



ARTICLE

# An Improved Res-Swin Fusion Network for Partial Discharge Pattern Recognition in Power Cables

Zihao Jia<sup>1</sup> and Jingrui Zhang<sup>1,2,\*</sup>

<sup>1</sup>Department of Instrumental and Electrical Engineering, Xiamen University, Xiamen, China

<sup>2</sup>Shenzhen Research Institute of Xiamen University, Xiamen University, Shenzhen, China

\*Corresponding Author: Jingrui Zhang. Email: zjrhust@gmail.com

Received: 21 May 2026; Accepted: 10 August 2026; Published: 22 October 2026

**ABSTRACT:** Accurate partial discharge (PD) pattern recognition is essential for assessing power cable insulation, yet it remains challenging due to complex discharge features and limited data availability in practical engineering. To address these issues, this paper first establishes a high-voltage experimental platform with four carefully designed physical defect models to construct a highly authentic Phase-Resolved Partial Discharge (PRPD) dataset. Subsequently, an improved Res-Swin fusion framework is proposed. It integrates a ResNet branch to decouple fine-grained local discharge textures and a Swin Transformer branch to model global phase-amplitude topologies. To alleviate overfitting in data-scarce scenarios, a multi-stage parameter freezing strategy is implemented. Furthermore, an Adaptive Chaotic Lévy-flight Sand Cat Swarm Optimization (ACLSCSO) algorithm is introduced to globally optimize branch-wise initial learning rates for the heterogeneous network. Extensive cross-validation experiments demonstrate that the proposed method achieves an average accuracy of 97.54%. The freezing strategy significantly reduces training time by 11.5%, while the synergistic optimization resolves the confusion between morphologically similar patterns, reducing the misclassification rate of “No Partial Discharge” as “Void Discharge” from 6.2% to 2.1%. Finally, attention visualizations show that the network’s focal regions are broadly consistent with characteristic PRPD regions, providing qualitative interpretability evidence for insulation condition assessment.

**KEYWORDS:** Deep learning; partial discharge pattern recognition; PRPD; Res-Swin; swarm optimization

## 1 Introduction

Power cables constitute a critical infrastructure for electric power transmission and distribution in modern power systems [1]. However, cable-related failures account for a staggering 81.1% of grid faults [2], a vulnerability particularly acute in aging high-voltage installations [3], leading to large-scale outages and severe safety hazards. Insulation degradation is widely recognized as the primary catalyst for such failures [4]. During long-term operation, localized electrical discharges—known as partial discharges (PD)—inevitably develop, acting as both a key indicator and an accelerator of the degradation process [5]. Consequently, continuous PD monitoring is indispensable for equipment condition assessment. Since distinct PD modalities strongly correlate with specific insulation defects (e.g., moisture triggers surface discharge, while isolated conductors generate suspended discharge) [6,7], the accurate classification of PD types provides vital information for defect localization and targeted maintenance [8].

PD detection techniques are broadly categorized into offline and online methodologies [9]. Although offline testing provides controlled environments with reduced background interference, online ultra-high-frequency (UHF) sensing has emerged as a superior continuous monitoring technique due to its robust resistance to electromagnetic noise [10]. The acquired complex signals are universally transformed into Phase-Resolved Partial Discharge (PRPD) maps, effectively characterizing amplitude, phase, and repetition density [11]. Nevertheless, traditional diagnostic approaches relying on manual feature extraction and subjective expert interpretation suffer from limited efficiency and poor robustness under complex on-site conditions [12].

With the rapid advancement of artificial intelligence, deep learning has fundamentally revolutionized PD pattern recognition [13,14]. Advanced visual diagnostic methods and continuous wavelet transform (CWT)-based approaches have further enriched this domain [15]. Specifically, Convolutional Neural Networks (CNNs) have been extensively utilized to automate feature extraction [16], with classic architectures like VGG [17] and ResNet [18] rigorously evaluated for PRPD interpretability. Modified architectures, such as CNN-Long Short-Term Memory (LSTM) networks, further address complex spatial-temporal interferences [19]. Despite these successes, standard CNNs are inherently constrained by their localized receptive fields. When confronted with morphologically similar PRPD patterns, they often fail to capture critical global phase-amplitude correlations, restricting their generalization ability.

To transcend the limitations of local feature extraction, attention-based architectures inspired by the foundational Transformer [20], such as the Swin Transformer [21], have been introduced to model long-range topological dependencies in PD diagnostics [22,23]. Recognizing the respective merits of different neural topologies, recent studies have explored hybrid frameworks—combining CNNs with Transformers or recurrent networks—to capture complementary multi-scale representations [24]. However, simply concatenating these heterogeneous modules frequently introduces severe overfitting on inherently small-scale PD datasets, highly redundant feature extraction, and prohibitive computational overhead that restricts practical deployment. Additionally, the diagnostic efficacy of these neural networks is exquisitely sensitive to hyperparameter configurations [25]. While swarm intelligence and Bayesian algorithms have successfully tuned single-architecture networks [26], and generative adversarial networks have mitigated data scarcity [27], existing optimizers frequently encounter premature convergence when attempting to globally tune high-dimensional, correlated hyperparameters across heterogeneous hybrid branches. Consequently, current PD methodologies face a distinct bottleneck: single networks lack comprehensive perception, whereas existing hybrids are critically hindered by overfitting, redundancy, and suboptimal tuning strategies.

Motivated by these persistent challenges, this paper proposes an improved Res-Swin fusion diagnostic framework grounded in highly authentic physical defect experiments. To systematically overcome the aforementioned limitations, a dual-stream architecture integrating ResNet and Swin Transformer synergistically decouples localized discharge textures from global phase dependencies. To alleviate overfitting and reduce computational overhead in data-scarce scenarios, a multi-stage training freezing strategy coupled with cosine learning rate decay is implemented. Finally, an ACLSCSO algorithm is introduced to optimize branch-wise initial learning rates. The experimental results indicate that the proposed method provides competitive performance in accuracy and computational efficiency.

## 2 Experimental Setup and Dataset Construction

An experimental platform is established to acquire partial discharge (PD) signals that reflect actual cable insulation conditions. The overall experimental circuit topology and the corresponding physical defect models are illustrated in Fig. 1.

