processes

MDPI

Article

Distribution Network Electrical Equipment Defect Identification Based on Multi-Modal Image Voiceprint Data Fusion and Channel Interleaving

An Chen [1,*], Junle Liu [1], Wenhao Zhang [1], Jiaxuan Lu [2], Jiamu Yang [2] and Bin Liao [2]

[1] China Southern Power Grid Guangdong Zhongshan Power Supply Bureau, Zhongshan 528400, China; junle_liu@163.com (J.L.); wenhao_zhangjddw@163.com (W.Z.)
[2] School of Electrical and Electronic Engineering, North China Electric Power University, Beijing 102206, China; jiaxuan_lu@ncepu.edu.cn (J.L.); jiamu_yang@ncepu.edu.cn (J.Y.); nathan@ncepu.edu.cn (B.L.)
* Correspondence: ccanchen@163.com

Abstract

With the explosive growth in the quantity of electrical equipment in distribution networks, traditional manual inspection struggles to achieve comprehensive coverage due to limited manpower and low efficiency. This has led to frequent equipment failures including partial discharge, insulation aging, and poor contact. These issues seriously compromise the safe and stable operation of distribution networks. Real-time monitoring and defect identification of their operation status are critical to ensuring the safety and stability of power systems. Currently, commonly used methods for defect identification in distribution network electrical equipment mainly rely on single-image or voiceprint data features. These methods lack consideration of the complementarity and interleaved nature between image and voiceprint features, resulting in reduced identification accuracy and reliability. To address the limitations of existing methods, this paper proposes distribution network electrical equipment defect identification based on multi-modal image voiceprint data fusion and channel interleaving. First, image and voiceprint feature models are constructed using two-dimensional principal component analysis (2DPCA) and the Mel scale, respectively. Multi-modal feature fusion is achieved using an improved transformer model that integrates intra-domain self-attention units and an inter-domain cross-attention mechanism. Second, an image and voiceprint multi-channel interleaving model is applied. It combines channel adaptability and confidence to dynamically adjust weights and generates defect identification results using a weighting approach based on output probability information content. Finally, simulation results show that, under the dataset size of 3300 samples, the proposed algorithm achieves a 8.96–33.27% improvement in defect recognition accuracy compared with baseline algorithms, and maintains an accuracy of over 86.5% even under 20% random noise interference by using improved transformer and multi-channel interleaving mechanism, verifying its advantages in accuracy and noise robustness.

Keywords: image and voiceprint data; multi-modal data fusion; channel interleaving; defect identification

Check for updates

Academic Editor: Jie Zhang

Received: 20 October 2025
Revised: 11 December 2025
Accepted: 15 December 2025
Published: 16 January 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

With the rapid advancement of smart grid, a large number of electrical equipment such as distribution transformer and high-voltage circuit breaker are connected to the distribution network [1]. However, the proliferation of massive electrical equipment has

also led to frequent grid operation failures, causing power supply interruptions and equipment damage [2,3]. In order to ensure the reliable and stable operation of power system, it is particularly important to prevent electrical equipment failures in advance. Electrical equipment defect identification can discover the defects and deal with them in a timely manner by monitoring the working state of electrical equipment based on multi-modal image voiceprint features. For example, image data features capture the surface cracks, corrosion and other information of electrical equipment, while voiceprint data features reflect the anomaly noise generated during equipment operation [4,5]. However, the distribution network operates in a complex environment characterized by high electromagnetic interference and lighting variations. These factors make image and voiceprint data features susceptible to noise, occlusion, and other disruptions, thereby increasing the difficulty of defect identification [6,7]. In addition, the applicability of specific features varies across different operation environments and fault types. For example, infrared thermal imaging is more suited for equipment in high-temperature environments, while voiceprint data is more appropriate for high-noise environments. How to fully utilize the complementary nature of multi-modal image and voiceprint features to improve the identification accuracy is an urgent problem to be solved [8].

The accuracy and robustness of equipment defect identification can be improved by extracting the core information in the image and voice print features of electrical equipment. Extensive research has been conducted in this area in academia and industry. In [9], Mei et al. proposed an infrared thermal imaging method based on DenseNet network, which significantly improves defect identification. In [10], a generalized feature embedding method was proposed by Luo et al. This method combines contrastive language-image pre-training (CLIP) and self-distillation with no labels version 2 (DINOv2) models for the fusion of infrared and visible images. Through feature embedding, the method significantly improves the detail retention and adaptability of the fused images while providing enhanced input quality for high-level tasks. In [11], Lin et al. proposed the use of an improved deep learning algorithm to process infrared images with arbitrary angles, which improves the ability to detect anomalies on the surface of the equipment. In [12], Qi et al. proposed a transformer fault voiceprint identification model based on Mel-frequency cepstral coefficients (MFCC) and deep learning. This model significantly improved the fault identification accuracy of transformer acoustic signals in low signal-to-noise ratio environments. In [13], Li et al. proposed a method based on blind source separation and convolutional neural networks for transformer fault diagnosis through voiceprint identification, effectively separating interference noise from fault noise. Despite the progress made in these studies, most of them focused on single modal data and failed to fully utilize the fusion of image and voiceprint features, thus limiting the comprehensive performance of defect identification.

Defect identification of electrical equipment based on multi-modal image and voiceprint data features can improve the accuracy and robustness. In [14], Zhang et al. proposed a Transformer-based multi-modal fusion method by combining image and voiceprint data, which incorporated a hybrid attention mechanism and a feature reconstruction strategy to accurately identify defects across various power grid equipment types. In [15], a method for evaluating the condition of electrical equipment based on image and voiceprint data fusion was proposed by Qiu et al. This method reduces the complexity of evaluation and maintenance costs for electrical equipment. In [16], a multi-modal fusion method based on image and voiceprint signals was proposed by Wang et al., in which the accuracy of power equipment fault detection is significantly improved by weighting the features through an adaptive attention mechanism. In [17], Bai et al. proposed a hierarchical classification method based on multi-modal feature fusion, which enhances the fine-grained defect identification capability of electrical equipment and addresses complex defect detection

in substation equipment. However, the aforementioned methods have not thoroughly investigated the impact of the complementarity and interleaving of image and voiceprint data features on defect identification. Consequently, these methods struggle to flexibly adjust decision-making basis when dealing with different environments and fault types, thereby resulting in poor generalization in identification.

Despite the progress made in existing research, distribution network electrical equipment defect identification still faces the following challenges. On the one hand, existing defect identification methods often rely on single-modality data, making them susceptible to interference and resulting in incomplete information. Furthermore, these methods fail to deeply explore the complementarity and interwoven nature of multimodal features. This leads to insufficient information utilization and poor adaptive capability, ultimately causing low identification accuracy and degraded model performance.

To address these challenges, a distribution network electrical equipment defect identification based on multi-modal image voiceprint data fusion and channel interleaving is proposed. First, an improved transformer-based multi-modal image and voiceprint data fusion method is introduced to achieve high-accuracy defect identification. In particular, two-dimensional principal component analysis (2DPCA) is used to construct the image feature set of distribution network electrical equipment. Unlike PCA and ICA, 2DPCA maintains the 2D spatial structure by processing matrices directly. This prevents the loss of critical information caused by flattening, which is essential for identifying surface defects such as cracks and corrosion. In comparison, a voiceprint feature model is constructed based on the Mel scale, effectively capturing relationships between human auditory characteristics and frequency units. Second, a multi-channel interleaving mechanism is proposed to enhance the robustness and applicability of the model in different environments. The channel interleaving mechanism establishes dynamic correlations among image, voiceprint, and fusion channels through interleaving neurons. When one channel exhibits low recognition confidence, these neurons activate complementary feature information from other high-confidence channels. Conversely, when a channel demonstrates high confidence, the mechanism reinforces its decision weight while preserving auxiliary judgments from other channels. The device image features, voiceprint features, and fused features are leveraged to calculate fault confidence values. Dynamic weight adjustments are developed based on channel adaptability to activate interleaved neurons. The fault confidence is defined as the probability of the defect class output by the channel divided by the number of defect categories. The dynamic channel weight is defined as the quotient of the confidence level of a channel and the sum of confidence levels of all channels. When a channel's confidence falls below threshold, a weight reduction mechanism is automatically triggered, reallocating its weight to higher-confidence channels. The interleaving neuron activation threshold is related to the confidence of channel. When the absolute difference in confidence between any two channels is less than it, the interleaving neuron activates and outputs the fused feature. This activation logic aims to enhance complementary correlations between high-confidence channel features.

The image identification, voiceprint identification, and fusion identification channels independently perform defect identification and output fault identification probabilities. The input information is expanded to include weighted device features, channel adaptability, and results from other channels. Interleaved influences are then applied to their outputs based on confidence values. The interleaved results are processed using a weight approach based on output probability information content to generate the defect identification outcome. Finally, the effectiveness and reliability of the proposed algorithm are validated through simulations. Comparison of this work with state-of-the-art approaches is shown in Table 1.

Table 1. The differences between the proposed work and other methods.

	Precision Decision-Making Capability Under Incomplete Information	Dynamic Channel Weight Adjustment	Complementarity Between Image and Voiceprint Features	Interleaving Between Image and Voiceprint Features
[14]	√	×	√	×
[15]	×	×	√	×
[16]	√	×	×	×
[17]	×	×	√	×
Proposed	√	√	√	√

Although references [14–17] have attempted multi-modal feature fusion, they still exhibit significant limitations in complementarity mining and interleaving design: Although reference [14] employs a hybrid attention mechanism, it only performs static concatenation at the feature level, lacking the ability to dynamically adjust modal weights according to defect types (e.g., bearing wear requires vibration-scratch synergy, partial discharge requires speckle-pulse correlation). Reference [15] achieves multi-source fusion but image and voiceprint features make independent decisions, lacking cross-modal interaction. Although reference [16] introduces adaptive attention, its weights are based solely on feature similarity, making it susceptible to noise interference and misjudgment of modal priority. Reference [17] proposes a hierarchical classification method but fails to design an inter-channel interleaving mechanism, making it difficult to correct deviations through cross-channel information. As shown in Table 1, existing multimodal methods either lack dynamic weight adjustment driven by channel adaptability or fail to establish an interleaving mechanism between image and voiceprint features. Consequently, they cannot achieve flexible identification where high-confidence channels dominate decision-making while low-confidence channels provide complementary support. The contributions of this paper are specified as follows.

