron core loosening, and the average recognition rate reaches 93.66%, achieving the expected goal. At the same time, compared with LSTM, the accuracy is improved by 14.9%, which confirms the feasibility of the CNN-LSTM network transformer voice print fault diagnosis technology.

3.4 The influence of different signal-to-noise ratios on transformer operating status identification results

In the practical transformer voice print fault diagnosis system, the collected sound signals are often disturbed by the surrounding environment, and whether the extracted feature parameters have good anti-noise property will have a great impact on the result of the voice print diagnosis system. Therefore, the following experiments were carried out to further test the anti-noise ability of MGCC. In this article, factory noise, speaking noise, vehicle noise, and white Gaussian noise were extracted from the Noisex-92 noise library. All the data lengths are 235 s, and the adoption rate is 19,980 Hz. The signal-to-noise ratios of these noises are set as −10 dB, −5 dB, 0 dB, 5 dB, and 10 dB, respectively. Then, background noise with a different signal-to-noise ratio was added to the original signal, and the MFCC, GFCC, and MGCC characteristic parameters of the transformer under three working conditions were obtained. These three characteristic parameters were input into the CNN network for classification and recognition, and the results obtained are shown in Figure 8.

It can be seen that under a high signal-to-noise ratio (SNR), the recognition rate of MFCC characteristic parameters is higher than that of GFCC. With the decrease in the SNR, the recognition rate of MFCC increases and the noise resistance is weaker, while the recognition rate of GFCC decreases slowly and the noise resistance is better. However, the accuracy of MGCC fusion parameters is higher than that of MFCC and CFCC, and the recognition rate tends to decline slowly as the signal-to-noise ratio decreases. Therefore, the MGCC feature parameters not only have the accuracy of MFCC feature parameters but also have the noise resistance of GFCC feature parameters, which is the most suitable feature parameter in the transformer fault diagnosis model.