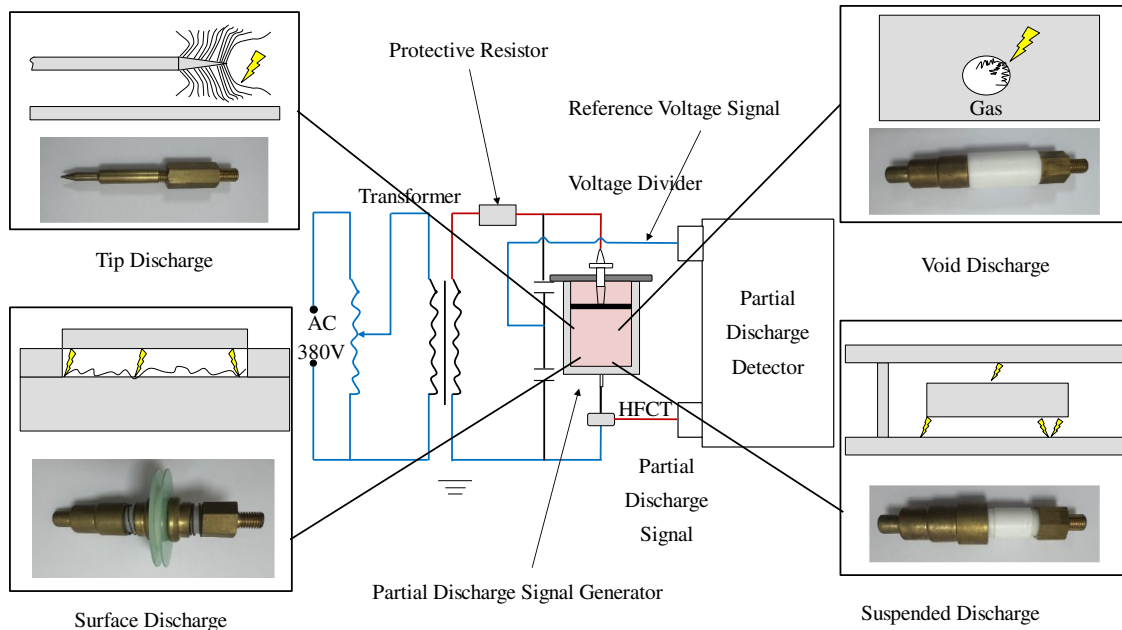


**Figure 1:** Experimental setup and physical defect models.

## 2.1 Experimental Platform and Defect Simulation

As shown in the central schematic of Fig. 2, the PD detection platform primarily consists of a voltage regulator, a high-voltage test transformer, a protective resistor, a capacitive voltage divider, and a specialized testing vessel. Rather than relying on synthetic datasets or public benchmarks, this study used PRPD data obtained from four representative physical defect models developed through an industrial collaboration. These models were designed to simulate typical insulation failures in practical power cables, including Tip Discharge caused by local electric-field enhancement, Void Discharge associated with internal air-gap defects, Surface Discharge along weakened insulation interfaces, and Suspended Discharge induced by isolated conductive components. The experimental platform photographs and representative defect-model images are provided in Figs. 1 and 2. However, fabrication-level parameters of these proprietary industrial specimens, including exact geometric dimensions, internal material composition, electrode spacing, tip radius, void geometry, and detailed surface-discharge configuration, were not authorized for publication or redistribution. Therefore, this study reports the maximum available non-proprietary experimental information and defines the reproducible scope at the PRPD-image, preprocessing, model-training, and evaluation-protocol levels.

During the experiments, an alternating current (AC) power supply of 10 kV at 50 Hz was applied to stress these defect models and trigger stable discharge events. To capture the discharge characteristics comprehensively, dual-sensing technology was employed. A high-frequency current transformer (HFCT, model SH-JF-60; frequency range: 0.3–100 MHz; sensitivity: 0.1 pC; transfer impedance: 18 mV/mA; aperture: 60 mm; IP68) was utilized to monitor grounding currents. Concurrently, an ultra-high-frequency sensor (UHF; frequency range: 300–1500 MHz; detection sensitivity: 10 pC; dynamic range: –85 to –10 dBm; Voltage Standing Wave Ratio (VSWR)  $\leq 3$ ) was deployed to capture the electromagnetic radiation emitted by PD events, with all signals recorded in dB units. The HFCT sensor was factory-calibrated according to IEC 60270 requirements for conventional partial discharge measurements. The UHF sensor was calibrated according to the manufacturer's specifications for electromagnetic detection, and its sensitivity was verified on-site prior to data acquisition using a calibrated discharge source of known magnitude.

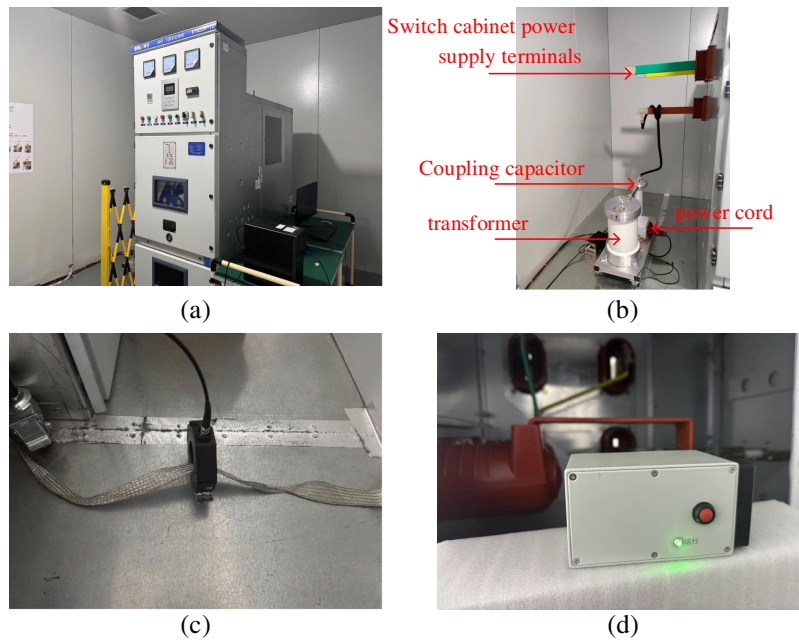


**Figure 2:** The experimental platform for partial discharge detection. (a) Experimental platform (b) Important components (c) HFCT (d) UHF sensor.

## 2.2 PRPD Spectrum Generation and Dataset Construction

In addition to the four defect conditions, a “No Partial Discharge” (No PD) baseline was also continuously recorded under the same sensing and acquisition configuration to ensure the comprehensiveness of the diagnostic model. It should be noted that the No PD category is not interpreted as an entirely noise-free signal. Low-level background responses may still appear because the measurement system, sensors, and acquisition environment remain active during baseline recording. Therefore, the diagnostic distinction of the No PD category is based on the absence of stable phase-resolved discharge clusters and high-density discharge structures, rather than on the complete absence of recorded signal energy.

During data acquisition, the HFCT signal was recorded over 1 s with 50,000,000 sampling points, corresponding to an original sampling rate of 50 MHz. Before PRPD map construction, the signal records were processed into phase-windowed amplitude sequences for subsequent statistical mapping. Since the exact front-end conditioning settings used in the industrial acquisition system cannot be fully disclosed, the present work focuses on the PRPD-level representation, preprocessing procedure, model-training configuration, and evaluation protocol. As an additional check, a representative 1-s HFCT record was reprocessed under decimation factors of 2, 4, 8, and 16 using anti-aliasing filtering before PRPD construction. The resulting PRPD maps showed stable morphology, with Structural Similarity Index Measure (SSIM) values ranging from 0.8296 to 0.8404 relative to the original 50 MHz signal-derived PRPD map.

Subsequently, the processed UHF signal records were converted into two-dimensional Phase-Resolved Partial Discharge (PRPD) maps using the multi-cycle statistical mapping procedure described in our previous work [28]. Representative PRPD spectra for the five insulation conditions are presented in Fig. 3. Each processed UHF record contains record-level metadata, including a record identifier, acquisition timestamp, device identifier, phase-window count, and cycle count. In the released dataset, each retained record follows a consistent structure of 50 power-frequency cycles and 60 phase windows, resulting in 3000 phase-windowed amplitude values for PRPD construction. Briefly, the 50 Hz power-frequency cycle was

used as the phase reference, and the amplitude values were statistically counted within phase-amplitude grids to obtain the pulse repetition density. In the resulting PRPD map, the horizontal axis represents the phase within one power-frequency cycle, the vertical axis represents the normalized amplitude range, and the color intensity represents the density of discharge events. Since the processed UHF records already contained phase-windowed amplitude values, no additional pulse-detection threshold was introduced during PRPD map drawing.

