

# Voiceprint recognition method of transformer based on LBT-ODF and MVN

Jiaqi Peng<sup>1</sup>, Yulin Ma<sup>2</sup>, Haiping Ye<sup>3</sup>, Xianrui Che<sup>4</sup>, Shou Li<sup>5</sup>, Bin Ai<sup>6</sup>

State Grid Qinghai Provincial Electric Power Company Haidong Power Supply Company, Haidong, China

<sup>1</sup>Corresponding author

**E-mail:** <sup>1</sup>934640100@qq.com, <sup>2</sup>124270163@qq.com, <sup>3</sup>myl123007@sina.com, <sup>4</sup>1208441733@qq.com, <sup>5</sup>94598897@qq.com, <sup>6</sup>460554340@qq.com

Received 28 August 2024; accepted 15 December 2024; published online 18 February 2025

DOI <https://doi.org/10.21595/jme.2024.24496>



Copyright © 2025 Jiaqi Peng, et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Abstract.** In order to comprehensively consider the time correlation of the signal and avoid converting the signal into the frequency domain, a higher classification success rate can be obtained. This paper proposes a decision-making model based on LBT-ODF space and multi-channel network for transformer voiceprint fault diagnosis. First, one-dimensional local features are extracted from the signal to acquire significant data of the temporal characteristics and a blended feature vector is derived as a result, then converting the signal from the time domain to a custom feature domain. Subsequently, a multi-channel one-dimensional convolutional neural network (1D-CNN) is used to process single feature and mixed feature channels respectively. At the same time, a multi-channel decision-making network is used to make majority decision-making judgments on the identification categories from each one-dimensional CNN model, and output the final fault category. The experimental outcomes demonstrate that this approach exhibits commendable recognition proficiency and resilience against noise. Consequently, the terminal model's identification precision escalates to a remarkable 96 %. This article provides a new idea for fault identification based on refined feature fusion measurement.

**Keywords:** voiceprint recognition, one-dimensional convolutional neural network, mechanical fault diagnosis, transformer.

## 1. Introduction

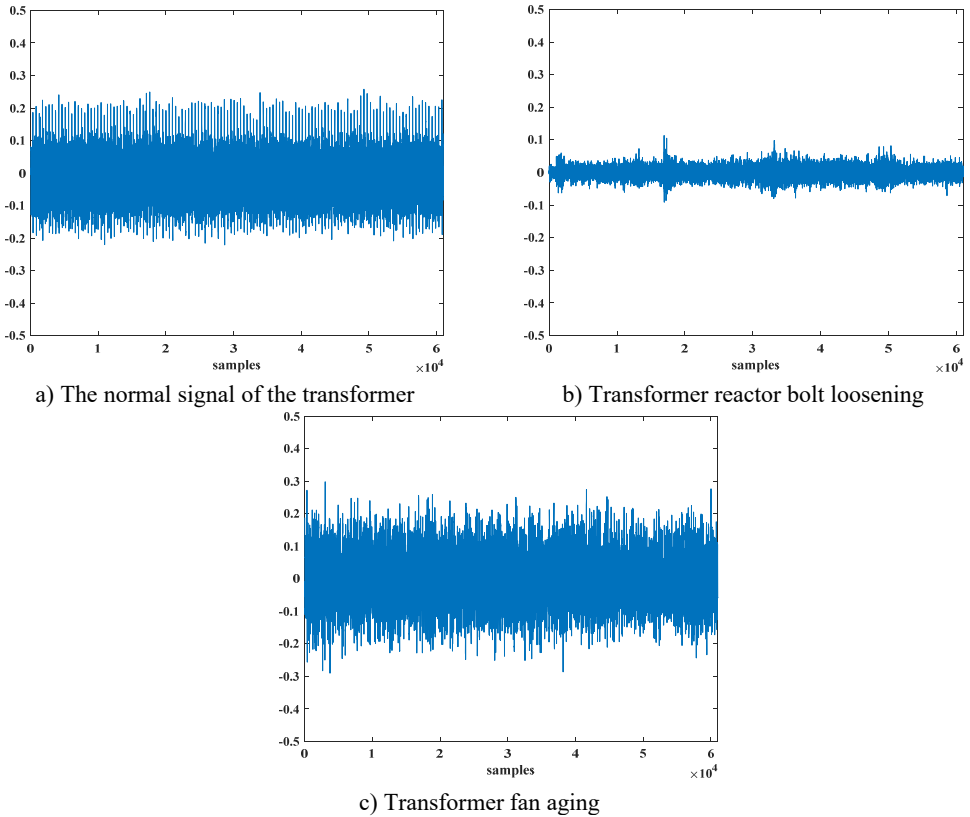
Transformers are key components in power systems, and their operating conditions directly affect the safety and stability of the entire power grid. Early detection and diagnosis of transformer faults are crucial for preventing accidents and reducing economic losses. Traditional fault detection methods rely on regular maintenance and manual inspections, which have issues such as long inspection cycles, high costs, and low accuracy. In recent years, with the development of measurement science and technology, fault detection methods based on acoustic signal recognition have become a research hotspot. These methods analyze the acoustic signals of transformers under different operating conditions, enabling rapid and accurate fault identification [1].

For example, Khan et al. proposed a method for acoustic signal detection using multi-channel acoustic sensors [2]. By placing multiple sensors at different positions on the transformer, they achieved multi-angle and multi-directional acoustic signal acquisition, significantly improving the accuracy of fault detection. This method not only captures weak acoustic signals but also enhances the robustness of detection through multi-channel data fusion.

In terms of signal preprocessing, Kim et al. studied various signal preprocessing techniques for transformer acoustic fault detection. They used adaptive filters and spectral subtraction to preprocess the raw audio signals, effectively reducing environmental noise and high-frequency interference, and improving signal quality (Kim et al. [3]). These preprocessing techniques provide a high-quality data foundation for subsequent feature extraction and analysis [4].

The normal signal of the transformer is shown in Fig. 1(a), Fig. 1(b) represents the voiceprint signal of the transformer reactor bolt loosening, and Fig. 1(c) represents the voiceprint signal of

the transformer fan aging.