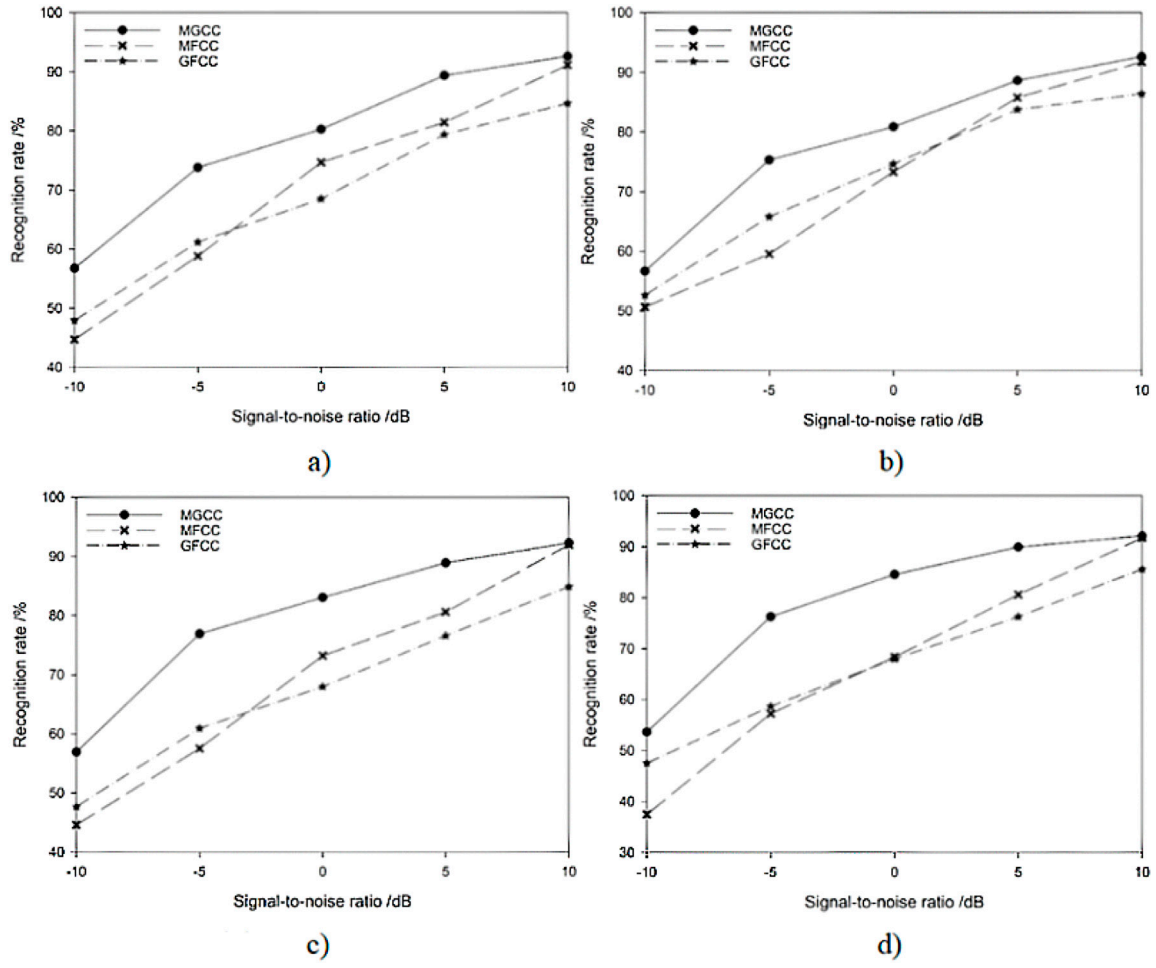

FIGURE 8
Identification results of feature parameters under different noise backgrounds. **(a)** Factory noise, **(b)** voice, **(c)** vehicle noise, and **(d)** Gaussian white noise.

TABLE 3 K-fold cross-validation.

Fold	Normal situation	Winding is loose	Core is loose	Average recognition rate
1	86%	92%	97%	91.7%
2	88%	94%	98%	93.3%
3	85%	91%	96%	90.7%
4	89%	93%	97%	93.0%
5	87%	95%	98%	93.3%
Mean ± standard deviation	87.0% ± 1.4%	93.0% ± 1.6%	97.2% ± 0.8%	92.4% ± 1.1%

3.5 K-fold cross-validation

Combined with the parameter settings, K-fold cross-validation is carried out to ensure the stability of the subsequent test results. The original data set (450 samples) is randomly divided into k subsets, and each fold contains approximately 90 (K = 5) samples. For the i-th fold (i = 1,2,. . ., K), the i-th subset is used as the validation set, and the remaining K-1 subsets are merged into the training set. The recognition rate of each fold is recorded, the average accuracy rate and standard deviation are calculated, and the stability of the model is evaluated. The results obtained after inspection are shown in Table 3.

It can be seen that the original fixed division result (average 93%) is close to the cross-validation mean (92.4%), and the standard

deviation is relatively low (±1.1%), indicating that the model has good stability, and the credibility of the original conclusion is relatively high.

4 Conclusion

To sum up, this article first describes the basic concepts of transformer vowel print fault diagnosis technology, including transformer sound acquisition and processing, sound signal feature extraction, and transformer fault type identification. Especially in the feature extraction link, detailed analysis and explanation are given to the two mainstream feature extraction parameters, MFCC and GFCC. The discussion lays a foundation for the subsequent empirical development. Second, the CNN network architecture and LSTM network architecture are given, and then the transformer fault diagnosis model architecture under the CNN-LSTM network is constructed. Finally, taking three different acoustic signals of transformer A, normal, winding fault, and iron core fault, as the research object, the transformer voice fault diagnosis technology based on a CNN-LSTM network is tested. The results show that under the CNN-LSTM classification model, the recognition rate of the transformer in state 0 is 88%. The recognition rate in state 1 is 93%. The recognition rate in state 2 is 98%. In the LSTM mode, the recognition rate of the transformer in state 0 is 73%. The recognition rate in state 1 is 83%. The recognition rate in state 2 is 83%, and the overall average is 79.3%. The feasibility of the CNN-LSTM-based network transformer voice print fault diagnosis technology is confirmed, and it can be used in practice.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

ZL: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review and editing. DY: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review and editing. BY: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review and editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. The authors declare that this study received funding from State Grid Shanxi Electric Power Company Science and Technology Project. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article, or the decision to submit it for publication.