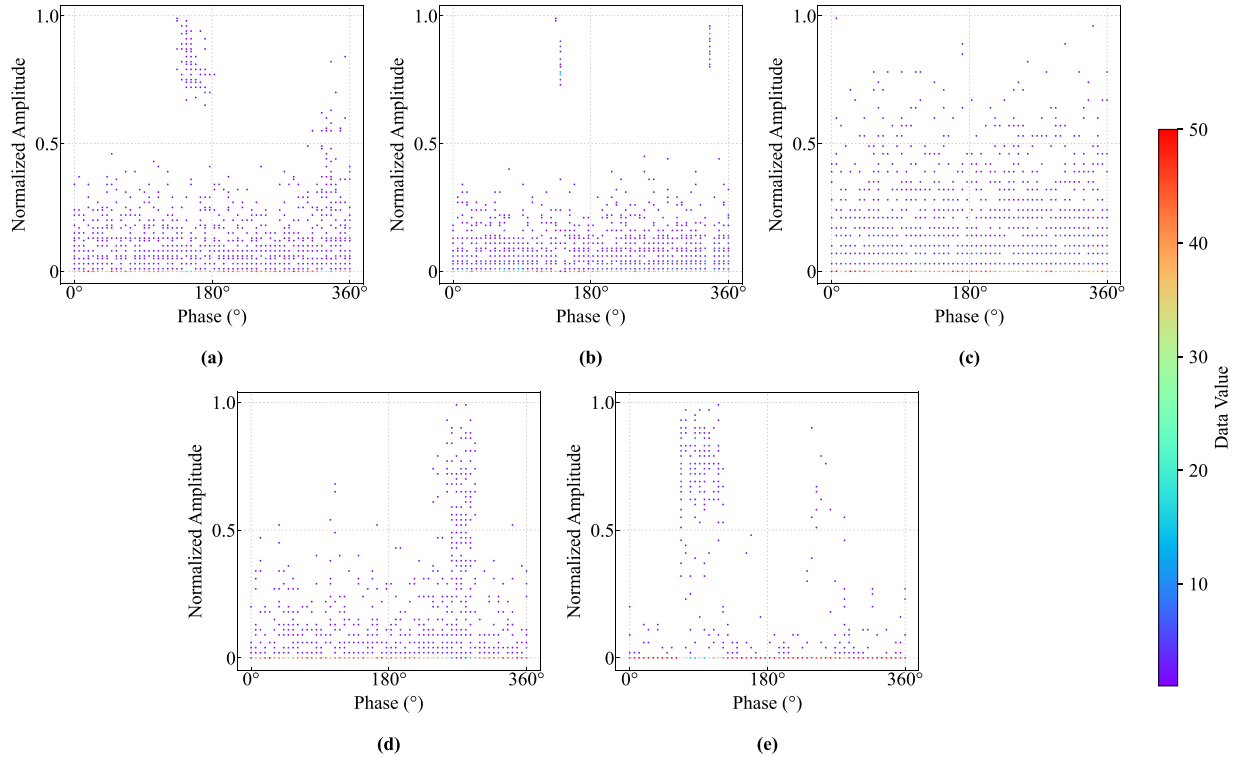


**Figure 3:** Generated PRPD spectra for the five categories of insulation conditions. (a) Tip discharge (b) void discharge (c) no partial discharge (d) suspended discharge (e) surface discharge.

Through this experimental procedure, a dataset comprising 1422 PRPD samples with a uniform resolution of  $515 \times 389$  pixels was generated. The detailed quantitative composition of the generated PRPD patterns across the five categories is summarized in Table 1. No fixed training/testing split was used as the final evaluation protocol. Instead, the final model evaluation was conducted using stratified five-fold cross-validation on the complete dataset, as described in Section 4.2.3.

**Table 1:** Quantitative composition of PRPD maps across different insulation conditions.

|                      | Number | Proportion (%) |
|----------------------|--------|----------------|
| Tip Discharge        | 294    | 20.68          |
| Surface Discharge    | 463    | 32.56          |
| No Partial Discharge | 240    | 16.88          |
| Void Discharge       | 58     | 4.08           |
| Suspended Discharge  | 367    | 25.81          |
| Total                | 1422   | 100.00         |

### 3 Proposed Hybrid Diagnostic Framework

The limitations of single-architecture models and the insufficient utilization of local and global feature information motivate the development of a Res-Swin fusion network. As illustrated in Fig. 4, a dual-stream parallel architecture is adopted in the proposed model, where a convolutional branch and an attention-based branch are integrated to construct a multi-dimensional feature representation for partial discharge diagnosis.

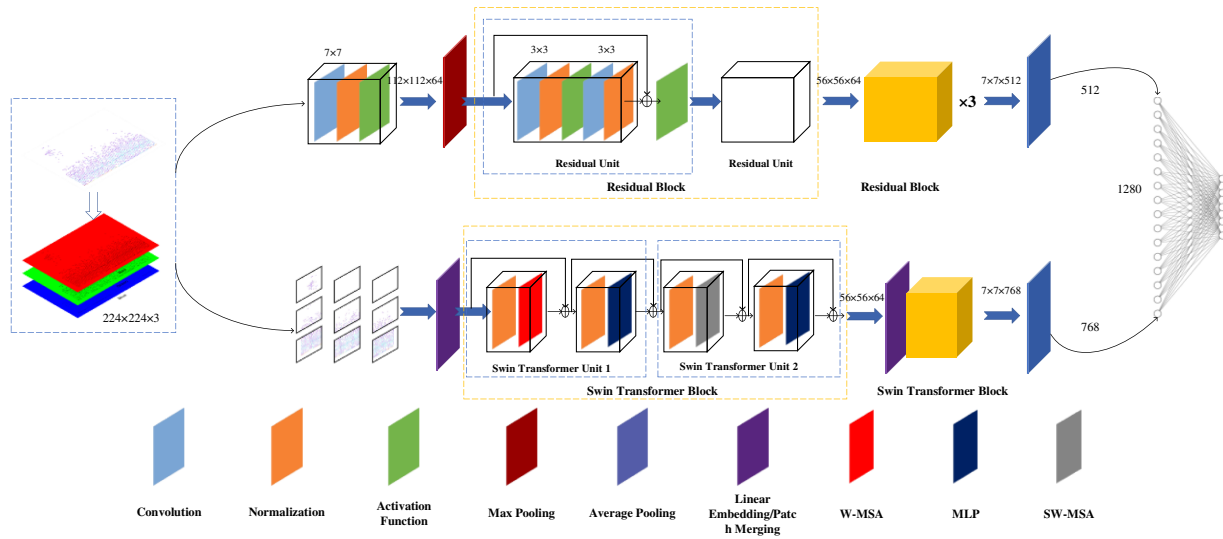


**Figure 4:** Architecture of the proposed Res-Swin fusion network.

#### 3.1 Dual-Stream Feature Extraction and Manifold Projection

The inherent complexity of PRPD maps requires a diagnostic model capable of decoupling fine-grained local discharge clusters from global phase-amplitude correlations. To achieve this, the proposed network processes the input PRPD images through two heterogeneous pathways, effectively constructing two distinct representation subspaces.

The convolutional branch employs ResNet18 [18] to establish a high-fidelity local morphological representation subspace. By utilizing a hierarchy of residual blocks, this branch acts as a specialized localized filter. It is particularly sensitive to the spatial distribution contours and edge textures of concentrated discharge clusters—features crucial for distinguishing morphologically similar defects like void and surface discharges. This pathway ultimately outputs a 512-dimensional local feature vector. In parallel, the Transformer branch utilizes the Swin Transformer [21] to map the global phase-amplitude topological modeling. By processing image patches through successive Window-based Multi-Head Self-Attention (W-MSA) and Shifted Window Multi-Head Self-Attention (SW-MSA) mechanisms, this branch transcends the limited receptive field of regular CNNs. It explicitly captures the long-range topological dependencies and periodic symmetric correlations across the entire 360-degree phase spectrum, yielding a 768-dimensional global feature vector.