**Fig. 1.** Normal signal and abnormal signal of the transformer

Analysis methodologies currently employed for transformer vibration signals predominantly feature techniques such as fast Fourier transform, variational mode decomposition [5], and collective empirical mode decomposition [6]. Additionally, wavelet packet decomposition and variational mode decomposition methods are also utilized in the analytical repertoire [7, [8], etc. For example, deep convolutional neural networks (DCNN) excel in processing complex vibration signals, effectively extracting features and performing fault diagnosis [9], [10]; long short-term memory networks (LSTM) demonstrate strong capabilities in time series data analysis, capturing long-term dependencies [11], [12]; autoencoders (Autoencoder) are used in unsupervised learning for feature extraction and dimensionality reduction, improving the efficiency and accuracy of signal processing [13], [14]. Compared with the voiceprint method, although the vibration signal identification method has no direct electrical connection with the transformer and has little interference, it requires high detection sensitivity and relies on the installation of the transformer box. The installation location requires further planning [15]. For voiceprint recognition, it is mainly based on traditional time-frequency and other feature information and artificial intelligence perception methods. Literature [16] by collecting the voiceprint or noise signals of dry-type transformers under typical defects such as loose cores and winding deformations for characteristic analysis, the identification of different fault states is achieved. However, the recognition results are susceptible to interference from transformer structure and environmental factors. And proposes a voiceprint recognition method for transformer mechanical faults based on fast incremental support vector data description and gated cycle units. However, the analysis of features is not sufficient. The characteristic values selected by the monitoring system are too

single. Neural networks can usually process time series data and mine the connections of data streams on time scales. The literature [17] combines unsupervised learning of stacked autoencoding (SAE) and supervised learning. Combined with the long short-term memory network (LSTM), a transformer diagnosis network framework based on SAE-LSTM is proposed. SAE is used for unsupervised pre-training to remove redundant information. Finally, the SAE training results are sent to the LSTM network for classification. Recognition, literature [18] uses a recurrent neural network architecture (RNN) to capture vibration time series characteristics and use it to determine the possibility of under-excitation, over-excitation and inter-turn faults of early transformers. Generally, convolutional neural networks are suitable for frequency domain information processing of signals. The 1D-CNN network generated by the variant can directly process time series data and feature selection without the need for frequency domain space to avoid excessive calculations and overly complex feature factors. problem, it is more suitable for fault identification of transformer sound signals. Applying LBP and LTP methods to one-dimensional sound signals facilitates the acquisition of crucial data regarding the local temporal features of sound, and at the same time, single features or mixed mining features need to be considered to ensure better recognition accuracy [19], [20].

This paper proposes an innovative network model (LBT-ODF-MVN) that integrates a custom feature space domain, multiple one-dimensional CNN modules, and a multi-channel decision module for the acoustic recognition of transformer mechanical faults. First, we extract local time-domain related features from the one-dimensional signal sequences of fault samples. To compensate for the loss of global information, this study introduces discrete and fitting features, which are embedded into the custom feature space LBT. This custom feature space not only considers the significance of different feature sequences but also captures their internal relationships, thereby better preserving the rich information in the data.

Furthermore, to fully utilize multi-scale information, we construct a Multi-View Network (MVN) model, which further enhances the fault recognition capability. Experimental results show that the proposed LBT-ODF-MVN model exhibits high efficiency and accuracy in fault recognition tasks. Specifically, by combining local time-domain features with multi-scale information, our model can more comprehensively capture fault characteristics, significantly improving recognition performance.

The innovations of this study lie in:

- Custom Feature Space: By embedding discrete and fitting features, we form a unique feature space LBT, which better represents and utilizes the complex information in the data.
- Multi-View Network Model: The MVN model allows us to integrate information from different scales, enhancing the robustness and accuracy of fault recognition.
- Multi-Channel Decision Module: Through the multi-channel decision module, we can effectively utilize the outputs from multiple one-dimensional CNN modules, further improving recognition performance.

These innovations not only advance the technology of acoustic recognition for transformer mechanical faults but also provide important technical support for practical engineering applications.

## 2. CNN transformer voiceprint recognition based on one-dimensional fusion features

The process of transformer fault voiceprint recognition primarily consists of two main components: feature extraction and algorithmic identification. In this paper, we construct a custom feature space using LBT-ODF features and propose a transformer voiceprint recognition method based on a multi-channel network model. The recognition process is illustrated in Fig. 2. To enhance the efficiency of the overall model, we preprocess the raw audio signals using a fifth-order Butterworth bandpass filter (25-400 Hz) to reduce noise. The filtered signals are then transformed from the time domain to individual feature vectors and mined feature vectors. We use LBT-ODF features to build a custom feature space, which includes Local Binary Patterns (LBP) and Local

Ternary Patterns (LTP) to capture local characteristics of the acoustic signals. Additionally, we employ one-dimensional Local Binary Patterns (1D-LBP) and one-dimensional Local Ternary Patterns (1D-LTP) to further refine the feature extraction. The multi-channel one-dimensional CNN structure includes four convolutional layer modules, an attention mechanism module, and a recognition module to process each feature vector. Finally, the designed multi-channel decision module performs a majority voting decision on the classification results obtained from each one-dimensional CNN model, generating the final result.

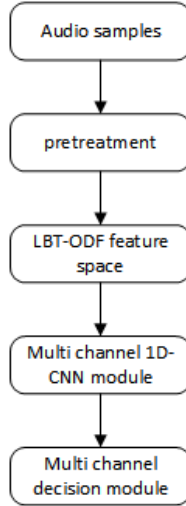


Fig. 2. Voiceprint recognition process

### 3. Preprocessing

Preprocessing is the fundamental step in our method. As mentioned in [21], this article uses a fifth-order Butterworth filter with a bandpass of 25-400 Hz to reduce noise from the transformer sound signal. For the design of the filter, the 1 dB frequency band is selected through attenuation. Fig. 3 represents the amplitude response, which shows that the response is initially flat and then gradually decreases.

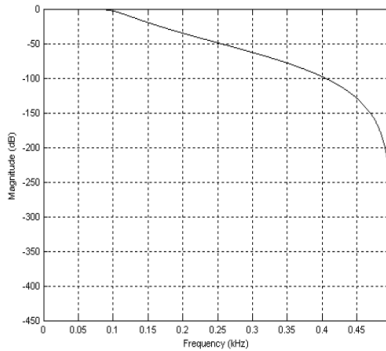


Fig. 3. Filter amplitude corresponding

### 4. LBT-ODF feature space

The feature space articulated herein is predominantly crafted from a blend of two categories of amalgamated data, encompassing local feature particulars and self-mined insights. These constitute the Local Binary Pattern features (LBP) and the Local Ternary Pattern information

(LTP), respectively. LBP was originally in article [22]. The approach entails partitioning the image into several segments to harvest local characteristics. Typically, LBP is delineated through three distinct circular vicinities. Illustrated in Fig. 4 are the trio of LBP operators, as referenced in [20].

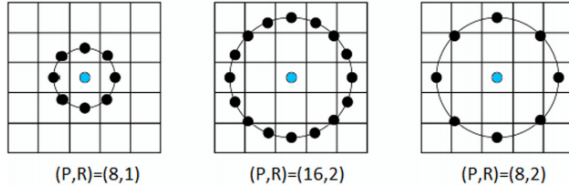


Fig. 4. Three domain definitions

Each pixel in the image is bestowed with an LBP label, constituted by a series of 1's and 0's. These binary labels are generated through a comparative analysis between each pixel and its surrounding  $N \times N$  neighborhood, centered on the focal pixel. Under the LBP scheme, a pixel is denoted as 1 should its value exceed or match that of the central pixel; conversely, it is flagged as 0 if lesser. A visual representation of pixels coded by the LBP operator can be observed in Fig. 5.