Conflict of interest

Authors LY, DY, and BY were employed by State Grid Shanxi Electric Power Company Ultra High Voltage Substation Branch Ultra High Voltage Beiyue Station.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Cao, L. T., Wei, H. B., Huang, Z., and Shi, M. L. (2023). LSTM neural network based reactor fault voiceprint recognition method. *J. Zhejiang Electr. Power* 2023(4), 114–120. doi:10.19585/j.zjdl.202304014

Chen, L. W., Li, J. H., and Jiang, X. (2025). Efficient multi-dimensional time series anomaly detection model based on graph convolution and transformer. *Ind. Control Comput.* 38(08), 109–110+113.

Chen, L., Yu, H., Qi, B., and Li, B. (2022). Principal component analysis and random forest algorithm fusion of transformer fault diagnosis method. *J. Transformer* 59(07), 23–28. doi:10.19487/j.cnki.1001-8425.2022.07.004

Chen, T., Leng, H. W., Li, X. S., and Chen, Y. F. (2022). Hierarchical fault diagnosis model of transformer based on gas analysis in oil and class overlap characteristics. *China Electr. Power* 55(07), 22–32+41.

Feng, Y. (2019). Measurement and analysis of transformer fault diagnosis based on voice print recognition technology. *Sci. Technol. Innov. Appl.* 13(06), 73–76.

Li, W., Zeng, F. Y., Wang, B., and Chen, Z. B. (2022). Application of voice print recognition technology based on MFCC weighted dynamic feature combination in underground cable protection. *Pow. Inf. Commun. Technol.* 20(5), 16–22. doi:10.16543/j.2095-641x.electric.power.ict.2022.05.003

Liu, L. R. (2022). Application of photoacoustic spectroscopy to gas monitoring in transformer oil. *Light Source Illumin* 2022(03), 86–88.

Liu, X. X., Ji, Y., and Liu, C. P. (2019). Voice print recognition based on LSTM neural network. *Comput. Sci.* 48(S2), 270–274.

Liu, Z. Y., Shi, Y. Y., Wang, W. R., Zeng, J. L., and Hu, S. N. (2022). Research on optoacoustic spectrum fault diagnosis system of transformer based on edge computing. *Electron. Devices* 45(04), 872–877.

Shao, Y. Y., Wang, X., and Peng, P. (2024). Intelligent diagnosis method for dry-type transformer status based on machine vision and auditory response. *Noise Vib. Control* 44(04), 199–204+235.

Shi, S. D., Huang, J. J., Zhu, X. X., Wang, Y., and Qian, B. Y. (2022). Based on voice print SDP - CNN transformer partial discharge pattern recognition. *Electr. Power Inf. Commun. Technol.* 20(10), 105–112.

Song, C., Xia, X., and Wang, X. Y. (2023). Transformer acoustic feature extraction and fault identification based on MFCC and CNN. *Electr. Eng.* 19(06), 49–54. doi:10.16543/j.2095-641x.electric.power.ict.2022.10.013

Zhang, K., Yang, K. J., Huang, W. L., Wang, C. L., Ji, K., Zhu, T. Y., et al. (2022). Based on the validation and test method of transformer condition of voiceprint recognition system. *Electr. Power Inf. Commun. Technol.* 9(01), 1–6.

Zhu, H. Y., Chen, G., and Li, M. G. (2025). Noise reduction method for network security alarm information of power monitoring system based on three-stage fusion of rules, statistics and transformer. *Hunan Electr. Power* 45(04), 126–132. doi:10.16339/j.cnki.jsjsyzdh.202201001