To bridge the semantic gap between these two orthogonal subspaces, the local and global vectors are seamlessly concatenated to form a 1280-dimensional unified descriptor. Subsequently, a Multi-Layer Perceptron (MLP) acts as a non-linear cross-dimensional recalibration module. It optimally fuses the heterogeneous features and projects them into a highly discriminative latent manifold, establishing robust mathematical decision boundaries for the final classification.

### 3.2 Multi-Stage Freezing Strategy for Limited Data

In practical electrical engineering applications, acquiring massive, fully annotated PD datasets is extremely challenging. Both the ResNet18 and Swin Transformer branches are initialized with ImageNet pre-trained weights, and the entire training pipeline follows a transfer learning paradigm. Directly fine-tuning such structurally complex dual backbones on limited domain-specific data inherently risks catastrophic forgetting and severe overfitting [29]. To mitigate this, a multi-stage training freezing strategy is implemented. By precisely regulating parameter gradients across different phases, this strategy acts as a dynamic implicit regularizer:

- Stage I: Synchronous Warm-up (Epochs 0 to 25): In the initial phase, all parameters of the dual branches, the fusion module, and the classifier head are updated simultaneously. This phase facilitates the rapid preliminary adaptation of the pre-trained weights to the specific domain distribution of the PRPD dataset.
- Stage II: Feature Anchoring and Head Calibration (Epochs 26 to 75): To prevent overfitting, the parameters of both the ResNet and Swin backbone networks are strictly frozen. Gradient updates are restricted solely to the MLP and classifier head. This forces the network to concentrate computational resources on refining the classification decision boundaries, relying on the anchored, stable feature extractors.
- Stage III: Global Asymptotic Refinement (Epochs 76 to 100): Finally, all backbone parameters are unfrozen and jointly fine-tuned. To prevent the destruction of learned features, this stage is governed by a cosine learning rate decay strategy, enabling a gentle, global asymptotic refinement of the entire network architecture.

### 3.3 Branch-Wise Hyperparameter Optimization via ACLSCSO

A critical challenge in training the proposed hybrid architecture is the disparate optimization dynamics of its components. The pre-trained convolutional layers, the self-attention blocks, and the newly initialized classification head exhibit fundamentally distinct sensitivities to initial learning rates. Uniformly applying a single global learning rate significantly restrains the model's optimization potential.

Selecting an appropriate optimization algorithm is crucial for this heterogeneous architecture. Traditional metaheuristic approaches, such as Particle Swarm Optimization (PSO) or Grey Wolf Optimizer (GWO), frequently suffer from premature convergence and become trapped in local optima within the high-dimensional, non-convex loss landscapes typical of dual-branch networks. Conversely, while Bayesian optimization is highly sample-efficient, it incurs prohibitive computational overhead when scaling to complex continuous search spaces with multiple interdependent hyperparameters across different network branches. To overcome these bottlenecks, this study adopts the Sand Cat Swarm Optimization (SCSO) framework as a baseline due to its superior inherent structural balance between low-frequency acoustic-guided exploration and high-frequency attack-driven exploitation. Standard Sand Cat Swarm Optimization (SCSO) [30] tends to suffer from premature convergence in complex non-convex loss landscapes, mainly due to the uneven distribution of the initial population. In ACLSCSO, a Tent chaotic map is incorporated into the population initialization stage to alleviate this issue. Benefiting from the ergodic characteristic of the chaotic sequence, the initialized agents achieve a more uniform coverage of the hyperparameter search space. The corresponding mapping function is defined as:

$$T(x) = \begin{cases} x/0.7 & x < 0.7 \\ (1-x)/0.3 & \text{otherwise} \end{cases} \quad (1)$$

$$X_j = X_{\min} + T(x_j) \cdot (X_{\max} - X_{\min}) \quad (2)$$

here, the initial value  $x_0$  is drawn from a uniform distribution over the open interval  $(0, 1)$ , and the chaotic sequence  $\{x_j\}$  is subsequently generated through iterative application of the Tent map in Eq. (1), producing an ergodic and uniformly distributed set of candidate positions that covers the hyperparameter search space defined by  $X_{\min}$  and  $X_{\max}$ . Furthermore, to enhance the algorithm's capability to escape local optima during the exploration phase, a Lévy flight strategy is incorporated, executing occasional long-distance jumps to aggressively expand the search radius.

Crucially, to maintain a strict balance between global exploration and local exploitation, an adaptive adjustment mechanism is introduced for the switching probability ( $p_{switch}$ ). This probability is dynamically regulated based on the relative fitness of the current individual, formulated as:

$$p_{switch} = p_{base} + \alpha \cdot \frac{L_{avg} - L_{current}}{L_{avg}} \quad (3)$$

here,  $\alpha$  is the adjustment coefficient constrained to the range  $(0, 1)$ , which governs the sensitivity of the switching probability to the relative fitness deviation between the current individual and the population average. In this mechanism, if an agent's current loss ( $L_{current}$ ) is lower than the population's historical average ( $L_{avg}$ ), the algorithm autonomously favors the hunting mode to conduct precise local refinement. Conversely, if the loss is comparatively high, it increases the probability of exploration to prevent stagnation. This tailored, dimension-aware optimization effectively guarantees that all heterogeneous branches converge harmoniously, thereby maximizing the ultimate diagnostic accuracy.

Through these strategies, ACLSCSO achieves a robust balance between global search capabilities and local convergence precision, making it highly suitable for optimizing the multi-branch learning rates of the Res-Swin model. Finally, the schematic diagram of the improved Res-Swin model for partial discharge pattern recognition under multiple strategies is illustrated in Fig. 5. The ACLSCSO optimization is executed exclusively during the offline training phase to determine the optimal initial learning rate combination. Once training is complete, the algorithm introduces no additional computational burden to the inference pipeline, leaving the real-time diagnostic latency solely determined by the forward pass of the trained network.

## 4 Experimental Analysis

### 4.1 Parameter Settings and Evaluation Metrics

To explicitly address computational complexity and ensure reproducibility, all experiments were conducted within the PyTorch 2.5.1 framework using Python 3.12.9 on a 64-bit Windows 11 workstation. The hardware architecture comprises an Intel Core i7-12700K CPU and an NVIDIA GeForce RTX 3080 GPU accelerated by CUDA 12.4. During the model optimization phase, the training process—incorporating a batch size of 8, multi-stage freezing, and the ACLSCSO algorithm—requires a peak GPU memory of approximately 4500 MB. Despite the intensive optimization process required during training, the final deployed Res-Swin network remains highly efficient for practical condition monitoring. During the testing and inference phase, the average inference time to process a single PRPD map (resolution:  $224 \times 224$ ) is 9.32 ms. It should be noted that ACLSCSO is used only during offline model development to determine the branch-wise initial learning rates. Once the model is trained, online diagnosis requires only a single forward pass of the fused Res-Swin network. Therefore, the reported inference latency reflects the deployed diagnostic stage rather than the offline optimization stage. Although the current results indicate practical potential for PRPD-map-level online monitoring, further field deployment should also consider acquisition-system integration, hardware constraints, and long-term validation under real operating conditions.