- Improved transformer-based multi-modal image and voiceprint data fusion method: To address the issue of traditional single-confidence methods being susceptible to noise interference and generating false high-confidence judgments, we designed an improved transformer model incorporating intra-domain self-attention units and an inter-domain attention mechanism. This model utilizes cross-domain attention to dynamically evaluate the correlation strength between multiple modalities, thereby enhancing the diagnostic system's accuracy. On one hand, the intra-domain self-attention unit addresses the traditional model's insensitivity to data variations by dynamically focusing on key features within a single modality. On the other hand, the cross-domain attention mechanism tackles the issue of isolated multimodal information in traditional models by establishing semantic correlations across modalities.
- Distribution network electrical equipment defect identification based on image and voiceprint multi-channel interleaving: To address the challenges of significant environmental noise fluctuations in distribution networks and the tendency of single-modal approaches to miss critical information, an adaptive interleaving mechanism for multimodal features is introduced. This mechanism employs cross-domain attention to dynamically assess the confidence level of each modality. It leverages the complementary strengths of image and voiceprint data to enhance identification accuracy while suppressing unreliable signals. This approach prevents misjudgments and ensures comprehensive system performance and robustness in complex environments.

2. Improved Transformer-Based Multi-Modal Image and Voiceprint Data Fusion Method

The transformer model processes and integrates multiple types of data simultaneously by feeding different modalities into a unified model structure. However, traditional transformer models suffer from limitations such as poor information extraction capabilities and high data requirements. The traditional transformer model is improved by incorporating intra-domain self-attention units and inter-domain attention mechanisms. Intra-domain attention focuses on key features within a single modality, while inter-domain attention builds semantic correlations across modalities. Their synergy achieves a fusion effect characterized by "single-modality purification and cross-modality complementarity". Furthermore, an improved transformer-based multi-modal image and voiceprint fusion method is proposed. Firstly, 2DPCA is employed to construct the image feature model, achieving grayscale image compression and extraction of key features both horizontally and vertically. Secondly, a voiceprint feature model is built based on the Mel scale, extracting MFCC and MFSC features through steps including FFT and Mel filtering. Finally, by combining intra-domain self-attention units and inter-domain cross-attention mechanisms, an improved Transformer model is designed to realize feature fusion. This approach enables long-term utilization and global interaction of data features, while effectively preserving complementary image and voiceprint data and better integrating multi-modal features.

The schematic diagram of improved transformer-based multi-modal image and voiceprint data fusion method is shown in Figure 1.

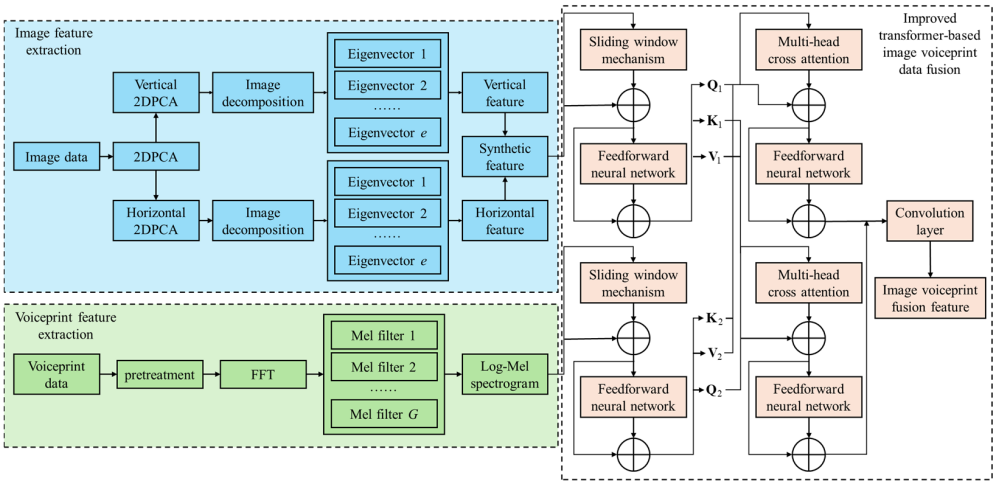


Figure 1. Improved transformer-based multi-modal image and voiceprint data fusion method.

2.1. Construction of Image Feature Models

2DPCA extracts features directly from two-dimensional matrices by performing horizontal and vertical compression on images, avoiding the structural information loss caused by converting images into one-dimensional vectors in traditional PCA and ICA. Traditional PCA flattens images into 1D vectors, inevitably losing the 2D spatial structure critical for identifying defects such as cracks and corrosion [18]. In contrast, 2DPCA performs dimensionality reduction directly on image matrices through horizontal and vertical projections, preserving pixel-level spatial relationships, edge contours, and texture distributions. By combining the compression results to compute eight statistical features, 2DPCA forms a compact yet discriminative feature representation that retains structural detail while reducing noise. Compared with traditional methods such as PCA and ICA, whose feature maps often appear over-concentrated and noisy, it preserves spatial relationships between pixels and effectively extracts discriminative local features. The method also suppresses noise

interference. Consequently, 2DPCA demonstrates higher robustness and identification accuracy in defect detection.

The image sample set is defined as $\mathbf{H} = \{\mathbf{H}_1, \ldots, \mathbf{H}_n, \ldots, \mathbf{H}_N\}$, where $\mathbf{H}_n$ represents the $\mathbf{H}_n$-th image sample. Assuming the dimensions of $\mathbf{H}_n$ are $i \times j$, the 2DPCA method applies a linear transformation to $\mathbf{H}_n$ for horizontal compression, reducing it into a smaller-dimensional matrix, denoted as

$$\mathbf{H}_n^{\text{hor}} = \mathbf{H}_n \mathbf{P}_{\text{opt}} \tag{1}$$

where $\mathbf{H}_n^{\text{hor}}$ represents the $i \times e$-dimensional result obtained after horizontal compression using 2DPCA. e denotes the number of projection axes. $\mathbf{P}_{\text{opt}}$ is the optimal projection matrix.

N_0 training samples $\hat{\mathbf{H}}_k$, i.e., $k = 1, 2, \ldots, N_0$ are selected from the image sample set of distribution network electrical equipment. These samples are used to construct the covariance matrix for 2DPCA, expressed as

$$\mathbf{C} = \frac{1}{N_0} \sum_{k=1}^{k=N_0} \left(\hat{\mathbf{H}}_k - \bar{\mathbf{H}} \right)^{\mathrm{T}} \left(\hat{\mathbf{H}}_k - \bar{\mathbf{H}} \right) \tag{2}$$

where $\bar{\mathbf{H}} = \frac{1}{N_0} \sum_{k=1}^{k=N_0} \hat{\mathbf{H}}_k$ represents the mean matrix of the training samples.

Similarly, the 2DPCA method is applied to $\mathbf{H}_n$ for vertical compression, reducing it into a matrix with smaller dimensions, expressed as

$$\mathbf{H}_n^{\text{ver}} = \mathbf{P}_{\text{opt}} \mathbf{H}_n \tag{3}$$

where the dimension of $\mathbf{H}_n^{\text{ver}}$ is $e \times j$.

$\mathbf{H}_n^{\text{hor}}$ and $\mathbf{H}_n^{\text{ver}}$ are represented as vectors, where $\mathbf{H}_n^{\text{hor}}$ is expressed as

$$\begin{cases} \mathbf{H}_n^{\text{hor}} = [\mathbf{h}_1, \ldots, \mathbf{h}_{e_0}, \ldots, \mathbf{h}_e] \\ \mathbf{h}_{e_0} = [h_{e_0 1}, h_{e_0 2}, \ldots, h_{e_0 i}]^{\mathrm{T}}, e_0 = 1, 2, \ldots, e \end{cases} \tag{4}$$

$\mathbf{H}_n^{\text{ver}}$ is expressed as

$$\begin{cases} \mathbf{H}_n^{\text{ver}} = [\mathbf{v}_1, \ldots, \mathbf{v}_{e_0}, \ldots, \mathbf{v}_e]^{\mathrm{T}} \\ \mathbf{v}_{e_0} = [v_{e_0 1}, v_{e_0 2}, \ldots, v_{e_0 j}], e_0 = 1, 2, \ldots, e \end{cases} \tag{5}$$

As shown in Figure 1, eight feature quantities are extracted from the horizontal 2DPCA image compression vector $\mathbf{h}_{e_0}$, i.e., $e_0 = 1, 2, \ldots, e$ and the vertical image compression vector $\mathbf{v}_{e_0}$, i.e., $e_0 = 1, 2, \ldots, e$. These features include mean, standard deviation, sharpness, sum of squared derivatives, sparsity, steepness, energy, and entropy. The final results consist of the horizontal 2DPCA image decomposition feature $\mathbf{S}_d^{\text{hor}}$, vertical 2DPCA image decomposition feature $\mathbf{S}_d^{\text{ver}}$, and composite feature $\mathbf{S}_d^{\text{agg}}$. The horizontal 2DPCA image decomposition feature $\mathbf{S}_d^{\text{hor}}$ is expressed as

$$\mathbf{S}_d^{\text{hor}} = \begin{bmatrix} \mu_{\mathbf{h}_1}, \sigma_{\mathbf{h}_1}, sh_{\mathbf{h}_1}, su_{\mathbf{h}_1}, sp_{\mathbf{h}_1}, st_{\mathbf{h}_1}, ene_{\mathbf{h}_1}, ent_{\mathbf{h}_1} \\ \mu_{\mathbf{h}_2}, \sigma_{\mathbf{h}_2}, sh_{\mathbf{h}_2}, su_{\mathbf{h}_2}, sp_{\mathbf{h}_2}, st_{\mathbf{h}_2}, ene_{\mathbf{h}_2}, ent_{\mathbf{h}_2} \\ \cdots \\ \mu_{\mathbf{h}_s}, \sigma_{\mathbf{h}_s}, sh_{\mathbf{h}_s}, su_{\mathbf{h}_s}, sp_{\mathbf{h}_s}, st_{\mathbf{h}_s}, ene_{\mathbf{h}_s}, ent_{\mathbf{h}_s} \end{bmatrix} \tag{6}$$

where $\mu_{\mathbf{h}_e}$ represents the mean of $\mathbf{h}_e$. $\sigma_{\mathbf{h}_e}$ represents the standard deviation of $\mathbf{h}_e$. $sh_{\mathbf{h}_e}$ represents the sharpness of $\mathbf{h}_e$. $su_{\mathbf{h}_e}$ represents the sum of squared derivatives of $\mathbf{h}_e$. $sp_{\mathbf{h}_e}$ represents the sparsity of $\mathbf{h}_e$. $st_{\mathbf{h}_e}$ represents the steepness of $\mathbf{h}_e$. $ene_{\mathbf{h}_e}$ represents the energy of $\mathbf{h}_e$. $ent_{\mathbf{h}_e}$ represents the entropy of $\mathbf{h}_e$.