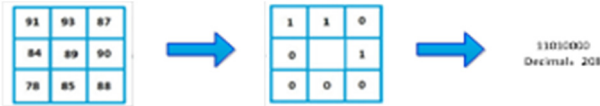


Fig. 5. LBP extraction process

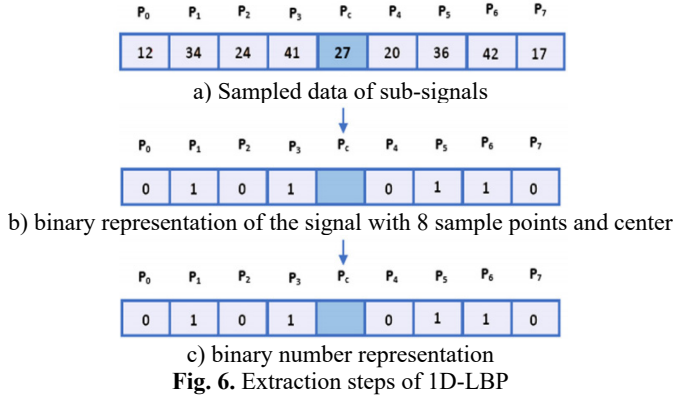
In an unprecedented move, the one-dimensional local binary pattern (1D-LBP) technique was adapted for use with speech signals, taking inspiration from the procedural framework of 2D-LBP in article [17]. The operational sequence mirrors that of its two-dimensional relative. Initially, a predefined number of samples are chosen, and the signal's pivotal point is established; within this work, a sample ensemble of 8 elements is adopted. Adhering to the principles of 2D-LBP, each successive signal value undergoes comparison against the central datum – those exceeding or equating to it receive a binary code of 1, while lesser values are denoted by 0. This comparative analysis is iteratively conducted across the entirety of the signal series. Ultimately, the central signal datum is supplanted by the aggregated, weight-adjusted sum of proximate values, leading to the derivation of a feature vector spanning  $1 \times N$  dimensions. The binary pattern relationship intrinsic to each signal value manifests thusly:

$$LBP_{P,R}(x,y) = \sum_0^{P-1} s(g_i - g_c)2^P, \quad (1)$$

where  $g_c$  is the center of the signal,  $P$  is the number of signal sample values in the neighborhood, and  $2^P$  is the coefficient of the signal sample value in each neighborhood. Fig. 6(a) delineates the magnitudes of both the neighboring and central samples. In Fig. 6(b), those domain samples surpassing the values of the  $P_c$  samples are annotated with a digit of 1, juxtaposed against the rest, which are inscribed as 0. Fig. 6(c) illustrates how the central signal value corresponds to the weighted summation of its surrounding values, referenced in relation to the central sample.

At the same time, rapid energy conversion often brings problems such as insufficient information extraction, and LBP is easily affected by noise and acoustic sample fluctuations [20]. In order to further improve the recognition ability of LBP, this article embeds the LTP feature into the feature vector and converts the signal from the time domain to the custom feature domain. LTP encapsulates the relational dynamics between a pixel and its domain, manifesting as “greater than”, “equal to”, or “less than”. When subjected to identical sampling parameters, LTP yields a richer, more intricate soundscape of features vis-à-vis 1D-LBP. Within the LTP framework,

threshold selection adopts a tripartite valuation scheme, contingent upon the interplay between the focal pixel and its neighbors. These values correspond to 0, 1, and -1, respectively, delineating the comparative outcomes.



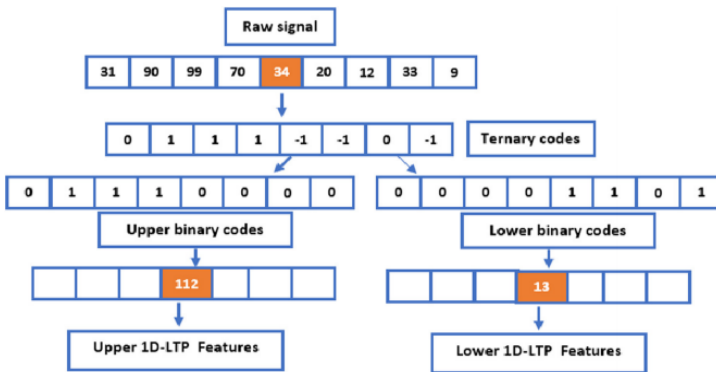
The Eq. (2) is usually chosen to calculate LTP characteristics:

$$S(q_i, p, t) = \begin{cases} -1, & q_i - (p - t) \leq 0, \\ 0, & (p + t) < q_i < (p - t), \\ 1, & q_i - (p + t) \geq 0, \end{cases} \quad (2)$$

where  $S(X)$  represents the ternary mode acoustic signal,  $q_i$  represents the domain sample value,  $p$  represents the center sample value, and  $t$  represents the threshold. The 1D-LTP features are garnered from the signal's upper and lower segments. These are meticulously computed and encoded, leveraging both lower and upper threshold values, as detailed subsequently Eq. (3a) and Eq. (3b):

$$TU = \sum_{i=0}^7 s^u(q_i, p, t)2^i, \quad (3a)$$

$$Tl = \sum_{i=0}^7 s^l(q_i, p, t)2^i. \quad (3b)$$



After obtaining three sets of feature vectors of a single sound signal data, in order to mine as much information as possible in the feature sequence, this paper defines three types of features according to the characteristics of feature calculation: original features, discrete features and

fitting features. Original features describe the original distribution of data, discrete features mainly describe the degree of dispersion and changing trends between data, and fitting features mainly show the contextual and implicit characteristics of the data.

If the overall LBT sequence is rough, there may be an increasing or decreasing trend locally. This increasing or decreasing trend feature can be extracted through the first-order difference operation: the second-order difference of the sequence can further represent continuous adjacent The change of concavity and convexity between value values; and the occupancy ratio is the proportion of the difference between the current value and the minimum value in the current sequence to the difference between the overall maximum and minimum values, which can make the abnormal data characteristics at the extreme points more obvious because of them They are all obtained by processing data using discrete statistical methods, so this article classifies first-order differences, occupancy ratios and concavity as discrete features. The fitting feature takes the “sliding window” as an opportunity to consider the integrity of the vector and the correlation between time series data, making up for the shortcomings of considering a single sample value in the original features and discrete features [19], so This paper selects kurtosis coefficient, coefficient of variation, and moving average as fitting features [18], [19].

Suppose the original feature vector is  $(q_1, q_2, \dots, q_m)$ , where  $m$  denotes the extent of the dataset and  $q$  is the value of the  $i$ -th time point.

(1) Original features.

The original data is defined as the original feature, and the original feature value of the sequence  $F = (f_1, f_2, \dots, f_m)$  is the same as  $(q_1, q_2, \dots, q_m)$ .

(2) Discrete characteristics.

1) The first-order difference is the difference between consecutive adjacent value values, which can represent the change of local monotonicity between adjacent data. The first-order difference  $D = (d_1, d_2, \dots, d_m)$  of the sequence is Eq. (4):

$$d_i = p_{i+1} - p_i, \quad i = 1, \dots, m - 1. \quad (4)$$

2) The occupancy ratio quantifies the ratio of the gap between a feature’s current measurement and the least value within the sequence, juxtaposed with the amplitude spanning from the highest to the lowest value across the board. This ratio is advantageous in highlighting the unique attributes of data points that fall into outlier categories at the boundaries.