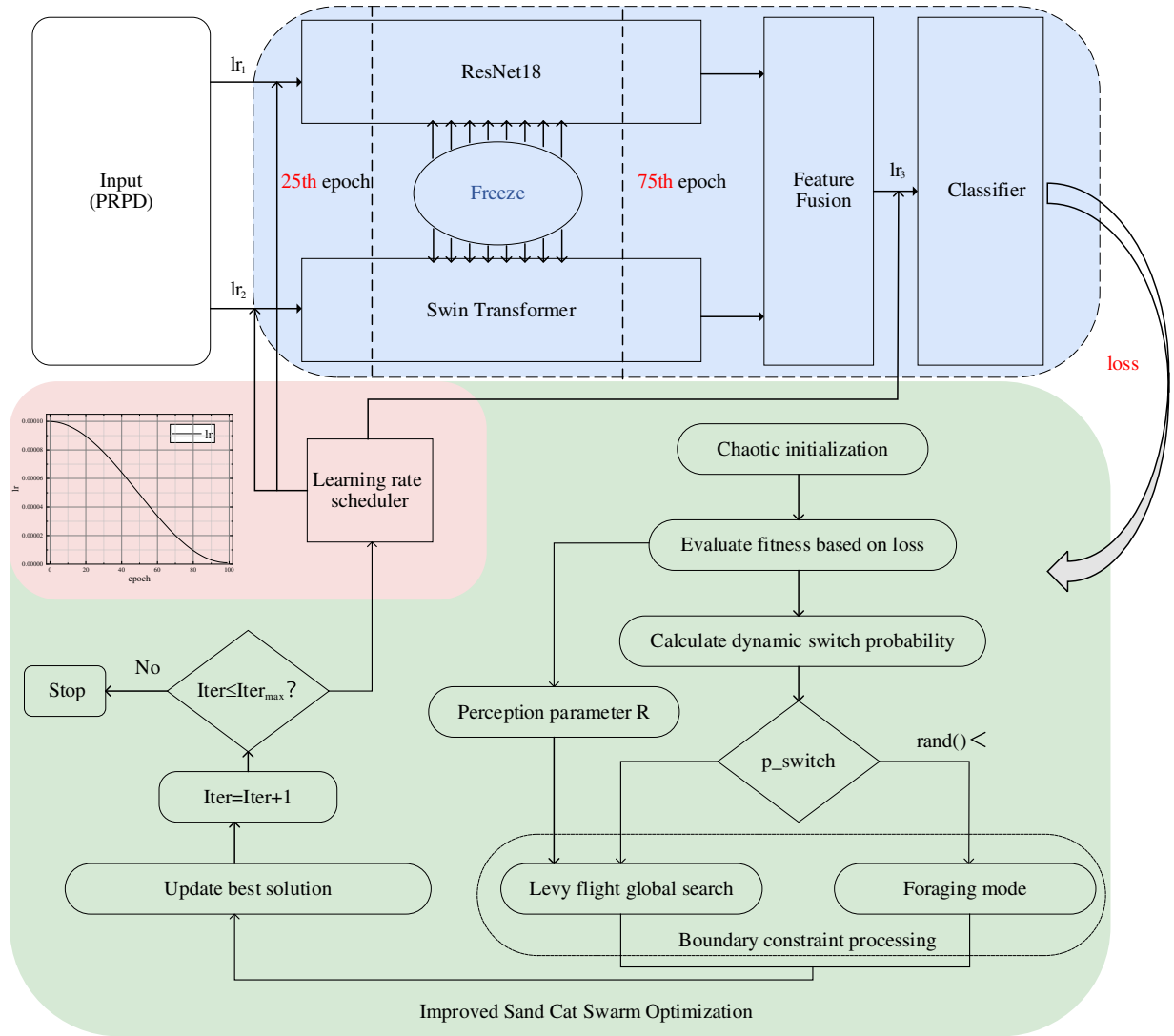


**Figure 5:** Architecture of the improved Res-Swin fusion model.

Based on extensive preliminary testing, the key hyperparameters for the proposed ACLSCSO algorithm and network training were explicitly defined. The swarm population size ( $N_{pop}$ ) and maximum iterations ( $Iter_{max}$ ) were set to 30 and 80, respectively. For the adaptive switching mechanism, the base probability ( $p_{base}$ ) and adjustment coefficient ( $\alpha$ ) were empirically set to 0.5 and 0.2. To ensure effective optimization, the search space boundaries for the initial learning rates of the ResNet branch, Swin branch, and classifier head were constrained to  $[10^{-5}, 10^{-2}]$ ,  $[10^{-6}, 10^{-3}]$ , and  $[10^{-5}, 10^{-2}]$ , respectively. These bounds were determined by two considerations. For the pre-trained backbone branches, transfer learning principles dictate that fine-tuning rates should be kept substantially smaller than those used for random initialization to preserve the learned feature representations; the Swin Transformer branch was assigned a narrower and lower range than ResNet18, reflecting its greater sensitivity to learning rate perturbations. For the randomly initialized classifier head, a relatively wider range extending to  $10^{-2}$  was prescribed to allow sufficient adaptability to the target domain distribution. Additionally, the network training utilized a batch size of 8 and a weight decay coefficient of  $5 \times 10^{-2}$  over 100 epochs.

Prior to being fed into the dual-branch network, all PRPD images are preprocessed to meet the standard  $224 \times 224$  input requirement of both backbones. For the training set, each PRPD image undergoes random resized cropping to  $224 \times 224$  followed by random horizontal flipping as training-time regularization. These operations are applied only during model training and are not used for validation or testing. Since the horizontal and vertical axes of PRPD maps represent phase and normalized amplitude, respectively, the augmented images are not interpreted as physically new discharge samples. Instead, they serve only to reduce overfitting by encouraging the network to learn robust phase-amplitude distribution patterns rather than memorizing fixed pixel-level positions. For the validation set, images are first resized to  $256 \times 256$  and then center-cropped to  $224 \times 224$  to provide a deterministic and physically consistent evaluation representation. All images are subsequently normalized using the ImageNet channel statistics (mean: [0.485, 0.456, 0.406], standard deviation: [0.229, 0.224, 0.225]).

Four standard evaluation metrics, including Accuracy, Precision, Recall, and F1-score, are adopted to quantitatively assess the performance of the proposed model in PRPD pattern recognition. Accuracy represents the overall proportion of correctly classified samples.

Precision and Recall characterize the prediction reliability and the model's sensitivity to target classes, respectively. In this work, per-class Recall is computed as  $TP/(TP + FN)$  for each individual category, and the Total Recall reported in the tables represents the macro-averaged Recall across all five classes, assigning equal weight to each category regardless of sample size. Furthermore, the F1-score, defined as the harmonic mean of Precision and Recall, provides a robust evaluation metric that is particularly suitable for datasets with class imbalance.

## 4.2 Experimental Results Analysis

### 4.2.1 Algorithm Optimization Results

Based on extensive experimental testing, the hyperparameters involved in the optimization process are provided in the [Appendix A \(Table A1\)](#). To ensure a fair and rigorous comparison, both the original SCSO and the proposed ACLSCSO were initialized with the identical set of populations. This was implemented by explicitly fixing the random seed to 42 prior to initialization, thereby eliminating the influence of random starting positions on the algorithmic performance. [Fig. 6](#) compares the convergence behaviors of the original SCSO and the proposed ACLSCSO algorithm in optimizing the learning rate combination for different branches of the fusion model. Benefiting from chaotic initialization, ACLSCSO exhibits a more uniform initial population distribution and achieves a lower initial objective value (0.84111) than SCSO (0.99466). Moreover, its convergence curve shows a faster decline during early iterations, indicating improved convergence efficiency. Ultimately, ACLSCSO reaches a significantly lower final objective value of 0.25972, compared with 0.33221 for SCSO, demonstrating improved search behavior and robustness against local optima. The optimal learning rates for the ResNet branch, Swin Transformer branch, and classifier head are determined as 0.000071, 0.00007, and 0.00071, respectively, validating that the pre-trained modules benefit from smaller fine-tuning rates, whereas the fusion head requires more aggressive learning. Overall, by integrating adaptive control, chaotic initialization, and Lévy flight, the ACLSCSO algorithm achieves an effective balance between exploration and exploitation, making it well suited for complex optimization tasks in partial discharge pattern recognition.