The vertical 2DPCA image decomposition feature $\mathbf{S}_d^{\text{ver}}$ is expressed as

$$\mathbf{S}_d^{\text{ver}} = \begin{bmatrix} \mu_{\mathbf{v}_1}, \sigma_{\mathbf{v}_1}, sh_{\mathbf{v}_1}, su_{\mathbf{v}_1}, sp_{\mathbf{v}_1}, st_{\mathbf{v}_1}, ene_{\mathbf{v}_1}, ent_{\mathbf{v}_1} \\ \mu_{\mathbf{v}_2}, \sigma_{\mathbf{v}_2}, sh_{\mathbf{v}_2}, su_{\mathbf{v}_2}, sp_{\mathbf{v}_2}, st_{\mathbf{v}_2}, ene_{\mathbf{v}_2}, ent_{\mathbf{v}_2} \\ \cdots \\ \mu_{\mathbf{v}_s}, \sigma_{\mathbf{v}_s}, sh_{\mathbf{v}_s}, su_{\mathbf{v}_s}, sp_{\mathbf{v}_s}, st_{\mathbf{v}_s}, ene_{\mathbf{v}_s}, ent_{\mathbf{v}_s} \end{bmatrix} \tag{7}$$

where $\mu_{\mathbf{v}_e}$ represents the mean of $\mathbf{v}_e$. $\sigma_{\mathbf{v}_e}$ represents the standard deviation of $\mathbf{v}_e$. $sh_{\mathbf{v}_e}$ represents the sharpness of $\mathbf{v}_e$. $su_{\mathbf{v}_e}$ represents the sum of squared derivatives of $\mathbf{v}_e$. $sp_{\mathbf{v}_e}$ represents the sparsity of $\mathbf{v}_e$. $st_{\mathbf{v}_e}$ represents the steepness of $\mathbf{v}_e$. $ene_{\mathbf{v}_e}$ represents the energy of $\mathbf{v}_e$. $ent_{\mathbf{v}_e}$ represents the entropy of $\mathbf{v}_e$.

Synthetic feature $\mathbf{S}_d^{\text{agg}}$ is expressed as

$$\mathbf{S}_d^{\text{agg}} = \left[\mathbf{S}_d^{\text{hor}}, \mathbf{S}_d^{\text{ver}}\right] \tag{8}$$

2.2. Construction of Voiceprint Feature Models

The one-dimensional waveform data of voiceprints contains only the time-domain information of the sound source signal. For more effective voiceprint identification, the one-dimensional waveform data must be converted into two-dimensional data in the time-frequency domain for analysis [19]. Distribution equipment fault noises are primarily concentrated in the 200–5000 Hz frequency range. The Mel-scale effectively simulates human ear sensitivity characteristics within this frequency band. By converting linear frequencies to Mel frequencies through Mel-filtering, it highlights frequency differences between fault noises and environmental interference. The process involves first converting voiceprint signals to the frequency domain via FFT. Then, Mel-filtering and logarithmic transformation extract Log–Mel spectra. First and second-order differentials supplement dynamic information. This finally forms a voiceprint feature model containing both frequency domain energy distribution and temporal dynamic variations. This model accurately captures acoustic fingerprints of equipment faults while avoiding environmental noise interference in feature discrimination.

During time-frequency domain analysis of sound data, it is transformed into a spectrogram to capture both time-domain and frequency-domain information of the sound. Human perception of frequency (Hz) is nonlinear, and the Mel scale effectively describes the relationship between human auditory characteristics and frequency units (Hz) [20]. The relationship between Mel frequency and Hz frequency is expressed as

$$f^{\text{Mel}} = 2595 \times \lg\left(\frac{f}{700}\right) \tag{9}$$

As shown in Figure 1, the voiceprint signal undergoes preprocessing steps such as pre-emphasis, framing, and windowing before being subjected to Fast Fourier Transformation (FFT). The transformation results of each frame are then superimposed to ultimately generate a two-dimensional spectrogram [21]. The FFT is expressed as

$$Y(k) = \sum_{b=b-1}^{B-1} x(m) e^{-j\frac{2\pi}{B}bk}, 0 \le k \le B \tag{10}$$

where B represents the number of FFT points, and k denotes the k-th point within an FFT. It is then converted into a power spectrum, expressed as

$$P(k) = \frac{|Y(k)|^2}{B} \tag{11}$$

The Mel filter is used to convert frequencies in the frequency domain into Mel frequencies [22]. The output of the Mel filter is expressed as

$$F_g(k) = \begin{cases} 0, k < f(g-1) \\ \frac{k-f(g-1)}{f(g)-f(g-1)}, f(g-1) < k \le f(g) \\ \frac{f(g+1)-k}{f(g+1)-f(g)}, f(g) < k \le f(g+1) \\ 0, k > f(g+1) \end{cases} \quad (12)$$

where g ranges from 1 to G, and G represents the number of Mel filters. The power spectrum of the voiceprint signal is processed through the filters to obtain the Mel spectrum, expressed as

$$S^{Mel}(g) = \sum_{k=f(g-1)}^{f(g+1)} F_g(k) P(k) \quad (13)$$

After computation, G outputs are obtained. Taking the logarithm of these outputs produces the Log–Mel spectrum. Since the Mel filtering process may result in the loss of some dynamic information in the audio signal, this loss is compensated for using first-order difference, expressed as

$$d_l = \frac{\sum_{r=1}^{R} r(\eta_{l+1} - \eta_{l-r})}{2\sum_{r=1}^{R} r^2} \quad (14)$$

where d_l represents the l-th first-order difference. η_l represents the l-th logarithmic spectral coefficient. R is typically set to 1 or 2, indicating the time interval for the first-order derivative. Computing the first-order difference twice yields the second-order difference, which can also compensate for the loss of dynamic information. Finally, combining the Log–Mel spectrum, first-order difference, and second-order difference results in the Log–Mel spectrum feature of the voiceprint signal, known as the log Mel-frequency spectral coefficients (MFSC) feature.

Compared to the MFS feature, the MFCC feature extraction involves an additional step of discrete cosine transformation (DCT). The Log–Mel spectrum is transformed into the cepstral domain using DCT, followed by first-order and second-order difference calculations, resulting in the MFCC, expressed as

$$MFCC = \sum_{g=g-1}^{G-1} S^{Mel}(g) \cos\left(\frac{r\pi\left(g+\frac{1}{2}\right)}{G}\right), 0 \le g \le G \quad (15)$$

This study jointly extracts MFSC and MFCC to balance robustness and information richness. MFCC yields compact 13-dimensional static coefficients via DCT, offering strong noise robustness and low redundancy that effectively mitigates Transformer overfitting, but inevitably loses high-dimensional details. In contrast, the 26-dimensional MFSC (G = 26) preserves richer time-frequency energy patterns suitable for Transformer's long-range modeling. The final voiceprint feature is formed by channel-wise concatenation in a 65-dimensional joint vector.

2.3. Improved Transformer-Based Multi-Modal Image and Voiceprint Data Fusion

An improved transformer model is constructed using intra-domain self-attention units and inter-domain cross-attention mechanisms to fuse image and voiceprint features. The specific process is described as follows:

Step 1: Information capture. The multi-head attention (MSA) mechanism considers the global feature distribution, helping the model capture information from multiple encoded subspaces.

Step 2: Feature tokens optimization. In order to optimize the feature tokens generated by MSA, a feed forward network (FFN) is applied, which consists of two multilayer perceptron (MLP) layers and one GELU activation function layer.

Step 3: Normalization. Layer normalization (LN) is performed after the execution of MSA and FFN.

Step 4: Intra-domain perception. Residual connections are deployed after these two modules. The intra-domain self-attention process is expressed as

$$\begin{aligned} \{\mathbf{Q}, \mathbf{K}, \mathbf{V}\} &= \{\mathbf{X}\omega^{Q}, \mathbf{X}\omega^{K}\mathbf{X}\omega^{V}\} \\ \mathbf{U} &= \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) \end{aligned} \tag{16}$$

Step 5: Inter-domain cross-attention interaction. In order to further explore shared information between different domains, the entire process of inter-domain cross-attention interaction is expressed as

$$\begin{aligned} \mathbf{U}_1 &= \text{Attention}(\mathbf{Q}_1, \mathbf{K}_2, \mathbf{V}_2) \\ \mathbf{U}_2 &= \text{Attention}(\mathbf{Q}_2, \mathbf{K}_1, \mathbf{V}_1) \end{aligned} \tag{17}$$

As shown in Figure 1, $\{\mathbf{Q}_1, \mathbf{K}_2, \mathbf{V}_2\}$ refers to the features in query domain 2 that are similar to $\mathbf{Q}_1$ in domain 1. Similarly, $\{\mathbf{Q}_2, \mathbf{K}_1, \mathbf{V}_1\}$ refers to the features in query domain 1 that are similar to $\mathbf{Q}_2$ in domain 2. After the S-Transformer processing, convolutional layers are used to integrate local information from different domains.

Step 6: Intra-domain and inter-domain mixed attention are cascaded to achieve effective fusion of complementary multi-modal features.

Qualitative interpretation of fusion mechanism are as follows.

(a) Complementary analysis of image and voiceprint modalities

Image features excel at capturing spatial structural information of equipment surfaces. They can accurately locate the physical position of defects on devices. However, they lack effective identification capability for internal hidden defects. Voiceprint features focus on acoustic signal changes during equipment operation. They capture abnormal noises generated by internal mechanical vibrations or electrical discharges. These features reflect the internal working status of equipment but cannot provide spatial location information of defects. They are also susceptible to environmental noise interference.

The proposed fusion model combines the advantages of both modalities. It uses the spatial positioning capability of images to compensate for the lack of location information in voiceprints. Simultaneously, it employs the hidden defect identification ability of voiceprints to fill the blind spots of image surface limitations. This approach avoids the one-sided identification caused by single-modality methods.

(b) Mechanism of Improved Transformer Fusion Module

Traditional Transformers often suppress certain modality features in multimodal fusion. They frequently fail to capture semantic relationships between modalities. Our improved Transformer addresses this issue through a dual-mechanism approach: intra-domain self-attention and inter-domain cross-attention.

The intra-domain self-attention module autonomously selects key features within a single modality. In the image domain, it focuses on defect areas like cracks and corrosion. It weakens redundant information such as equipment background and lighting variations. In the voiceprint domain, it locks onto frequency bands corresponding to abnormal noises.

It filters out irrelevant signals like environmental electromagnetic interference and airflow noise. This ensures the purity of single-modality features.

The inter-domain cross-attention module establishes semantic correlations across modalities. When identifying composite defects, it links spatial features from images with frequency features from voiceprints. This enables the model to understand that multiple noise sources originate from the same fault. Thus, it achieves genuine deep fusion.