3) Concave-convexity represents the change of concavity and convexity between consecutive adjacent values, and is defined by the second-order difference. Concave-convexity is a supplement to the first-order difference:

$$S_i = p_{i+2} - 2p_{i+1} + p_i, \quad i = 1, \dots, m - 2. \quad (5)$$

(3) Fitting features.

The fitting feature is based on the sliding window. Let the data in the sliding window be  $win$  and the size of the window be  $w$ . The vectors are processed using the “delay embedding method”:

$$win = [p_i, p_{i+1}, \dots, p_{i+w-1}], \quad (6a)$$

$$p_w = [win_1, win_2, \dots, win_{m-w+1}], \quad (6b)$$

1) Kurtosis coefficient:

$$K_i(w) = E \left[ \left( \frac{win_i - \mu}{\sigma} \right)^4 \right], \quad i = 1, 2, \dots, m - w + 1. \quad (7)$$

It describes whether the time series is heavy-tailed or light-tailed relative to the normal distribution, where  $E$  represents the mathematical expectation and  $w$  represents the size of the

sliding window. Time series with high kurtosis tend to have heavier tails or outliers, making them better at detecting sudden spikes or dips.

2) Coefficient of variation:

$$C_i(w) = \frac{\sigma}{\mu}, \quad i = 1, 2, \dots, m - w + 1. \quad (8)$$

It describes the dispersion of a sequence in a sliding window, indicates the extent to which the data is concentrated on the average value, and reflects the overall change of a sequence.

3) The average moving line describes the average value of the corresponding sequence in the sliding window. It weakens irregular changes in time series and reveals changing trends among data:

$$z_{ik}^l = w_k^{lT} x_i^l + b_k^l. \quad (9)$$

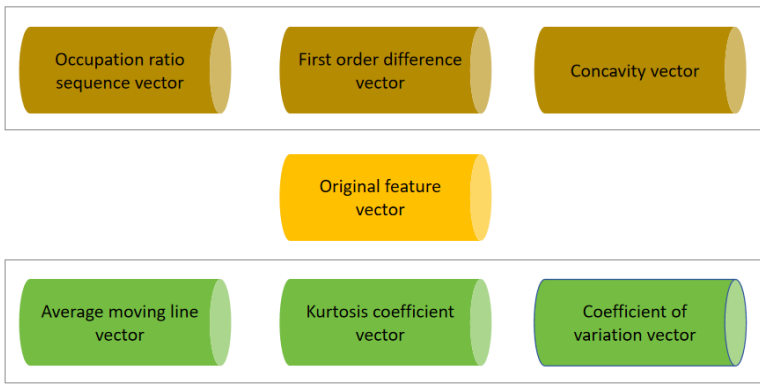


Fig. 8. Custom feature space

## 5. MVN model

After completing the construction of the multi-dimensional feature domain, taking into account the meaning and intrinsic relationship of different feature sequences, in order to fully integrate multi-scale information, this paper constructs an MVN model to complete the fault identification task. The model structure is shown in Fig. 9.

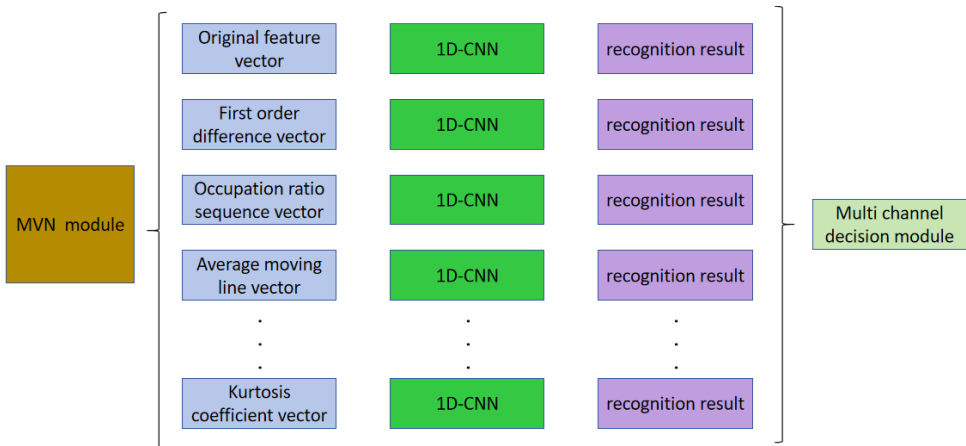


Fig. 9. MVN model



Featuring multi-dimensional feature vectors as its input, the one-dimensional CNN construct is anchored by four convolution sections, an Attention Mechanism (AM) module, alongside a categorization unit delivering early results. Thereafter, a dedicated decision-making segment furnishes the terminal recognition outcomes.

### 5.1. Convolution module

The one-dimensional convolutional layer [23] uses a one-dimensional filter with a preset width to locally convolute the input data in the order of steps, and then outputs the corresponding result. The calculation method of convolution is Eq. (10):

$$y_i^l = \sum_{j=0}^{k-1} x_{i+j-bk/2c}^{l-1} \cdot w_j^{l,k} + b^{l,k}, \quad (10)$$

where  $x_i^l$  is the input signal centered on position  $i$  of the  $l$ -th layer,  $w_k^{l,T}$  and  $b_k^l$  are the weight matrix and bias term of the  $k$ -th filter of the  $l$ -th layer respectively, and  $z_{ik}^l$  is the  $k$ -th feature map of the  $l$ -th layer. The kernel sizes for the quartet of convolution modules are standardized at 32, 32, 64, and 64 correspondingly. Given the shared parameter framework, each layer's convolution kernel count synchronizes with its respective convolutional layer's feature map quantity. Thus, reduced kernel numbers in superficial layers capture fundamental, lower frequency attributes, and increased kernel counts in more profound layers harvest a plethora of higher frequency features. This approach minimizes parameters, safeguarding against excessive model training on noise. Post-convolution, Batch Normalization (BN) is a pivotal inclusion, expediting training, stabilizing gradients, and regularizing, thereby amplifying convergence rates and network sturdiness. A modified linear unit, acting as the activation function post-BN, bolsters non-linearity. The pooling layer refines extracted information into a condensed, vectorial representation, preserving salience whilst discarding trivialities. Architecturally, it diminishes succeeding layer inputs via subsampling, alleviating computation and endowing shift tolerance. Max pooling, by electing maximal values within regions, following the secondary convolution segment, safeguards textural traits, with its operation defined as Eq. (11):

$$p_{ik} = \max_{(o,p) \in R_{ik}} (y_{koq}), \quad (11)$$

where  $y_{koq}$  is the element at  $(o, p)$  in the pooling region  $R_{ik}$ , representing the local neighborhood near the  $k$ -th feature map position  $i$ ,  $p_{ik}$  is the relevant output after the pooling operation. This article sets the size and stride of the pooling kernel to 2.