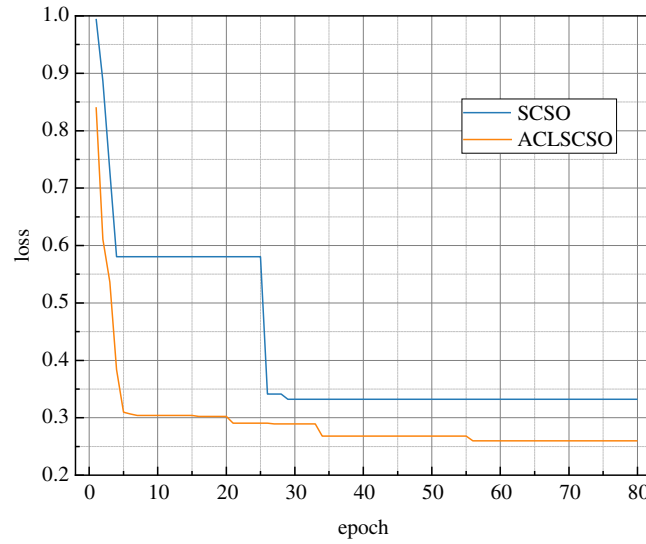


**Figure 6:** Optimization process of the three-branch learning rate loss function using the improved SCSO algorithm.

#### 4.2.2 Res-Swin Model Classification Effectiveness

This section systematically evaluates the classification efficacy of the proposed Res-Swin architecture, with the evolution of generalization accuracy illustrated in Fig. 7. During the initial training phase, single-branch models exhibit distinct limitations; notably, the ResNet model incurs the highest training loss due to its restricted receptive field. In contrast, by synergizing ResNet’s local texture extraction with Swin Transformer’s global context modeling, the fusion model effectively mitigates the “cold start” problem. It directly achieves a significantly lower initial loss of 1.0658 (compared to 1.1074 for the Swin baseline) alongside a remarkable initial accuracy of 82.62%. As training progresses, this synchronized advantage is consistently maintained. The underlying optimization loss descends smoothly to a minimal final value of 0.2165, strictly outperforming the Swin model’s 0.2599. This superior mathematical optimization directly translates into classification robustness, driving the validation accuracy to a peak of 96.10%. The strong coherence between the deeper loss convergence and the higher asymptotic accuracy confirms that the dual-branch fusion intrinsically learns a comprehensive and discriminative feature representation, ensuring highly reliable convergence. Asymptotic accuracy confirms that the dual-branch fusion does not merely memorize data, but intrinsically learns a more comprehensive and discriminative feature representation. Overall, the stable accuracy trajectory and the corresponding steep descent in training loss highlight the fusion strategy’s effectiveness in enhancing training stability and ensuring reliable convergence.

#### 4.2.3 Improved Res-Swin Model Classification Performance

To evaluate the generalization performance of the proposed improved Res-Swin model, stratified five-fold cross-validation was performed on the complete dataset of 1422 PRPD maps. The distribution summarized in Table 1 represents the full dataset composition rather than a fixed training/testing split. During cross-validation, the complete sample pool was divided into five stratified folds, and the class proportion of each fold was kept as consistent as possible with the overall dataset distribution. In each iteration, four folds were used for model training and the remaining fold was used for validation. This procedure was repeated five times until each fold had served once as the validation subset. The reported performance values were then averaged across the five validation folds. This approach provides a consistent record-level evaluation protocol for assessing model performance under the current dataset structure.

The cross-validation was conducted on the retained PRPD record pool with stratified class proportions. Each PRPD map was generated from an individual processed UHF record rather than from image-level cropping or duplication. Therefore, the reported results represent record-level stratified cross-validation performance on the current PRPD dataset.

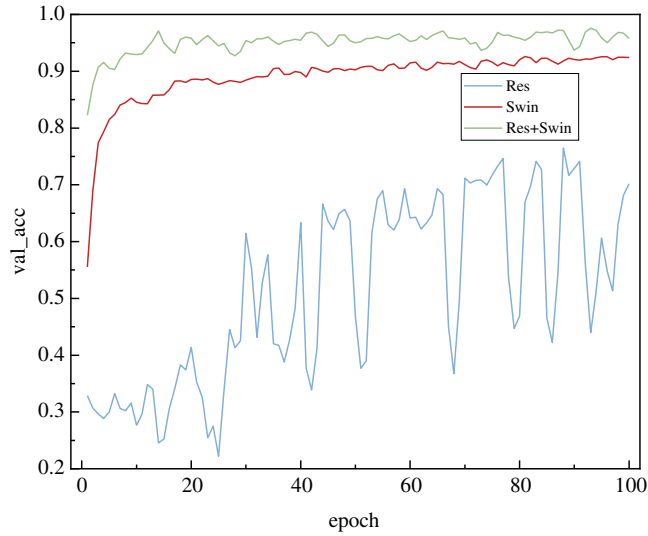


**Figure 7:** Accuracy curve of partial discharge classification.

The cross-validation results demonstrate the model's consistent stability; across the five folds, the improved Res-Swin model exhibits consistent convergence, achieving validation accuracies of 97.77%, 97.19%, 97.32%, 97.18%, and 98.24%, respectively. The resulting average accuracy of 97.54% confirms that the model's performance is robust to data variations and not dependent on specific data splits. Building on this solid foundation, systematic ablation studies were further conducted to quantify the contributions of each proposed strategy.

The synergistic effects of the proposed improvement strategies are further verified through analysis of the training dynamics by examining the accuracy evolution in Fig. 8 together with the computational efficiency metrics reported in Table 2. The baseline Res-Swin model (RS, green curve) shows gradual learning but suffers from noticeable fluctuations in the later stages, converging to a final accuracy of 0.9645. Introducing cosine learning rate decay (RSc, blue curve) significantly smooths the optimization landscape, accelerating the descent from local optima and improving the final accuracy to 0.9681 while mitigating oscillatory behavior. Independently, the application of the ACLSCSO-optimized Res-Swin model (RSA) yields the most dramatic improvement in the "cold start" phase. By precisely optimizing the initial hyper-parameters, the model achieves a notable initial accuracy of 86.17%. More importantly, this configuration simultaneously drives the initial validation loss down to 0.40378. This numerical advantage clearly indicates that ACLSCSO effectively positions the model in a highly favorable region of the parameter space right from the start, circumventing early-stage gradient struggles. Most critically, when Cosine Decay, Freezing, and ACLSCSO are jointly applied (RScfA, purple curve), the model exhibits the best overall performance, seamlessly combining rapid early-stage adaptability with robust long-term stability. Consequently, RScfA achieves the highest final accuracy of 0.9754. Notably, this best-performing configuration in this experiment also achieves a notably low final loss of 0.07576, suggesting that the joint strategy helps reduce prediction uncertainty under the current evaluation setting.