3. Power Quality Disturbance Identification Method Based on Dual-Channel Time-Frequency Feature Fusion Network

The complementary and interleaving characteristics of image and voiceprint features are often overlooked by traditional defect identification methods for distribution network electrical equipment. This oversight limits the effective utilization of complementary information, leading to reduced identification accuracy. Moreover, these methods struggle to adapt decision-making criteria flexibly when faced with varying environments and fault types, leading to poor generalization and adaptability. An interleaving approach for three neural network channels—image, voiceprint, and image voiceprint fusion is introduced. This approach effectively utilizes the complementarity of multi-modal data features, enabling more accurate defect identification. Additionally, a mechanism of interleaved neurons and channel adaptability analysis is incorporated, allowing the output of one network to influence the decision-making process of others. Networks with higher accuracy are assigned greater weight. This interleaving mechanism is combined with a strategy for adaptive weight adjustment. It enables the system to flexibly handle multi-modal data under diverse operation conditions. As a result, precise defect identification for electrical equipment is achieved.

3.1. Multi-Channel Interleaving Architecture of Image and Voiceprint Neural Networks

A total of Q types of electrical equipment with A types of internal defect features are considered, represented as the set $\mathbf{Q} = \{1, \ldots, q, \ldots, Q\}$, where q denotes the q-th type of electrical equipment. The fault set is represented as $\mathbf{D} = \left\{d_{q,1}, \ldots, d_{q,a}, \ldots, d_{q,A}\right\}$, where $d_{q,a}$ indicates the a-th fault of the q-th type of equipment. The number of horizontal projection axes is set to 750 (determined based on cross validation), and the feature dimension after vertical compression is 640. The multi-channel interleaving architecture of image voiceprint neural networks is shown in Figure 2, with the specific structure described as follows.

In the architecture, the image channel extracts spatial defect features using an Enhanced Transformer + Convolution architecture: pre-processed inspection images pass through convolutional layers for low-level texture extraction and a Transformer encoder for high-level structural modeling and long-range spatial correlation. A fully connected layer outputs defect probabilities and confidence scores. The voiceprint channel captures time–frequency features of failure noise via a Mel Filter + Enhanced Transformer' architecture: acoustic signals are transformed into Log–Mel spectra, refined by temporal convolution for local frequency cues, and then modeled by a self-attention encoder for global temporal dependencies. The output layer yields defect probability distributions. The fusion channel integrates multimodal semantics using cross-domain attention + interleaving neurons: cross-domain attention aligns image and acoustic representations, and interleaving neurons perform fine-grained element-level fusion. The final probabilities are normalized using attention weights, addressing the limitations of simple concatenation approaches.

Channel adaptability analysis enhances the suitability of electrical equipment defect identification across various environments, allowing networks with higher accuracy in specific conditions to receive greater weight in the final decision. For example, voiceprint detection is well-suited for dynamic monitoring environments, such as operating motors or

transformers. By monitoring changes in acoustic signals in real time, voiceprint detection can identify early signs of faults, such as wear or looseness, effectively preventing severe equipment damage. On the other hand, image detection is more appropriate for identifying defects with clear visual features, such as surface cracks, corrosion, or dirt on equipment like distribution boxes, transformers, and cable joints. Leveraging image processing technology, image detection enables multidimensional analysis to improve defect identification accuracy, especially in tasks requiring high efficiency and repeatability. This interleaving mechanism and the strategy of adaptive weight adjustment enable the system to flexibly handle multi-modal data under varying operation conditions, thereby achieving precise defect identification for electrical equipment.

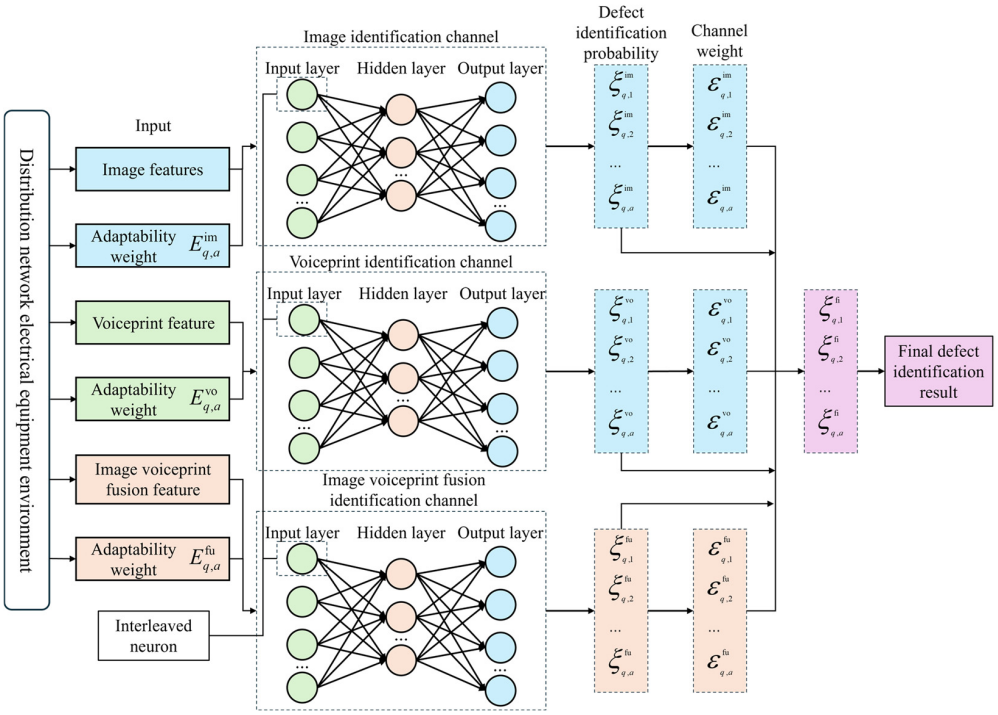


Figure 2. Multi-channel interleaving architecture of image and voiceprint neural networks.

When a network in fault detection exhibits a significantly higher probability for a specific fault compared to others, it indicates that the network has higher accuracy in identifying that particular fault. In this case, the network's output can fully activate the interleaved neurons, enabling them to acquire more reliable knowledge and conclusions. If the network's output probability distribution lacks a clear bias, it indicates higher uncertainty in the identification results. Under such circumstances, the activation level of the interleaved neurons is low, making it impractical to rely solely on a single network's judgment. Consequently, the interleaved neurons are more inclined to consider the opinions of both networks comprehensively. By extracting valid information from the network outputs, the interleaved neurons form a more holistic and reliable judgment, thereby enhancing the system's adaptability in complex environments.

3.2. Distribution Network Electrical Equipment Defect Identification Based on Image and Voiceprint Channel Interleaving

Channel adaptability refers to the ability of the three channels to identify various types of faults for different equipment. Higher channel adaptability indicates that the channel is better suited for defect identification of a specific fault in the given scenario. It can be determined based on historical data or expert knowledge. Through channel adaptability analysis, adaptability weights $E^{\mathrm{im}}_{q,a}$, $E^{\mathrm{vo}}_{q,a}$, and $E^{\mathrm{fu}}_{q,a}$ are derived, enabling the

neural network to adapt to varying environments and fault types. In specific environments, the higher the accuracy of a network channel, the greater the weight it is assigned in the final identification process.

Scenario-adaptive logic of multi-channel interleaving mechanism are introduced as follows.

The multi-channel interleaving mechanism dynamically adjusts each channel's contribution based on the scenario. When strong electromagnetic interference exists, voiceprint channel features become susceptible to noise pollution. In this case, the interleaving mechanism reduces the decision weight of the voiceprint channel. It instead relies more on the spatial features of the image channel and the comprehensive judgment of the fusion channel. This avoids misjudgments caused by noise.

In low-light environments, the visual feature quality of the image channel degrades. The mechanism then increases the weights of both the voiceprint and fusion channels. It utilizes the voiceprint's insensitivity to lighting conditions to compensate for image deficiencies.

Under normal environmental conditions, the mechanism balances the weights across image, voiceprint, and fusion channels. It fully leverages the complementary advantages of each modality. This ensures reliable identification results can be stably output across different scenarios, moving beyond fixed empirical weight allocation.

Based on the fault probabilities $\xi_{q,a}^{\mathrm{im}}$, $\xi_{q,a}^{\mathrm{vo}}$, and $\xi_{q,a}^{\mathrm{fu}}$ output by the three networks, the confidence levels $\Psi_{q,a}^{\mathrm{im}}$, $\Psi_{q,a}^{\mathrm{vo}}$, and $\Psi_{q,a}^{\mathrm{fu}}$ for the fault of the q-th type of equipment are expressed as

$$\Psi_{q,a}^{\mathrm{im}} = \xi_{q,a}^{\mathrm{im}} - \frac{1}{A}\sum_{a=1}^{QA} \xi_{q,a}^{\mathrm{im}} \tag{18}$$

$$\Psi_{q,a}^{\mathrm{vo}} = \xi_{q,a}^{\mathrm{vo}} - \frac{1}{A}\sum_{a=1}^{QA} \xi_{q,a}^{\mathrm{vo}} \tag{19}$$

$$\Psi_{q,a}^{\mathrm{fu}} = \xi_{q,a}^{\mathrm{fu}} - \frac{1}{A}\sum_{a=1}^{QA} \xi_{q,a}^{\mathrm{fu}} \tag{20}$$

The higher the confidence value, the stronger the network's confidence in its output, and the more accurate its defect identification results for the equipment.

By weighting the confidence values calculated by the network with channel adaptability, a comprehensive probability distribution analysis can be obtained. The results of the probability analysis are then fed into the interleaved neurons to derive adaptive weights $\rho_{q,a}^{\mathrm{im}}$, $\rho_{q,a}^{\mathrm{vo}}$, and $\rho_{q,a}^{\mathrm{fu}}$, expressed as

$$\rho_{q,a}^{\mathrm{im}} = \Phi\left(\frac{\Psi_{q,a}^{\mathrm{im}}}{\Psi_{q,a}^{\mathrm{im}} + \Psi_{q,a}^{\mathrm{vo}} + \Psi_{q,a}^{\mathrm{fu}}}\right) E_{q,a}^{\mathrm{im}} \tag{21}$$

$$\rho_{q,a}^{\mathrm{vo}} = \Phi\left(\frac{\Psi_{q,a}^{\mathrm{vo}}}{\Psi_{q,a}^{\mathrm{vo}} + \Psi_{q,a}^{\mathrm{vo}} + \Psi_{q,a}^{\mathrm{vo}}}\right) E_{q,a}^{\mathrm{vo}} \tag{22}$$

$$\rho_{q,a}^{\mathrm{fu}} = \Phi\left(\frac{\Psi_{q,a}^{\mathrm{fu}}}{\Psi_{q,a}^{\mathrm{fu}} + \Psi_{q,a}^{\mathrm{fu}} + \Psi_{q,a}^{\mathrm{fu}}}\right) E_{q,a}^{\mathrm{fu}} \tag{23}$$

where $\Phi(x)$ can be Leaky ReLU activation function.