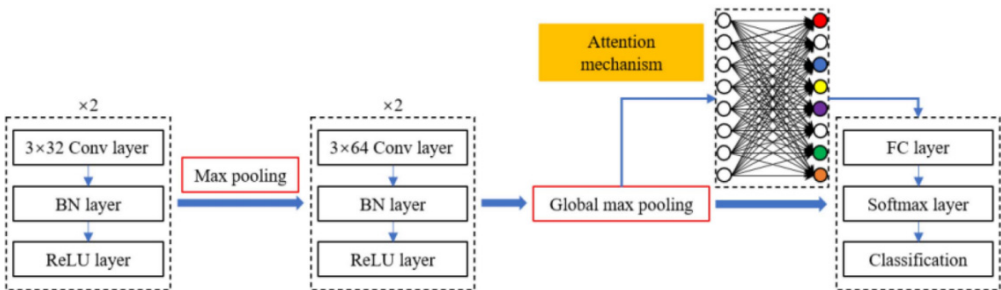


Fig. 10. 1D-CNN

Global max pooling, through its selection of the supreme element across all feature maps, prunes the fully connected layer's parameter set in both the Attention Mechanism and classification modules, proficiently averting overfitting whilst concurrently conserving

computational power and time. Consequently, a global max pooling level is instituted post-quartet convolution and pre-AM module:

$$g_k = \max(y_k). \quad (12)$$

Among them,  $y_k$  is the activation value of the  $k$ -th feature map, and  $g_k$  is the relevant output of the pooling operation.

## 5.2. Attention mechanism (AM)

Within the convolutional module, an array of features is distilled from a multi-faceted feature cosmos. However, in a complex noise ecosystem, the signal waveform profiles gathered across different epochs and geographies manifest heterogeneities. While some features serve as valuable indicators for identifying and classifying transformer mechanical malfunctions, others may act as confounding factors, impairing the model's generalizability and sturdiness. Therefore, this treatise introduces an Attention Modulation (AM) component, crafted to dynamically adjust feature mapping weights. This empowers the model to filter out superfluous features whilst concentrating on those bearing significant data. The AM component encompasses a fully connected layer and a sigmoid layer. The fully connected layer amalgamates every neuron from the preceding layer, intertwining it with each neuron in the current layer to engender a holistic semantic understanding. Thus, the neuron count in the fully connected layer of the AM module aligns with the number of output feature maps. Suppose the sequence of activation values following the application of the global max-pooling operation for each feature map is  $G = [g_1, g_2, \dots, g_k]$ , and  $K$  is the number of feature maps. Initially, feed the feature sequence into the FC layer, then acquire the attention mechanism weights, denoted as  $S$ , for each feature map via a sigmoid activation layer [23]. This procedure unfolds as Eq. (13):

$$S = \varepsilon(WG + b) = (s_1, s_2, \dots, s_k). \quad (13)$$

Among them,  $W$  and  $b$  are the weight matrix and bias vector of the FC layer respectively,  $\varepsilon()$  is as the sigmoid operator:

$$\varepsilon(x) = \frac{1}{1 + e^{-x}}. \quad (14)$$

Among them,  $x$  is the output of the fully connected layer. Post nonlinear activation, the attention mechanism weights assume values within the range of (0-1). Subsequently, these weights  $S$  are element-wise multiplied with the original feature sequence, yielding a feature sequence  $G$  that has been filtered through the AM process.

## 5.3. Identification module

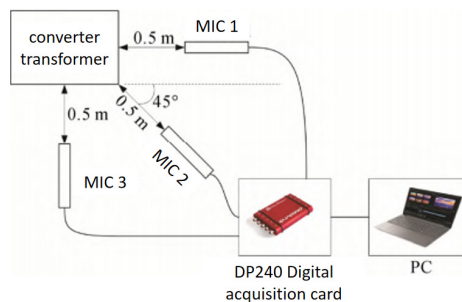
The convolution module extracts depth information from the multi-dimensional feature space, and uses the AM module to adaptively weight the obtained information to obtain useful classification information. The AM-weighted information sequence is input into the recognition module for recognition. This component encompasses fully connected layers alongside softmax layers. Within the identification module, the neuron count in the fully connected layers mirrors the category number of the feature sequences, fixed at a figure of two. The concluding layer of the CNN constitutes the output layer. Specifically, for categorization endeavors utilizing CNN, the softmax layer customarily serves as the terminal stratum.

## 5.4. Multi-channel decision-making module

To address the issue of diminished recognition accuracy in the holistic model, usually the mel spectrum feature information and local feature information of loose core and loose coil are very similar, which reduces the overall accuracy of the single-channel system. The input feature sequences in different channels have different effects on the environment. The response is different. Therefore, when used alone to analyze fault signal waveforms in deep or shallow channels, the overall accuracy is lower. Hence, a multi-channel decision module is introduced, utilizing the multi-dimensional feature sequence derived from the bespoke feature space architecture detailed herein. This module enacts a voting procedure amongst the forecasted classes output by each channel-specific 1-D CNN component, adhering to a majority voting principle. The category garnering over fifty percent of the vote is then designated as the conclusive recognition outcome. Should parity occur in the tally for core and coil defect classifications, precedence is given to the prediction made by the first-order difference feature channel, which is henceforth adopted as the ultimate categorization. This mechanism ensures that the determination of each fault type is a collaborative effort, incorporating insights from diverse channels, thereby circumventing potential extremities and bolstering both the precision and resilience of fault discernment.

## 6. Experiments and results

The test object of this study is a custom-made real-type oil-immersed converter transformer, model D-800/35. The design parameters and structural characteristics of this transformer allow it to simulate various operating conditions found in actual operation. The experimental setup includes the following main components: the transformer body, acoustic sensors, and sensor placement. The transformer body is a D-800/35 type oil-immersed converter transformer, capable of simulating different mechanical fault scenarios. The acoustic sensors consist of high-performance microphones and high-precision acquisition cards. The frequency response range of the microphones is 3.5-20000 Hz, allowing them to capture a wide range of acoustic signals. The acquisition card model is DP240, with a maximum sampling frequency of 102.4 kHz, ensuring high-resolution and high-precision data. The experimental setup is as shown. To comprehensively capture the acoustic signals at different positions on the transformer, we placed three microphones on the front, side, and diagonal corners of the transformer. Each microphone is positioned at the same height as the center of the transformer and is 0.5 meters away from the transformer's surface. This multi-point placement provides more comprehensive and accurate acoustic data.