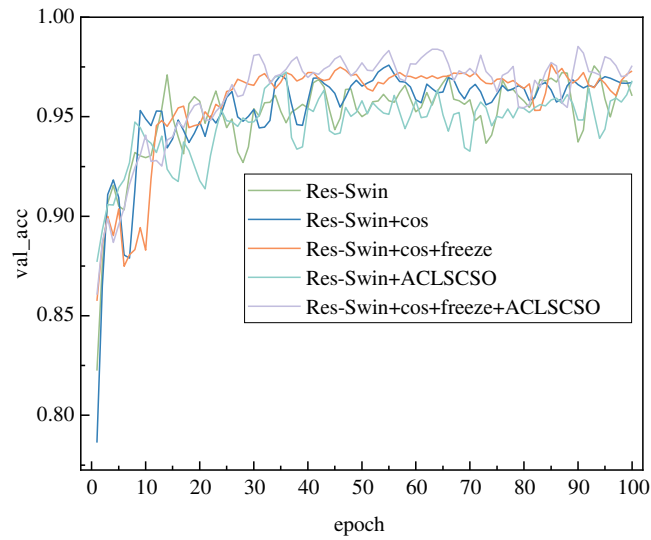


**Figure 8:** Accuracy curves of fusion model validation set under different configurations.

**Table 2:** Evaluation of recognition effectiveness of various graph types at 100 iterations (%).

| Model | Train Time (min) | Index     | Tip Discharge | Suspended Discharge | No PD  | Void Discharge | Surface Discharge | Total  |
|-------|------------------|-----------|---------------|---------------------|--------|----------------|-------------------|--------|
| RS    | 86               | Accuracy  | 0.931         | 1                   | 0.9167 | 0.9091         | 0.9863            | 0.9645 |
|       |                  | Recall    | 0.931         | 1                   | 0.9167 | 0.9091         | 0.9863            | 0.9486 |
|       |                  | Precision | 0.9818        | 1                   | 0.9565 | 0.7692         | 0.9474            | 0.931  |
|       |                  | F1-score  | 0.9558        | 1                   | 0.9362 | 0.8333         | 0.9664            | 0.9383 |
| Rsc   | 87               | Accuracy  | 0.9828        | 1                   | 0.9167 | 0.8182         | 0.9726            | 0.9681 |
|       |                  | Recall    | 0.9828        | 1                   | 0.9167 | 0.8182         | 0.9726            | 0.938  |
|       |                  | Precision | 0.95          | 1                   | 1      | 0.75           | 0.9595            | 0.9319 |
|       |                  | F1-score  | 0.9661        | 1                   | 0.9565 | 0.7826         | 0.966             | 0.9342 |
| RScf  | 77               | Accuracy  | 1             | 1                   | 0.9167 | 0.8182         | 0.9726            | 0.9716 |
|       |                  | Recall    | 1             | 1                   | 0.9167 | 0.8182         | 0.9726            | 0.9415 |
|       |                  | Precision | 0.9508        | 0.9892              | 0.9778 | 0.75           | 1                 | 0.9336 |
|       |                  | F1-score  | 0.9748        | 0.9946              | 0.9462 | 0.7826         | 0.9861            | 0.9369 |
| RSA   | 87               | Accuracy  | 1             | 1                   | 0.9167 | 0.9091         | 0.9452            | 0.9681 |
|       |                  | Recall    | 1             | 1                   | 0.9167 | 0.9091         | 0.9452            | 0.9542 |
|       |                  | Precision | 0.9667        | 1                   | 0.9565 | 0.6667         | 1                 | 0.918  |
|       |                  | F1-score  | 0.9831        | 1                   | 0.9362 | 0.7692         | 0.9718            | 0.9321 |
| RScfA | 78               | Accuracy  | 1             | 1                   | 0.9583 | 0.9091         | 0.9459            | 0.9754 |
|       |                  | Recall    | 1             | 1                   | 0.9583 | 0.9091         | 0.9459            | 0.9627 |
|       |                  | Precision | 0.9355        | 1                   | 1      | 0.8333         | 0.9859            | 0.9509 |
|       |                  | F1-score  | 0.9667        | 1                   | 0.9787 | 0.8696         | 0.9655            | 0.9561 |

Beyond accuracy, computational cost is a critical factor for practical engineering deployment. As detailed in Table 2, the multi-stage freezing strategy significantly enhances training efficiency. By selectively regulating parameter updates, the configuration with freezing (RScf) achieves an 11.5% reduction in training time compared to its unfrozen counterpart (RSc). Even for the fully optimized RScfA model—which

involves complex iterative swarm optimization steps—the total training time remains lower than the baseline RS model. This result suggests that the freezing mechanism reduces the computational overhead of backpropagation without noticeably compromising feature learning.

Table 2 and Fig. 9 present a comprehensive evaluation of recognition performance across specific PD categories. In Fig. 9, the horizontal axis corresponds to the predicted labels, while the vertical axis denotes the ground-truth categories, thereby explicitly visualizing inter-class confusion patterns. While the baseline RS model performs well on distinct patterns like Suspended Discharge (approaching 100% accuracy), it struggles with morphologically similar categories, specifically showing a low precision of 76.92% for Void Discharge. To explicitly evaluate the impact of the proposed hyperparameter optimization, a direct comparison is made between the unoptimized RScf model and the optimized RScfA framework. The RScf baseline employs only cosine decay and layer freezing, whereas the RScfA framework further incorporates the ACLSCSO algorithm. Although the unoptimized RScf model achieves a respectable total accuracy of 97.16%, it still struggles with overlapping phase-amplitude features, yielding an accuracy of only 81.82% for Void Discharge and 91.67% for the “No PD” state. By introducing the ACLSCSO optimization, the RScfA model dynamically identifies the optimal learning rate combinations for the dual-branch network. Consequently, the total accuracy increases to 97.54%. More notably, this optimization substantially resolves the most persistent inter-class confusion: the accuracy for the challenging Void Discharge class jumps from 81.82% to 90.91%, and the accuracy for “No PD” improves from 91.67% to 95.83%. In summary, the combined application of Cosine Decay, Freezing, and ACLSCSO does not merely provide incremental gains; it creates a robust synergy that simultaneously maximizes overall accuracy, minimizes training cost, and resolves critical misclassification bottlenecks in challenging discharge modes.

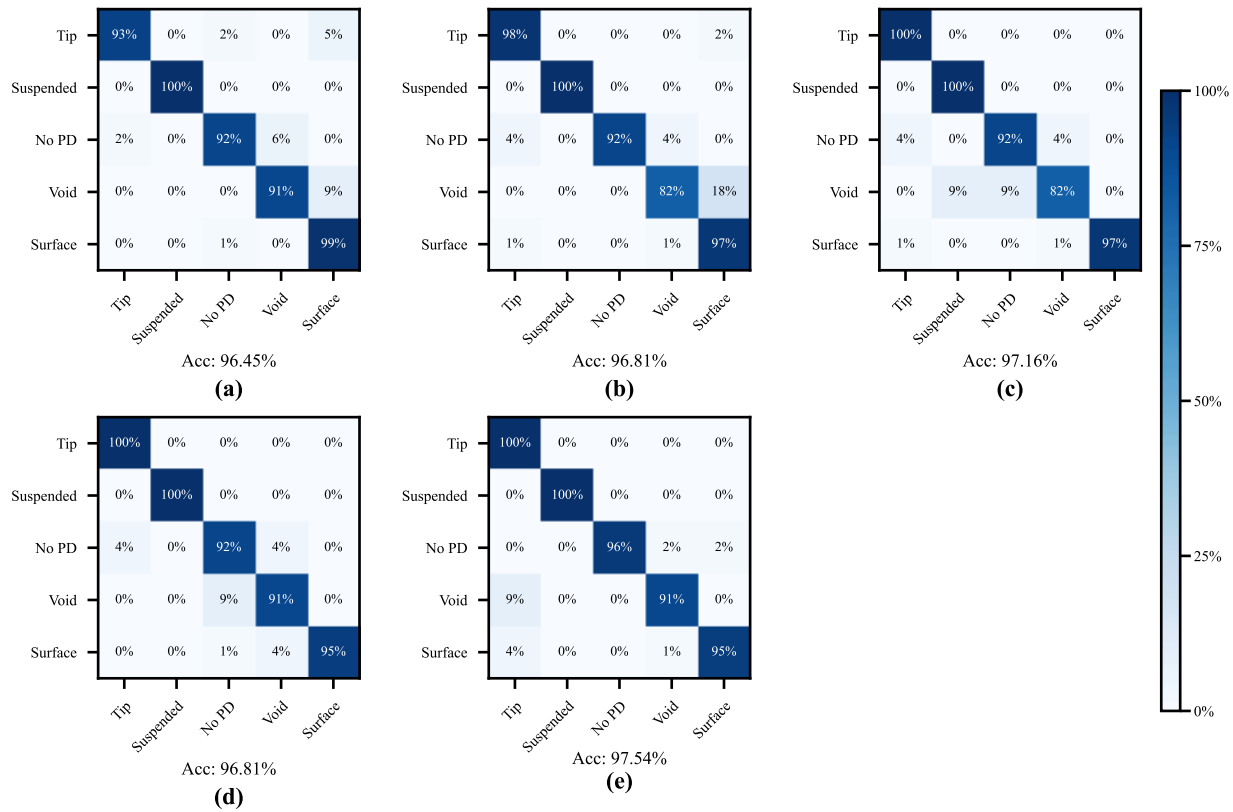


**Figure 9:** Confusion matrix of recognition accuracy for each graph type during 100 iterations. (a) RS (b) Rsc (c) RScf (d) RSA (e) RScfA.