In each channel, the interleaved neurons initially remain inactive. This ensures that each channel makes independent judgments without interference, maintaining the accuracy and independence of the initial results. When a channel outputs its results, the probability distribution and confidence information are transmitted to the interleaved neurons of other channels for activating them. Adaptive weight values influence the final defect

identification results across the three channels. When a channel exhibits low confidence and poor adaptability, the calculated adaptive weight value decreases. This results in a lower activation level of the corresponding interleaved neurons, which consider the results of all three channels comprehensively. Conversely, the probability distribution becomes more reliant on a single channel when its confidence is high. This mechanism interleaves the image, voiceprint, and fusion channels, allowing their outputs to influence each other. This improves the accuracy of the final defect identification for electrical equipment.

After inputting the three channels into the interleaved neurons using the aforementioned weighting method, the channel output results are weighted using a weighting approach based on output probability information content. The steps are as follows:

Step 1: Normalize the output results of the two neural networks using the range normalization method [23].

Step 2: Calculate $\theta_{q,a}^{\mathrm{im}}$, $\theta_{q,a}^{\mathrm{vo}}$, and $\theta_{q,a}^{\mathrm{fu}}$, where $\theta_{q,a}^{\mathrm{im}}$ represents the proportion of the normalized probability of the q-th type of electrical equipment experiencing the a-th type of fault in the image channel output among all fault types. Similarly, $\theta_{q,a}^{\mathrm{vo}}$ represents the proportion of the normalized probability in the voiceprint channel output. $\theta_{q,a}^{\mathrm{fu}}$ represents the proportion of the normalized probability in the fusion channel output for the same fault scenario.

Step 3: Based on the above proportions, calculate $\beta_{q,a}^{\mathrm{im}}$, $\beta_{q,a}^{\mathrm{vo}}$, and $\beta_{q,a}^{\mathrm{fu}}$. $\beta_{q,a}^{\mathrm{im}}$ represents the validity index of the image channel for the q-th type of electrical equipment. $\beta_{q,a}^{\mathrm{vo}}$ represents the validity index of the voiceprint channel. $\beta_{q,a}^{\mathrm{fu}}$ represents the validity index of the fusion channel. A higher channel validity indicates a greater degree of dispersion in the defect probability output, suggesting that the channel provides more accurate probability predictions for the occurrence of defects in specific scenarios. As a result, the channel will be assigned with greater weight in the comprehensive interleaved identification. The calculation process is expressed as

$$\beta_{q,a}^{\mathrm{im}} = \delta \sum_{a=1}^{QA} \theta_{q,a}^{\mathrm{im}} \ln\left(\theta_{q,a}^{\mathrm{im}}\right) = \delta \sum_{a=1}^{QA} \left[\Theta\left(\xi_{q,a}^{\mathrm{im}}\right) \cdot \ln\left(\Theta\left(\xi_{q,a}^{\mathrm{im}}\right)\right)\right] \tag{24}$$

$$\beta_{q,a}^{\mathrm{vo}} = \delta \sum_{a=1}^{QA} \theta_{q,a}^{\mathrm{vo}} \ln\left(\theta_{q,a}^{\mathrm{vo}}\right) = \delta \sum_{a=1}^{QA} \left[\Theta\left(\xi_{q,a}^{\mathrm{vo}}\right) \cdot \ln\left(\Theta\left(\xi_{q,a}^{\mathrm{vo}}\right)\right)\right] \tag{25}$$

$$\beta_{q,a}^{\mathrm{fu}} = \delta \sum_{a=1}^{QA} \theta_{q,a}^{\mathrm{fu}} \ln\left(\theta_{q,a}^{\mathrm{fu}}\right) = \delta \sum_{a=1}^{QA} \left[\Theta\left(\xi_{q,a}^{\mathrm{fu}}\right) \cdot \ln\left(\Theta\left(\xi_{q,a}^{\mathrm{fu}}\right)\right)\right] \tag{26}$$

where δ is a constant, typically set to $\delta = \frac{1}{\ln A}$. Based on the channel validity indices, the weights of the voiceprint, image, and fusion channels for the q-th type of equipment, $\varepsilon_{q,a}^{\mathrm{im}}$, $\varepsilon_{q,a}^{\mathrm{vo}}$, and $\varepsilon_{q,a}^{\mathrm{fu}}$, are derived and expressed as

$$\varepsilon_{q,a}^{\mathrm{im}} = \frac{1 + \beta_{q,a}^{\mathrm{im}}}{\left(1 + \beta_{q,a}^{\mathrm{im}}\right) + \left(1 + \beta_{q,a}^{\mathrm{vo}}\right) + \left(1 + \beta_{q,a}^{\mathrm{fu}}\right)} \tag{27}$$

$$\varepsilon_{q,a}^{\mathrm{vo}} = \frac{1 + \beta_{q,a}^{\mathrm{vo}}}{\left(1 + \beta_{q,a}^{\mathrm{im}}\right) + \left(1 + \beta_{q,a}^{\mathrm{vo}}\right) + \left(1 + \beta_{q,a}^{\mathrm{fu}}\right)} \tag{28}$$

$$\varepsilon_{q,a}^{\mathrm{fu}} = \frac{1 + \beta_{q,a}^{\mathrm{fu}}}{\left(1 + \beta_{q,a}^{\mathrm{im}}\right) + \left(1 + \beta_{q,a}^{\mathrm{vo}}\right) + \left(1 + \beta_{q,a}^{\mathrm{fu}}\right)} \tag{29}$$

Step 4: Derive the probability distribution $\xi_{q,a}^{\mathrm{fi}}$ output by the interleaved neural network based on the weights $\varepsilon_{q,a}^{\mathrm{im}}$, $\varepsilon_{q,a}^{\mathrm{vo}}$, and $\varepsilon_{q,a}^{\mathrm{fu}}$, expressed as

$$\xi_{q,a}^{\mathrm{fi}} = \varepsilon_{q,a}^{\mathrm{im}} \xi_{q,a}^{\mathrm{im}} + \varepsilon_{q,a}^{\mathrm{vo}} \xi_{q,a}^{\mathrm{vo}} + \varepsilon_{q,a}^{\mathrm{fu}} \xi_{q,a}^{\mathrm{fu}} \tag{30}$$

The fault with the highest probability is selected as the final defect identification result. By introducing interleaved neurons and adaptive weight settings, this method dynamically adjusts the sensitivity to faults based on the outputs and confidence levels of the networks. This enhances the accuracy of defect identification for electrical equipment across diverse environments.

4. Simulation Experiment

We constructed a distribution network electrical equipment dataset comprising 3300 samples, including 3000 defective samples and 300 defect-free samples. The dataset integrates multimodal data, with voiceprints captured by an 8-array linear isometric microphone and images acquired using a 12-megapixel camera. Each sample includes one voiceprint sample and one image sample. The total dataset (3300 samples) was split in a 7:1:2 ratio into a training set, validation set, and test set using a fixed random seed of 42. The image samples are labeled with location and category using the LabelImg tool (version 1.8.6). The duration of each voiceprint sample is 0~5 s, and the voiceprint data in the samples are labeled with categories by Python (version 3.8) code and exported to a csv. file, so that each category of audio file has an independent ID. The dataset covers several practical application scenarios. It includes operation environments of the distribution network, such as humidity, high temperature, and electromagnetic conditions. Additionally, the dataset includes images and voiceprints captured from different angles.

A total of 10 types of electrical equipment are considered, including motors, transformers, distribution boxes, and cable joints. The dataset also includes 4 types of defective features, such as partial discharges, insulation aging, poor contact, and bearing wear. The neural network uses an initial learning rate of 10^{-3} and a cross-entropy loss function. It is trained with the adaptive moment estimation optimizer. The rest of the parameter settings are shown in Table 2 [14,17,24]. All image and voiceprint data have undergone preprocessing including denoising and normalization.

Table 2. Simulation parameters.

Name	Value	Explanation	Name	Value	Explanation
Q	10 types	Number of electrical equipment types	A	4 categories	Number of defect categories
N	3300 pieces	Total number of samples	$2DPCA\text{-}N_0$	750 pieces	Number of training samples
G	26 pieces	Number of Mel Filters	δ	0.558 (constant)	Entropy constant
Learning rate	0.001	Initial learning rate	Depth	12 floors	Transformer depth
T_frame	25 ms	Window length	T_hop	10 ms	Frame shift
e	[1~4] pieces	Number of 2dpca projection axes	fs	16 kHz	Audio sampling rate
i/j	640 dimensions	Image horizontal/vertical dimensions	N_FFT	512	FFT size
M	3	Number of channels			

To verify the effectiveness of the proposed algorithm, five algorithms are selected for comparison.

Baseline 1 [25]: Baseline 1 is a substation equipment defect identification algorithm based on a deep convolutional neural network (DCNN). This algorithm extracts image features of electrical equipment using the DCNN architecture. It then applies transfer learning to identify defects in different types of electrical equipment.

Baseline 2 [5]: Baseline 2 is an electrical equipment defect identification algorithm based on the hidden Markov model (HMM). It extracts voiceprint features of electrical equipment under different operation conditions. Using these voiceprint features and the time-frequency domain characteristics of HMM, the algorithm classifies and identifies electrical equipment defects.

Baseline 3 [17]: Baseline 3 is a hierarchical classification algorithm based on multi-modal feature fusion. This algorithm enhances the fine-grained defect identification capability of electrical equipment by leveraging multi-modal feature fusion. It is designed to address the detection problem of complex defects in substation electrical equipment.

Baseline 4 [14]: Baseline 4 is a Transformer-based multi-modal power grid equipment defect identification algorithm. The algorithm fuses image and voiceprint data while constructing both a hybrid attention mechanism and a feature reconstruction strategy. This approach enables accurate defect identification across various power grid equipment types.

Baseline 5 [15]: Baseline 5 is a power equipment state assessment method based on multi-source fusion (MSF). This algorithm extracts multi-source feature data of power equipment under different operating states, and realizes intelligent assessment and classification recognition of the state of key distribution network equipment based on a feature-level fusion framework and an adaptive weight mechanism.

Baseline 1 and Baseline 2 do not consider the fusion and interleaving of image data and voiceprint data. Baseline 3 although adopting hierarchical label guidance, ignores the collaborative mechanism of multi-channel interleaved neurons and fails to realize real-time confidence interaction and output probability reconstruction among image channel, voiceprint channel and fusion channel. Baseline 4 despite introducing the attention mechanism, lacks a dynamic weight adjustment strategy driven by channel adaptability, making it difficult to adaptively optimize decision weights according to defect types and environmental noise. Baseline 5 achieves multi-source feature fusion, but lacks a cross-modal attention interaction mechanism, limiting its ability to deeply explore the spatio-temporal correlations between image and voiceprint features. All algorithms are evaluated on the same dataset.