**Fig. 11.** Experimental layout

To ensure the accuracy and reliability of the experimental data, we adopted the following data collection methods: First, the sampling frequency was set to the maximum sampling rate of 102.4 kHz. A high sampling rate captures more detailed information, enhancing the resolution and accuracy of the data. Second, data collection from all sensors is synchronized to ensure temporal

consistency of the acoustic signals from different positions, which aids in subsequent data analysis and fault diagnosis. During the experiment, we strictly controlled the experimental environment to minimize external noise interference. The laboratory was kept quiet, and other machinery was avoided to ensure the purity of the collected acoustic signals. Each experiment lasted 10 minutes, with acoustic signals recorded every minute. During data recording, we used professional data acquisition software to monitor data quality and collection status in real-time, ensuring the integrity and reliability of the data. Finally, the raw data collected were preprocessed, including noise reduction and filtering steps, to reduce the impact of environmental noise and improve signal quality. The preprocessed data were used for subsequent feature extraction and analysis.

A total of 9180s of valid audio were collected. These audios were then classified, labeled, and classified into data sets, and the audios were divided into 6 situations: excitation, current, loose core excitation, loose core current, loose coil excitation, and loose coil current. Within this experimental setup, the proportion of the training dataset to the testing dataset approximates a ratio of 8:2.

To more accurately quantify the uncertainty of the measurements, we used the Standard Error (SE) in the experimental results to assess the stability of the measurement outcomes. The standard error measures the standard deviation of the sample mean, reflecting the variability of the sample mean. A smaller standard error indicates a more precise estimate of the sample mean and more stable measurement results. The formula for calculating the standard error is shown Eq. (15):

$$SE = \frac{s}{\sqrt{n}} \quad (15)$$

where  $s$  is the sample standard deviation,  $n$  is the sample size.

The confidence interval (CI) provides a range within which the true population mean is likely to fall. A 95 % confidence interval means that we are 95 % confident that the true population mean lies within this interval. The narrower the confidence interval, the less uncertainty there is about the measurement results:

$$CI = \bar{x} \pm t_{\alpha/2, n-1} \left( \frac{s}{\sqrt{n}} \right), \quad (16)$$

where  $\bar{x}$  is the sample mean,  $t_{\alpha/2, n-1}$  is the critical value of the  $t$ -distribution,  $s$  is the sample standard deviation, and  $n$  is the sample size.

The coefficient of variation (CV) is used to measure the relative variability of the data. It is calculated as the standard deviation divided by the mean, expressed as a percentage. A smaller coefficient of variation indicates lower relative variability and more consistent measurement results:

$$CV = \left( \frac{s}{\bar{x}} \right) \times 100 \%. \quad (17)$$

The standard errors for the measurement results of excitation fault, current fault, core loosening excitation fault, core loosening current fault, coil loosening excitation fault, and coil loosening current fault are 0.028, 0.021, 0.017, 0.012, 0.016, and 0.024, respectively.

The 95 % confidence intervals for the measurement results of excitation fault, current fault, loose core excitation fault, loose core current fault, loose coil excitation fault, and loose coil current fault are 0.94, 0.98, 0.97, 0.93, and 0.92, respectively.

The coefficients of variation (CV) for the measurement results of excitation fault, current fault, loose core excitation fault, loose core current fault, loose coil excitation fault, and loose coil current fault are 0.051, 0.047, 0.023, 0.042, and 0.036, respectively.

It can be seen that the measurement errors are extremely small, fully meeting the experimental

requirements.

This paper first evaluates the rationality of the self-defined feature space, and conducts comparative experiments by constructing a multi-channel network and a single-channel 1D-CNN network. In addition, this article selects two traditional machine learning methods: traditional SVM and KNN, as well as the neural network deep learning methods of Piczak-CNN [24] and MixupCNN [23] for comparative experiments.

This paper compares the proposed LBT-ODF-MVN model with traditional machine learning methods, including Support Vector Machine (SVM) and K-Nearest Neighbors (KNN), as well as deep learning methods such as Piczak-CNN [20] and MixupCNN [21].

The LBT-ODF-MVN model proposed in this paper demonstrates significant superiority in transformer voiceprint recognition tasks, achieving an accuracy of 96.07 %. In contrast, traditional Support Vector Machines (SVM) perform well on small-scale datasets but experience a decline in performance when handling large-scale and high-dimensional data, with an accuracy of only 72.7 %. K-Nearest Neighbors (KNN) algorithms, although simple and easy to use, are sensitive to the scale and dimensionality of the data, leading to higher computational complexity, with an accuracy of 78.3 %. Piczak-CNN performs well in environmental sound classification tasks but has room for improvement in specific domain fault recognition tasks, achieving an accuracy of 83.7 %. MixupCNN enhances the model's generalization capability through data augmentation techniques but still has room for improvement in specific domain fault recognition tasks, achieving an accuracy of 92.2 %. One-dimensional Convolutional Neural Networks (1D-CNN) perform well in handling time-series data but require further optimization in feature selection and processing, achieving an accuracy of 93.2% with raw feature sequences and 93.7 % with first-order difference sequences. The LBT-ODF-MVN model, through its custom feature space and multi-channel decision mechanism, is better able to capture and utilize key features in time-series data, thereby significantly improving recognition accuracy and robustness. These characteristics make the LBT-ODF-MVN model excel in transformer voiceprint recognition tasks, clearly outperforming other methods.

**Table 1.** Fault identification comparison results

Model	Accuracy
SVM	72.7 %
KNN ( $k = 5$ )	78.3 %
Piczak-CNN	83.7 %
MixupCNN	92.2 %
1D-CNN (original feature sequence)	93.2%
1D-CNN (first-order difference sequence)	93.7 %
LBT-ODF-MVN	96.07 %

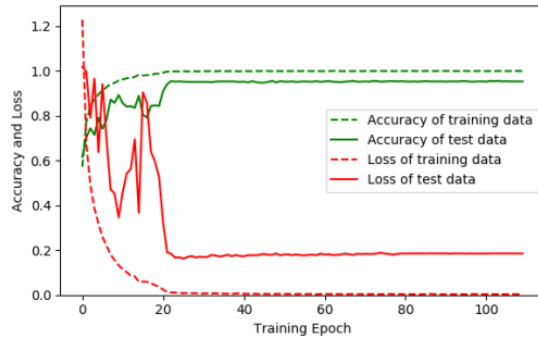
The learning rate hits its nadir at the 20th and 80th epochs, spanning an overall training period of 110 epochs. Employing the stochastic gradient descent optimizer with a momentum of 0.9, and utilizing a 10-fold cross-validation technique, the workflow proceeds as follows: Initially, the dataset of 8732 samples is shuffled and subsequently partitioned into nine folds of 875 samples each, and one fold of 857 samples. The mean outcome of the 10-fold cross-validation is a commendable 95.7 %, while the zenith of accuracy for the model is recorded at 96.07 %. A comparative overview delineated in Table 1 underscores the superiority of our proposed model over those put forth by the majority of alternative research endeavors, alongside conventional machine learning methodoologies.

Focusing on a selection of prototypical experimental outcomes, as illustrated in Table 2, reveals an outstanding capacity to distinguish loose coil faults, achieving a success rate surpassing 95 %. Performance on other fault types appears less impressive. Analysis of the data suggests that audio signals characterized by higher autocorrelation coefficients and Root Mean Square (RMS) values tend to yield superior recognition rates.

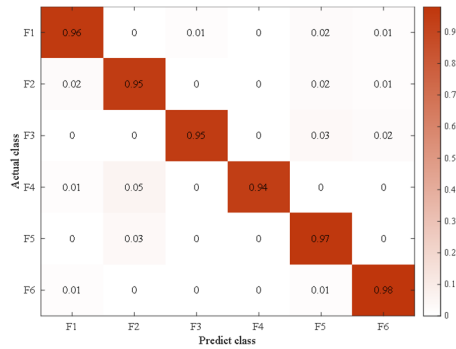
**Table 2.** Voiceprint diagnosis results of multiple faults in transformers

	Excitation	Current	Loose core excitation	Core loosening current	Loose coil excitation	Coil loose current
Zero crossing rate	0.125	0.259	0.182	0.158	0.326	0.101
Autocorrelation RMS	0.014	0.007	0.006	0.009	0.006	0.015
Accuracy	1.595	3.03	0.914	4.437	2.299	5.660
Accuracy	0.972	0.946	0.946	0.928	0.957	0.977

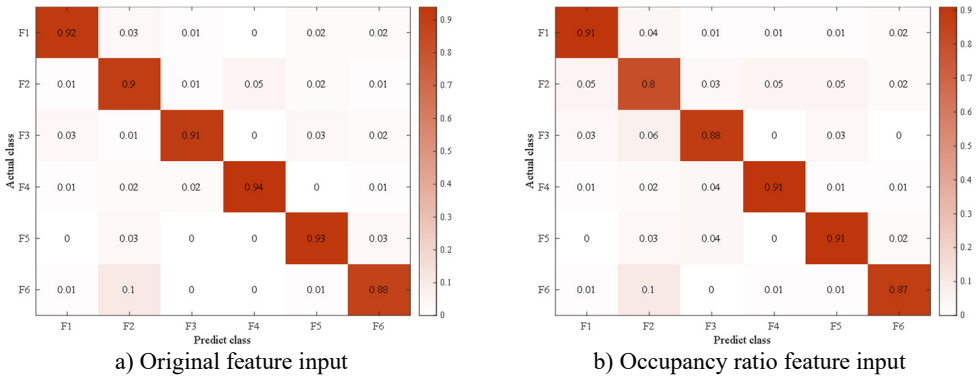
The Fig. 12 describes the training loss curve and category mixing matrix of the LBT-ODF-MVN model. Observe that in the initial 20 epochs, the employment of a larger learning rate instigates persistent oscillations in the loss, which simultaneously undergoes a swift decline. Post the 20th epoch, the model institutes a learning rate reduction, thereby stabilizing and embarking on a quest for the optimal solution.

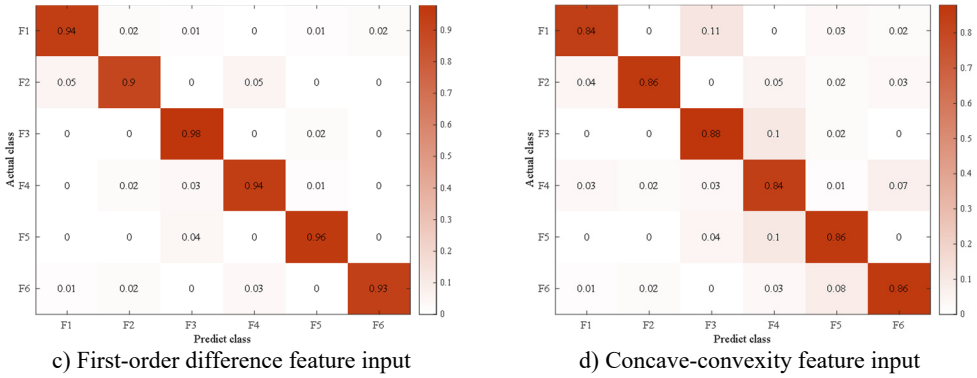


**Fig. 12.** Training curves of accuracy and loss of the proposed LBT-MVN on the transformer dataset



**Fig. 13.** Mixing matrix





**Fig. 14.** Fuzzy matrix for recognition of different feature inputs in a single channel

In order to verify the LBT-ODF-MVN model in this article when the votes are consistent in the multi-channel decision-making module, this article sets the first-order difference sequence as the final recognition result. On the basis of a single channel and excluding the decision-making module, the following conclusions are drawn by comparing the recognition accuracy of each feature sequence. Since the fitting features play a complementary role in the overall local information and have less impact on the subsequent final judgment, this article selects the original features. and discrete feature sequences as single-channel model inputs. By comparing the first-order difference features, the output results are more accurate and meet the setting requirements.

## 7. Conclusions

This paper proposes a network model (LBT-ODF-MVN) that contains a custom feature space domain, multiple one-dimensional CNN modules and a multi-channel decision module for voiceprint recognition of transformer mechanical faults. First, local time-domain related features are extracted from the one-dimensional signal sequence of the fault sample. In order to make up for the global lost hidden information, this paper embeds discrete features and fitting features to form a custom feature space LBT, taking into account the meaning and meaning of different feature sequences. Internal relationship, in order to fully integrate multi-scale information, this paper constructs an MVN model to complete the fault identification task. The results show the efficiency and accuracy of this article's model and custom feature domain. In future work, other newly proposed and improved 1D-CNN models will be considered to integrate into the overall framework to obtain better recognition methods.

## Acknowledgements

The authors have not disclosed any funding.

## Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

## Author contributions

Peng Jiaqi: mathematical model and the simulation techniques. Ma Yulin: spelling and grammar checking as well as virtual validation; Ye Haiping: software. Che Xianrui: writing-original draft preparation. Li Shou: resources. Ai Bin: experimental validation.

## Conflict of interest

The authors declare that they have no conflict of interest.