Figure 3 shows that the proposed method converges at the 48th iteration, reducing the number of convergence iterations by 22.18% and 39.12% compared to Baseline 3 and Baseline 4, respectively, while the final accuracy is improved by 3.51% to 12.96% compared to the five baseline algorithms. The core reasons lie in the "dual-attention collaborative optimization" of the improved Transformer: Firstly, the intra-domain self-attention unit focuses on key single-modal features through Equation (16), reducing interference from irrelevant features during training. Secondly, the inter-domain attention mechanism mines shared information between image and voiceprint features through Equation (17), accelerating the collaborative optimization of the feature space. In contrast, although Baseline 4 employs Transformer, it lacks the precise filtering of single-modal features by intra-domain attention, leading to feature redundancy during training and slower convergence. Baseline 1, relying solely on image features, suffers from insufficient information dimensions, resulting in early convergence but a 12.96% lower final accuracy. The proposed method achieves dual improvement in convergence speed and accuracy through "single-modal purification + cross-modal collaboration."

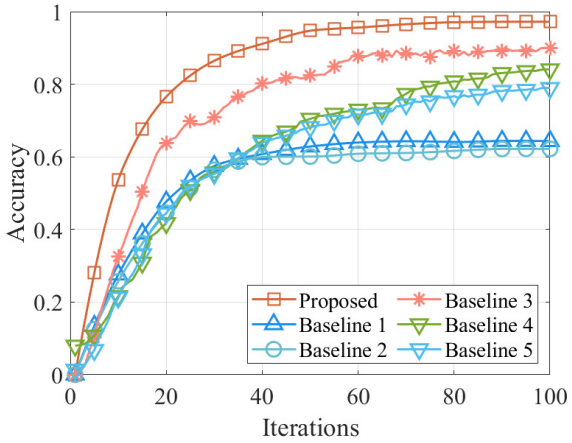


Figure 3. The accuracy of electrical equipment defect identification versus iterations.

Figure 4 shows that the proposed algorithm reduces the loss value by 25.21% compared to the best-performing baseline algorithm at the 100th iteration, while maintaining a lower loss level throughout the entire process. The key lies in the "bias complementary correction" of multi-modal features: Firstly, when blurred images exist in the training samples, voiceprint features can supplement defect information through inter-domain cross-attention, preventing the model from falling into local optima due to single-modal bias. Secondly, the dynamic weights of the multi-channel interleaving mechanism reduce the weight of the blurred image channel, thereby minimizing interference from low-quality features in loss calculation. In contrast, Baseline 1, limited by single-modal information, is prone to loss fluctuations caused by image noise during training. Baseline 2, on the other hand, exhibits slow loss reduction due to voiceprint features being affected by electromagnetic interference. The proposed algorithm achieves rapid and stable loss reduction through multi-modal complementarity and dynamic weight adjustment.

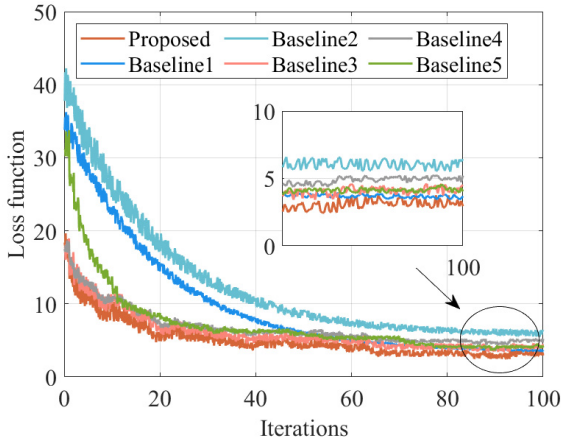


Figure 4. The loss function values of different algorithms versus iterations.

Figure 5 shows that after incorporating image and voiceprint fusion, the proposed algorithm reduces the loss value by 70.33% compared to the case using only image features, and achieves faster convergence speed. The core reason lies in the "feature discriminability enhancement" of the fusion mechanism: Firstly, image features and voiceprint features form a "visual-auditory" complementarity—such as the "outer race scratches" and "200–300 Hz periodic vibrations" of bearing wear. When either feature is used alone, the model easily confuses "wear" with "normal", resulting in persistently high loss. Secondly, the inter-domain cross-attention mechanism of the improved Transformer binds the two types of features through Equation (17), significantly enhancing the defect discriminability of the

fused features and rapidly reducing classification errors during training. In contrast, when using only image features, the model cannot capture vibration information, leading to a high misjudgment rate for "slight wear" and difficulty in reducing the loss. This proves the core value of the fusion mechanism for loss optimization.

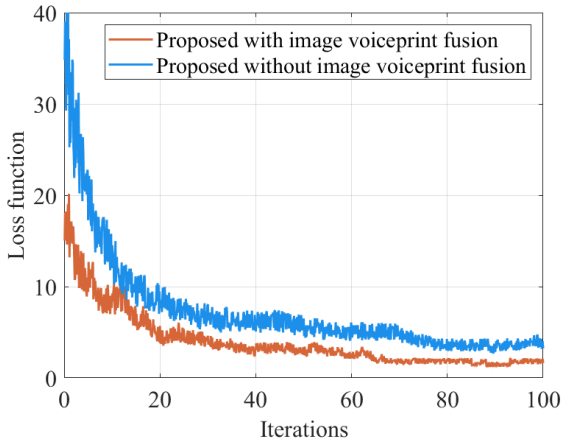


Figure 5. Influence of image voiceprint fusion on the performance of the proposed algorithm.

The core identification metrics of "Image-only", "Voiceprint-only", "Our Fusion Model", and five baseline models are compared in Table 3. The results demonstrate that our fusion model achieves significantly higher accuracy, recall, and F1-score than single-modal approaches. Its accuracy shows a 7.2% improvement over the best single-modal method. It also outperforms all baseline models, achieving a 1.3% accuracy gain over Baseline 4. These results prove that multimodal fusion effectively overcomes single-modal limitations. Furthermore, the improved Transformer fusion mechanism surpasses existing multimodal methods.

Table 3. Comparison of core identification metrics across different model types.

Model Type	Precision (%)	Recall (%)	F1-Score (%)
Image-only	58	81.4	67.7
Voiceprint-only	57	81.9	67.2
Baseline1	63	83.6	71.9
Baseline3	90	96.2	93.0
Baseline4	85	93.9	89.2
Baseline5	79	92.5	85.2
Our Fusion Model	99	99.4	99.2

To verify the necessity of "intra-domain self-attention" and "inter-domain cross-attention" in the improved Transformer, ablation experiments were conducted on the original dataset. The results are shown in Table 4. Removing either component caused significant performance degradation (accuracy dropped by 4.3–5.9%). The convergence speed also slowed down. This proves that "intra-domain self-attention for key feature selection and inter-domain cross-attention for establishing cross-modal correlations" is the core of the fusion model's high performance. It further validates the rationality of the improved design.

Table 4. Comparison of core identification metrics across different fusion model variants.

Model Variant	Precision (%)	Recall (%)	F1-Score (%)
Full model (Proposed)	99.0	99.4	99.2
Proposed without intra-domain self-attention	92.1	96.8	94.4
Proposed without inter-domain cross-attention	90.5	96.4	93.4
Traditional transformer (No Improvement)	88.3	94.7	91.4

Figure 6 shows the accuracy of electrical equipment defect identification versus the number of 2DPCA projection axes. Since Baseline 2 identifies equipment defects only through the voiceprint features of electrical equipment, it is not included in this comparative experiment. It can be seen that, as the number of 2DPCA projection axes increases, the accuracy of defect identification improves for both the proposed algorithm and all baseline algorithms. This result indicates that using 2DPCA to extract image features of electrical equipment is effective for identifying defects. Moreover, the proposed algorithm consistently achieves higher accuracy compared to the baselines. When compared with Baseline 1, Baseline 3, Baseline 4, and Baseline 5, the accuracy of the proposed algorithm is improved by 35.43%, 8.30%, 13.01%, and 19.71%, respectively. A larger number of 2DPCA projection axes can significantly improve the accuracy of electrical equipment defect identification but may lead to an increase in computational complexity. The proposed algorithm uses image data, voiceprint data, and fused data for electrical equipment defect identification, avoiding the decrease in identification accuracy caused by the limitations of single data. It is worth noting that the proposed algorithm, when using the minimum number of projection axes, has already approached the performance level of some baseline algorithms when using the maximum number of projection axes. This means that under the condition of ensuring the same identification accuracy, the proposed algorithm can reduce the computational load of feature dimensions by approximately 75%, providing a more efficient solution for practical engineering applications.

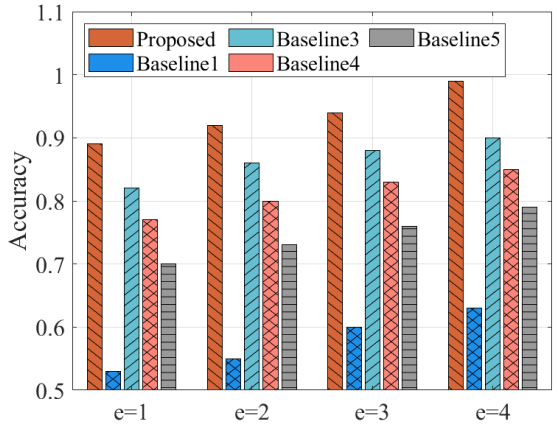


Figure 6. The accuracy of electrical equipment defect identification versus the number of 2DPCA projection axes.

Random noise of 5% to 20% is added to the voiceprint and image samples of partial discharge faults in the dataset, with noise amplitudes of 50 pC and 80 pC, respectively. The electrical equipment defect identification results of the six algorithms after adding random noise are shown in Table 5. It can be seen that with the increase in noise content and noise amplitude, the accuracy of the three algorithms for the identification of electrical equipment defects has decreased, in which the proposed algorithm consistently achieves the highest accuracy for the identification of electrical equipment defects. This is because

the proposed algorithm adopts a multi-channel interleaving mechanism and a weight-adaptive adjustment strategy, which can dynamically allocate weights according to channel confidence. When a certain channel has low confidence, it integrates the results of multiple channels; when a certain channel has high confidence, it relies more on the result of a single channel. This mechanism enables the system to dynamically adjust the weight distribution between channels, flexibly handle multi-modal data under different working conditions, and has good resistance to random noise interference.

Table 5. Electrical equipment defect identification results under noisy environment.