## References

- [1] H. Ma et al., "Vibration characteristic analysis and natural frequency identification of transformer," *High Voltage Apparatus*, Vol. 59, No. 6, pp. 82–92, Jun. 2023, <https://doi.org/10.13296/j.1001-1609.hva.2023.06.010>
- [2] J. Peng et al., "Analysis of a phase B failure to operate in a 110 kV outdoor SF6 circuit breaker," *High Voltage Apparatus*, Vol. 59, No. 3, pp. 216–220, 2023, <https://doi.org/10.13296/j.1001-1609.hva.2023.03.029>, doi:10.13296/j.1001-1609.hva.2023.03.029
- [3] S. Wu, S. Ji, J. Sun, N. Liang, T. Zhao, and S. Dai, "Vibration monitoring and variation law of converter transformer in operation," *High Voltage Engineering*, Vol. 48, No. 4, pp. 1561–1570, Apr. 2022, <https://doi.org/10.13336/j.1003-6520.hve.20201674>
- [4] Z. Pan, J. Deng, B. Zhou, and Z. Xie, "Comparison and analysis of characteristics of vibration signal of converter transformer and AC transformer," *High Voltage Apparatus*, Vol. 58, No. 1, pp. 122–129, Jan. 2022, <https://doi.org/10.19487/cnki.1001-8425.2020.03.009>
- [5] X. Wei et al., "Reduce the noise of transient electromagnetic signal based on the method of SMA-VMD-WTD," *IEEE Sensors Journal*, Vol. 22, No. 15, pp. 14959–14969, Aug. 2022, <https://doi.org/10.1109/jsen.2022.3184697>
- [6] S. Zhao et al., "Detection of interturn short-circuit faults in DFIGs based on external leakage flux sensing and the VMD-RCMDE analytical method," *IEEE Transactions on Instrumentation and Measurement*, Vol. 71, pp. 1–12, Jan. 2022, <https://doi.org/10.1109/tim.2022.3186061>
- [7] R. Li et al., "Identification method of AC variable frequency motor rotor broken bar fault based on parameter optimization variational mode decomposition," *Journal of Electrical Engineering and Technology*, Vol. 36, No. 18, pp. 3922–3933, 2021, <https://doi.org/10.19595/j.cnki.1000-6753.tces.200572>
- [8] H. Ma et al., "Online fault diagnosis method for loose windings of transformer based on multi-feature voiceprint pattern," *Journal of Electrical Machines and Control*, Vol. 27, No. 5, pp. 76–87, 2023, <https://doi.org/10.15938/j.emc.2023.05.009>
- [9] X. Jiang et al., "Digitalization transformation of power transmission and transformation under the background of new power system," *High Voltage Engineering*, Vol. 48, No. 1, pp. 1–10, 2022, <https://doi.org/10.13336/j.1003-6520.hve.20211649>
- [10] L. Xie et al., "Prediction model of dissolved gas in transformer oil based on variational mode decomposition and gated recurrent unit neural network," *High Voltage Technology*, Vol. 48, No. 2, pp. 653–660, 2022, <https://doi.org/10.13336/j.1003-6520.hve.20201808>
- [11] Y. Shao, X. Wang, P. Peng, G. Yuan, and N. Ke, "Research on defect detection method of power equipment based on acoustic imaging technology," *China Measurement and Test*, Vol. 47, No. 7, pp. 42–48, Jul. 2021.
- [12] Xiong Q., J. Zhao, Z. Guo, X. Feng, H. Liu, and L. Liu, "Mechanical defects diagnosis for gas insulated switchgear using acoustic imaging approach," *Applied Acoustics*, Vol. 174, p. 10778, 2021.
- [13] L. Zhao, S. Wang, Y. Yang, Y. Jin, W. Zheng, and X. Wang, "Detection and rapid positioning of abnormal noise of GIS based on acoustic imaging technology," in *IET Conference Proceedings*, Vol. 2021, No. 5, pp. 653–657, Oct. 2021, <https://doi.org/10.1049/icp.2021.2368>
- [14] W. Si et al., "Research on loose detection of transformer core based on acoustic imaging and image processing," *High Voltage Apparatus*, Vol. 57, No. 11, pp. 180–186, Nov. 2021, <https://doi.org/10.13296/1001-1609.hva.2021.11.023>
- [15] Y. Shao, X. Wang, P. Peng, T. Gu, and J. Li, "Research and application on typical abnormal noise of 10 kV dry-type transformer," *Transformer*, Vol. 58, No. 5, pp. 82–87, May 2021, <https://doi.org/10.19487/j.cnki.1001-8425.2021.04.006>
- [16] L. Qian, Y. Jingjing, X. Zitao, X. Kezheng, and S. Huazhong, "Modeling and structural optimization of acoustic imaging sensor unit for detecting abnormal noises of dry-type transformer," in *International Conference on Power System Technology (POWERCON)*, pp. 2388–2392, Dec. 2021, <https://doi.org/10.1109/powercon53785.2021.9697795>
- [17] Z. Li, "Research on acoustic imaging method of dry-type transformer abnormal sound defects based on frequency domain features," Chongqing University, 2022.



- [18] Z. Zhang, Y. Wang, Z. Li, and J. Liu, "Localization of mechanical and electrical defects in dry-type transformers using an optimized acoustic imaging approach," *Plos one*, Vol. 18, No. 11, p. e0294674, 2023.
- [19] X. Du, J. Xie, J. Wu, G. Ding, S. Yan, and X. Wu, "Analysis and Investigation of a 500 kV transformer with oil chromatogram abnormality caused by on-site installation defect," *Transformer*, Vol. 58, No. 10, pp. 70–72, Oct. 2021, <https://doi.org/10.19487/j.cnki.1001-8425.2021.10.012>
- [20] H. Zhang et al., "LSTM network transformer fault diagnosis based on SMOTE and Bayes optimization," *China Electric Power*, Vol. 56, No. 10, pp. 164–170, 2023, <https://doi.org/10.19651/j.cnki.emt.2107176>
- [21] Y. Lu et al., "Transformer defect diagnosis method based on voiceprint features and integrated learning," *Power Engineering Technology*, Vol. 42, No. 5, pp. 46–55, 2023.
- [22] Sz Wenrong et al., "Research on transformer core loosening detection based on acoustic imaging and image processing," *High Voltage Electrical Apparatus*, Vol. 57, No. 11, pp. 180–186, 2021.
- [23] Q. Qin et al., "Research and system realisation of acoustic imaging and sound source positioning technology for power equipment," *Popular Electricity*, Vol. 38, No. 6, pp. 40–41, 2023.
- [24] L. C. Hong, Z. H. Chen, Y. F. Wang, M. Shahidehpour, and M. H. Wu, "A novel SVM-based decision framework considering feature distribution for power transformer fault diagnosis," *Energy Reports Amsterdam*, Vol. 8, pp. 9392–9401, Nov. 2022, <https://doi.org/10.1016/j.jegy.2022.07.062>



**Peng Jiaqi** obtained a master's degree from Lanzhou Jiaotong University in 2015. I am currently working at Haidong Power Supply Company of State Grid Qinghai Electric Power Company. The current research direction is high-voltage electrical testing and power supply quality research.



**Ma Yulin** obtained a bachelor's degree from Lanzhou University of Technology in 1996. He is currently working at Haidong Power Supply Company of State Grid Qinghai Provincial Electric Power Company. The current research direction is the study of live working in power transmission.



**Ye Haiping** obtained his bachelor's degree from North China Electric Power University in 2009. He is currently working at Haidong Power Supply Company of State Grid Qinghai Provincial Electric Power Company. The current research direction is high-voltage electrical testing research.



**Che Xianrui** obtained a bachelor's degree from Changsha University of Science and Technology in 2009. She is currently working at Haidong Power Supply Company of State Grid Qinghai Provincial Electric Power Company. The current research direction is secondary substation technology.



**Li Shou** obtained his bachelor's degree from Xi'an University of Technology in 2006. He is currently working at Haidong Power Supply Company of State Grid Qinghai Provincial Electric Power Company. The current research direction is relay protection technology.



**Ai Bin** obtained his master's degree from Northeast Electric Power University in 2014. He is currently working at Haidong Power Supply Company of State Grid Qinghai Provincial Electric Power Company. The current research direction is the development of digital distribution networks.