Random Noise		The Accuracy of Electrical Equipment Defect Identification (%)					
Noise Content (%)	**Noise Amplitude (pC)**	**The Proposed Algorithm**	**Baseline 1**	**Baseline 2**	**Baseline 3**	**Baseline 4**	**Baseline 5**
5	50	96.37	60.42	59.91	88.60	82.64	78.09
	80	95.21	59.38	58.03	86.77	82.32	76.52
10	50	93.10	58.13	55.85	84.79	79.73	73.63
	80	92.44	57.30	54.97	84.09	79.60	72.54
20	50	87.86	52.85	50.93	79.74	74.52	69.01
	80	86.57	51.27	49.80	78.09	73.87	67.88

Figure 7 demonstrates that the proposed algorithm achieves the highest identification accuracy across all four defect types, with distinct mechanistic advantages for different defects: firstly, for partial discharge, the inter-domain attention mechanism enhances the 'light spot-pulse' correlation, improving accuracy by 14.01% over Baseline 4; secondly, for insulation aging, interleaved neurons amplify weak feature responses through Equations (21)–(23), increasing accuracy by 12.19% compared to Baseline 2; thirdly, for poor contact, the dynamic weight mechanism reduces blurred image channel weights and relies more on voiceprint features, achieving an 36.05% accuracy improvement over Baseline 1; finally, for bearing wear, 2DPCA captures outer race scratches while the Mel scale extracts vibration features, with interleaved neurons fusing both feature types to reach 93.87% accuracy, representing improvements of 35.91% and 33.29% over Baseline 1 and Baseline 2, respectively. In contrast, Baseline 1 exhibits high misjudgment rates for bearing wear due to lacking voiceprint vibration information, while Baseline 3, though multimodal, lacks dynamic weight adjustment to adapt to dominant feature types across different defects, resulting in reduced accuracy for certain defects.

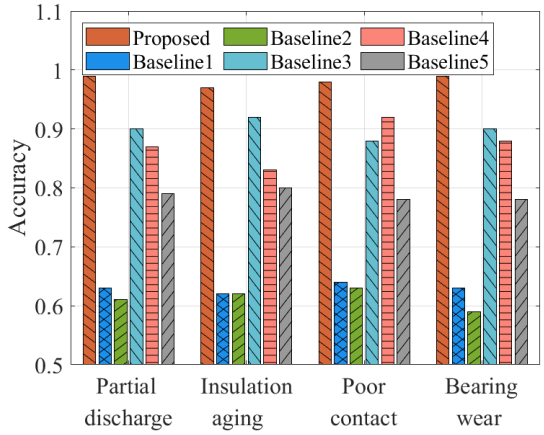


Figure 7. Identification results versus different defect types.

Figure 8 shows single-dimensional relative bullseye degree analysis. The comprehensive performance of each algorithm is demonstrated based on three core evaluation metrics: Recall, Precision, and F1. The calculation of Recall, Precision, and F1 are expressed as

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{31}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{32}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{33}$$

where TP is the number of samples correctly predicted as positive and FN is the number of samples incorrectly predicted as negative (actually with defects but misjudged as defect-free). Recall is a metric to measure the model's ability to correctly identify positive samples, that is, the proportion of actually defective equipment that is correctly detected by the model. Precision measures the proportion of samples predicted as positive by the model that are actually positive, that is, the proportion of equipment claimed as "defective" by the model that are truly defective. FP refers to the number of samples incorrectly predicted as positive (actually defect-free but misjudged as defective). F1 is the harmonic mean of Recall and Precision, balancing the performance of both.

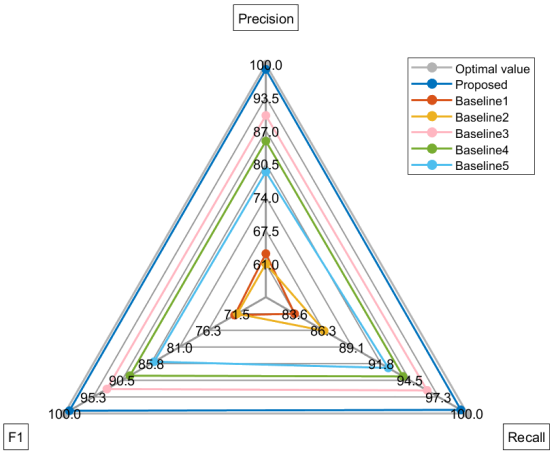


Figure 8. Single-dimensional relative bullseye degree analysis.

The proposed algorithm exhibits excellent performance close to the optimal value in all three dimensions, with its Recall, Precision, and F1 all outperforming those of all baseline algorithms. As a comprehensive measure of Recall and Precision, the F1 is a concise summary of the information in the bullseye chart. Compared with the single-modal Baseline 1 and Baseline 2, the F1 of the proposed algorithm is increased by 3.77% and 4.58%, respectively. Compared with Baseline 3, which performs the second best, the F1 of the proposed algorithm is increased by 0.9%. Compared with Baseline 4 and Baseline 5, the F1 of the proposed algorithm is increased by 1.97% and 2.12%, respectively. This is mainly due to the fact that the proposed algorithm deeply explores the complementary information of image and voiceprint features through the improved intra-domain and inter-domain attention mechanism of Transformer, and dynamically optimizes the decision-making process by combining the adaptive weight strategy of the multi-channel interleaving model. This enables the model to accurately identify defect targets while significantly reducing the risk of missed detection, and finally achieves high-precision identification of electrical equipment defect types.

Figure 9 shows that the proposed algorithm achieves an average identification accuracy of 93.73% for the four defect types, with misjudgment rates below 2% between

insulation aging and poor contact. The key lies in the "dynamic weight error correction" of multi-channel interleaving: Firstly, when processing easily confused defects, channel adaptability analysis calculates the channel confidence for both defect types—in insulation aging scenarios, the image channel confidence reaches 0.82, and the dynamic weight mechanism increases its weight to 55%; in poor contact scenarios, the voiceprint channel confidence reaches 0.85, with its weight raised to 58%, thereby reducing confusion through "defect-dominant channel enhancement". Secondly, interleaved neurons reduce activation intensity for low-confidence channels, avoiding interference from invalid information in decision-making. In contrast, traditional methods, due to fixed weights, cannot adjust channel contributions according to defect types, resulting in misjudgment rates of 8–10% for insulation aging and poor contact, which proves the proposed dynamic weight mechanism's effectiveness in suppressing inter-class confusion.

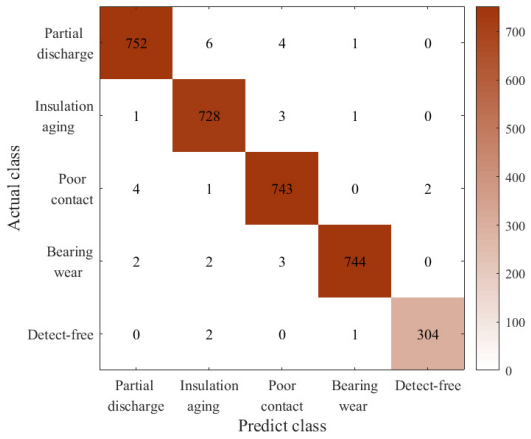


Figure 9. Confusion matrix of defect types for the proposed algorithm.

To further illustrate the model's limitations in complex industrial scenarios, a qualitative error analysis of typical misclassified cases reveals that errors mainly arise from feature coupling between similar defects, environmental interference, and sensor constraints. Partial discharge is often mistaken for insulation aging due to similar carbonized textures or for poor contact because discharge hissing overlaps with arcing sounds in noisy conditions; insulation aging is confused with bearing wear through low-frequency vibration resemblance or with poor contact when discoloration mimics thermal marks under poor lighting; poor contact may be labeled as partial discharge due to shared sizzling/corona acoustics or missed entirely when visually occluded and acoustically faint; bearing wear is misjudged as insulation aging from lubricant/rust textures or as poor contact when severe clicking mimics sparking; defect-free equipment generates false positives for aging due to shadows/dirt being interpreted as deterioration textures or for bearing wear when wind or adjacent machinery noise is misread as friction. These errors underscore the challenges of tight visual-acoustic coupling and limited robustness in real industrial environments.

Table 6 comprehensively evaluates the performance of the proposed algorithm and baseline algorithms in terms of operational efficiency. The proposed algorithm only takes 80.34 s to complete model convergence, which is slightly slower than Baseline 1 and Baseline 2 with relatively simple models. It reduces the convergence time by 14.72% compared with Baseline 3 and by 28.41% compared with Baseline 5, which has the slowest convergence time. This is mainly attributed to the innovative multi-modal feature fusion mechanism. Through the collaborative training of image and voiceprint channels, the feature space is rapidly optimized during the iteration process. Moreover, the proposed algorithm still maintains excellent stability under the condition of 70% memory usage. Its multi-channel interactive training architecture effectively balances computing resources

and performance requirements. Furthermore, the proposed algorithm is 36.4% better than Baseline 3 with the same memory usage, which verifies the significant role of the dynamic weight allocation strategy in accelerating model optimization.

Table 6. The comparisons between the proposed algorithm and other baselines in terms of operational efficiency.

Algorithms	Runtime	Memory Overhead	Convergence Time
Proposed	312.35 s	70%	80.34 s
Baseline 1	267.43 s	60%	71.28 s
Baseline 2	258.85 s	55%	66.79 s
Baseline 3	351.26 s	70%	94.91 s
Baseline 4	337.03 s	66%	107.55 s
Baseline 5	323.79 s	63%	112.21 s

Figure 10 demonstrates the stability performance under adversarial conditions where the audio channel encounters severe noise interference during time steps 60–90, causing its error rate to spike drastically. In this scenario, Baseline 3 exhibits degradation with accuracy plummeting to 75.9%—it lacks multi-channel interleaved neuron collaboration and real-time confidence interaction, allowing the image channel to monopolize adaptive weights; Baseline 4 shows moderate degradation with minimum accuracy of 70.6% and fails to fully recover, as it lacks channel adaptability-driven dynamic weight adjustment to optimize decision weights under noise. In contrast, the proposed algorithm maintains remarkable stability with minimum accuracy of 92.1%, outperforming the baselines by 16.2% and 21.5%, respectively—this is because of its dual-stability mechanism. Firstly, the dynamic weight adjustment mechanism, guided by both channel adaptability and output confidence, ensures that weight allocation reflects each channel's suitability to the current scenario while preventing excessive reliance on channels that initially exhibit high confidence. Secondly, the combination of multi-channel interleaving and adaptive weighting helps stabilize the numerical behavior of the model and mitigates overconfidence issues.

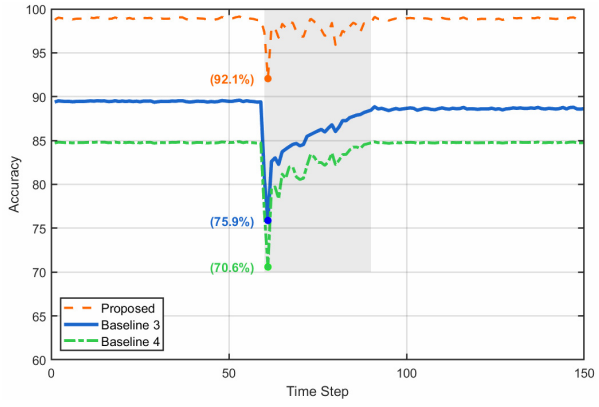


Figure 10. Robustness Comparison Under Adversarial Noise.

5. Conclusions

To address the problems of low accuracy and poor generalization of existing defect identification methods for electrical equipment in distribution networks, a distribution network electrical equipment defect identification method based on image voiceprint multi-modal data fusion and channel interleaving was proposed. Firstly, unified heterogeneous signal representation is achieved using MFCC and 2DPCA to transform acoustic frequency-

domain and vibration time-domain signals into an isomorphic feature space, resolving the fusion challenge of different physical quantities. Secondly, an adaptive multi-modal fusion mechanism is designed through a multi-channel feature entanglement module that captures cross-modal time-frequency correlations and dynamically adjusts acoustic-vibration feature weights according to equipment types, achieving equipment-specific intelligent fusion. Thirdly, long-range modeling of fault evolution is realized via improved Transformer architecture to capture causal dependencies of progressive faults in time series, enhancing early-stage fault detection sensitivity.

Simulation results show that the proposed algorithm has better performance in defect identification of electrical equipment in distribution networks. Compared to the unimodal method that only uses image or voiceprint data (Baseline 1) and the multimodal method without channel interleaving and adaptive attention mechanism (Baseline 2), the proposed algorithm converges at the 48th iteration, with the number of iterations required for convergence increasing by 20.83% and 14.58%, respectively. When the number of iterations reaches 200, the accuracy of electrical equipment defect identification increases by 6.27% and 12.96%, respectively. Furthermore, the proposed algorithm achieves higher recognition accuracy for partial discharge by 3.26% and 4.39%, for insulation aging by 3.37% and 12.19%, for poor contact by 8.05% and 10.59%, and for bearing wear by 11.91% and 3.29%, respectively, compared to Baseline 1 and Baseline 2. Future research can further enhance the model's noise resistance by integrating noise reduction techniques and reinforcement learning methods to improve defect identification under extreme conditions. We will also advance field deployment and optimization in two phases: firstly, a 6-month trial in three climate-specific substations using calibrated anti-drift microphones and waterproof cameras to build a lab-field cross-domain dataset with real interference; secondly, applying domain adaptation and continual learning to align data distributions and update parameters dynamically for reliable on-site defect identification.

Author Contributions: Conceptualization, A.C. and J.L. (Jiaxuan Lu); Funding acquisition, A.C. and J.L. (Junle Liu); Investigation, W.Z., J.L. (Jiaxuan Lu) and B.L.; Methodology, A.C., J.L. (Jiaxuan Lu) and B.L.; Writing—original draft, J.L. (Junle Liu), J.L. (Jiaxuan Lu) and J.Y.; Writing—review and editing, A.C., J.L. (Junle Liu) and B.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Science and Technology Project of China Southern Power Grid Company under Grant Number GDKJXM20240115.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: Authors An Chen, Junle Liu, and Wenhao Zhang were employed by the company China Southern Power Grid Guangdong Zhongshan Power Supply Bureau. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The Science and Technology Project of China Southern Power Grid Company had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Chaves, T.R.; Martins, M.A.I.; Martins, K.A.; Macedo, A.F. Development of an Automated Distribution Grid with the Application of New Technologies. *IEEE Access* **2022**, *10*, 9431–9445. [CrossRef]
2. Zhou, N.; Xu, Y. A Prioritization Method for Switchgear Maintenance based on Equipment Failure Mode Analysis and Integrated Risk Assessment. *IEEE Trans. Power Deliv.* **2023**, *39*, 728–739. [CrossRef]
3. Hatziargyriou, N.; Milanovic, J.; Rahmann, C.; Ajjarapu, V.; Canizares, C.; Erlich, I.; Hill, D.; Hiskens, I.; Kamwa, I.; Pal, B.; et al. Definition and Classification of Power System Stability–Revisited & Extended. *IEEE Trans. Power Syst.* **2022**, *36*, 3271–3281.

4. Huang, X.; Zhang, X.; Zhang, Y.; Zhao, L. A Method of Identifying Rust Status of Dampers based on Image Processing. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 5407–5417. [CrossRef]
5. Yu, Z.; Wei, Y.; Niu, B.; Zhang, X. Automatic Condition Monitoring and Fault Diagnosis System for Power Transformers based on Voiceprint Recognition. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 9600411. [CrossRef]
6. Liu, Y.; Zhao, Y.; Wang, L.; Fang, C.; Xie, B.; Cui, L. High-Impedance Fault Detection Method based on Feature Extraction and Synchronous Data Divergence Discrimination in Distribution Networks. *J. Mod. Power Syst. Clean. Energy* **2022**, *11*, 1235–1246. [CrossRef]
7. Zhao, S.; Blaabjerg, F.; Wang, H. An Overview of Artificial Intelligence Applications for Power Electronics. *IEEE Trans. Power Electron.* **2020**, *36*, 4633–4658. [CrossRef]
8. Li, J.; Song, G.; Yan, J.; Li, Y.; Xu, Z. Data-Driven Fault Detection and Classification for MTDC Systems by Integrating HCTSA and Softmax Regression. *IEEE Trans. Power Deliv.* **2021**, *37*, 893–904. [CrossRef]
9. Mei, B.; Han, R.; Jiang, X.; Wang, Y.; Yin, D. Failure Detection of Infrared Thermal Imaging Power Equipment based on Improved DenseNet. In Proceedings of the 2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Shanghai, China, 23–25 October 2021; pp. 1–5.
10. Luo, Y.; Wang, F.; Liu, X. Infrared and Visible Image Fusion via General Feature Embedding from CLIP and DINOv2. *IEEE Access* **2024**, *12*, 99362–99371. [CrossRef]
11. Lin, Y.; Zhang, W.; Zhang, H.; Bai, D.; Li, J.; Xu, R. An Intelligent Infrared Image Fault Diagnosis for Electrical Equipment. In Proceedings of the 2020 5th Asia Conference on Power and Electrical Engineering (ACPEE), Chengdu, China, 4–7 June 2020; pp. 1829–1833.
12. Qi, X.; Shi, L.; Li, X.; Hao, C.; Ji, S.; Chai, F.; Han, D. Transformer Voiceprint Feature Extraction and Fault Recognition Based on MFCC and Deep Learning. In Proceedings of the 2023 IEEE 5th International Conference on Power, Intelligent Computing and Systems (ICPICS), Shenyang, China, 14–16 July 2023; pp. 820–825.
13. Min, L.; Huamao, Z.; Annan, Q. Voiceprint Recognition of Transformer Fault Based on Blind Source Separation and Convolutional Neural Network. In Proceedings of the 2021 IEEE Electrical Insulation Conference (EIC), Denver, CO, USA, 7–28 June 2021; pp. 618–621.
14. Zhang, G.; Du, Z.; Zhang, Y.; Wang, B.; Chen, J.; Zhang, X. Construction and Optimization of Multimodal Perception Model for Defect Identification of Power Grid Equipment. *Artif. Intell. Sci. Eng.* **2024**, *2*, 36–41.
15. Qiu, J.; Liang, Y.; Cheng, X.; Zhao, X.; Ma, L. State Assessment of Key Equipment of Microgrid based on Multi-Source Data Fusion Method. *IEEE Access* **2022**, *10*, 710–714.
16. Ye, X.; Milanovic, J. Composite Index for Comprehensive Assessment of Power System Transient Stability. *IEEE Trans. Power Syst.* **2022**, *37*, 2847–2857. [CrossRef]
17. Bai, Y.; Wang, L.; Gao, W.; Ma, Y. Multi-Modal Hierarchical Classification for Power Equipment Defect Detection. *J. Image Graph.* **2024**, *29*, 2011–2023. [CrossRef]
18. Zhou, G.; Xu, G.; Hao, J.; Chen, S.; Xu, J.; Zheng, X. Generalized Centered 2-D Principal Component Analysis. *IEEE Trans. Cybern.* **2019**, *51*, 1666–1677. [CrossRef]
19. Zhao, H.; Yue, L.; Wang, W.; Zeng, X. Research on End-to-End Voiceprint Recognition Model based on Convolutional Neural Network. *J. Web Eng.* **2021**, *20*, 1573–1586. [CrossRef]
20. Lai, S.-C.; Hung, Y.-H.; Zhu, Y.-C.; Wang, S.-T.; Huang, Q.-X.; Sheu, M.-H.; Juang, W.-H. Hardware Accelerator Design of DCT Algorithm with Unique-Group Cosine Coefficients for Mel-Scale Frequency Cepstral Coefficients. *IEEE Access* **2022**, *10*, 79681–79688. [CrossRef]
21. Li, Q.; Yang, A.; Guo, P.; Xin, X. High accuracy and low complexity frequency offset estimation method based on all phase FFT for M-QAM coherent optical systems. *IEEE Photonics J.* **2021**, *14*, 7200806. [CrossRef]
22. Ushijima, T.; Tachioka, Y.; Uenohara, S.; Furuya, K. Sparse Independent Vector Analysis based on Mel Filter. In Proceedings of the 2020 IEEE 9th Global Conference on Consumer Electronics (GCCE), Kobe, Japan, 13–16 October 2020; pp. 555–556.
23. Liu, F.; Cui, Y.; Masouros, C.; Xu, J.; Han, T.X.; Eldar, Y.C.; Buzzi, S. Integrated Sensing and Communications: Toward Dual-Functional Wireless Networks for 6G and Beyond. *IEEE J. Sel. Areas Commun.* **2022**, *40*, 1728–1767. [CrossRef]
24. Guan, X.; Gao, W.; Peng, H.; Shu, N.; Gao, D.W. Image-Based Incipient Fault Classification of Electrical Substation Equipment by Transfer Learning of Deep Convolutional Neural Network. *IEEE Can. J. Electr. Comput. Eng.* **2021**, *45*, 1–8. [CrossRef]
25. Han, K.; Xiao, A.; Wu, E.; Guo, J.; Xu, C.; Wang, Y. Transformer in Transformer. *arXiv* **2021**, arXiv:2103.00112.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.