To rigorously validate the reliability and statistical significance of the proposed improvements, a detailed statistical analysis across the five-fold cross-validation is presented in Table 3. Rather than relying solely on mean performance, assessing the Standard Deviation and the 95% Confidence Interval provides deeper insights into model stability under data variations. Furthermore, paired Student's *t*-tests were conducted to explicitly quantify the statistical significance of the performance gains achieved by the final optimized framework relative to all other ablation configurations.

**Table 3:** Statistical significance analysis of five-fold cross-validation accuracy (%).

| Model | Mean (%) | Standard Deviation | 95% Confidence Interval | <i>p</i> -Value |
|-------|----------|--------------------|-------------------------|-----------------|
| RS    | 96.45    | 0.36               | $96.45 \pm 0.31$        | $p = 0.004$     |
| RSc   | 96.81    | 0.34               | $96.81 \pm 0.30$        | $p = 0.003$     |
| RScf  | 97.16    | 0.29               | $97.16 \pm 0.25$        | $p = 0.026$     |
| RSA   | 96.81    | 0.46               | $96.81 \pm 0.40$        | $p = 0.004$     |
| RScfA | 97.54    | 0.46               | $97.54 \pm 0.40$        | —               |

As observed, the baseline RS model exhibits notable volatility with a 95% Confidence Interval of  $96.45 \pm 0.31\%$ . The sequential introduction of Cosine Decay and the multi-stage Freezing strategy progressively stabilizes the training dynamics. Among the evaluated ablation configurations, the RScfA framework achieves the highest mean accuracy of 97.54% while maintaining a tightly constrained 95% Confidence Interval of  $97.54 \pm 0.40\%$ . The paired *t*-tests yield exact *p*-values of 0.004, 0.003, 0.026, and 0.004 when comparing the optimized RScfA framework to the RS, RSc, RScf, and RSA models, respectively. These *p*-values are below the commonly used significance threshold of 0.05, indicating that the performance differences between RScfA and the compared ablation variants are statistically significant under the current five-fold cross-validation setting. Nevertheless, the paired *t*-test is interpreted as supplementary statistical evidence rather than absolute proof of model superiority. Together with the mean accuracy, standard deviation, confidence interval, and class-wise results, these findings suggest that the combined use of cosine decay, multi-stage freezing, and ACLSCSO contributes to more stable classification performance on the current PRPD dataset.

In addition to the overall five-fold statistical analysis, class-wise confidence intervals were further estimated to quantify the uncertainty associated with each insulation condition, particularly for the minority Void Discharge category. The confidence intervals were calculated using binomial proportions based on the class-wise sample counts and the five-fold average class-wise accuracy of the proposed RScfA framework. As shown in Table 4, the majority classes have relatively narrow confidence intervals because of their larger sample sizes. In contrast, the Void Discharge category achieves a five-fold class-wise accuracy of 90.91%, while its estimated 95% confidence interval is wider, ranging from 81.02% to 97.14%. This result indicates that the performance estimate for Void Discharge should be interpreted with appropriate statistical caution because only 58 PRPD maps are available for this category.

#### 4.2.4 Comparison with Other Architectures

To provide a clearer context within the literature, comparative evaluations examine the proposed RScfA framework against four representative alternative topologies. The benchmarked models encompass early convolutional networks such as VGG16 [17], dense-connected convolutional architectures represented by DenseNet121 [31], foundational pure self-attention mechanisms exemplified by the Vision Transformer (ViT)-B/16 [32], and recent advanced hybrid designs such as MaxViT-Tiny [33]. To ensure a fair comparison,

all baseline models were trained on the same PRPD dataset using the same stratified five-fold cross-validation protocol, input size, epoch number, validation preprocessing, and evaluation metrics. The detailed training configurations of the baseline architectures and the proposed RScfA framework are provided in the [Appendix A \(Table A2\)](#), and the classification metrics with training time overhead are reported in [Table 5](#).

**Table 4:** Estimated class-wise 95% confidence intervals for the proposed RScfA framework.

| Insulation Condition | Number of PRPD Maps | Accuracy | 95% Confidence Interval |
|----------------------|---------------------|----------|-------------------------|
| Tip Discharge        | 294                 | 1        | 0.9875–1                |
| Suspended Discharge  | 367                 | 1        | 0.9900–1                |
| No Partial Discharge | 240                 | 0.9583   | 0.9247–0.9798           |
| Void Discharge       | 58                  | 0.9091   | 0.8102–0.9714           |
| Surface Discharge    | 463                 | 0.9459   | 0.9213–0.9648           |

**Table 5:** Performance comparison of the proposed RScfA framework with other deep learning architectures.

| Model    | Time (min) | Index     | Tip Discharge | Suspended Discharge | No PD  | Void Discharge | Surface Discharge | Total  |
|----------|------------|-----------|---------------|---------------------|--------|----------------|-------------------|--------|
| Vgg      | 93         | Accuracy  | 0.9828        | 0.9891              | 0.9583 | 0.6364         | 0.9452            | 0.9574 |
|          |            | Recall    | 0.9828        | 0.9891              | 0.9583 | 0.6364         | 0.9452            | 0.9024 |
|          |            | Precision | 0.8906        | 1                   | 0.9583 | 0.7692         | 0.9583            | 0.9215 |
|          |            | F1-score  | 0.9344        | 0.9945              | 0.9583 | 0.7778         | 0.9517            | 0.9234 |
| Densenet | 85         | Accuracy  | 0.9828        | 0.9891              | 0.8958 | 0.8182         | 0.9459            | 0.9716 |
|          |            | Recall    | 0.9828        | 0.9891              | 0.8958 | 1              | 0.9595            | 0.9654 |
|          |            | Precision | 0.9508        | 0.9891              | 0.9773 | 0.7857         | 0.9459            | 0.9409 |
|          |            | F1-score  | 0.9748        | 0.9891              | 0.9348 | 0.8333         | 0.9730            | 0.9529 |
| Vit      | 110        | Accuracy  | 0.8276        | 0.837               | 0.9583 | 0.0909         | 0.8219            | 0.8227 |
|          |            | Recall    | 0.8276        | 0.837               | 0.9583 | 0.0909         | 0.8219            | 0.7071 |
|          |            | Precision | 0.6575        | 0.9167              | 0.902  | 0.8333         | 0.8219            | 0.8596 |
|          |            | F1-score  | 0.7328        | 0.8750              | 0.9293 | 0.1667         | 0.8219            | 0.7051 |
| Maxvit   | 105        | Accuracy  | 0.9655        | 1                   | 0.9583 | 0.8182         | 0.9233            | 0.9574 |
|          |            | Recall    | 0.9655        | 1                   | 0.9583 | 0.8182         | 0.9233            | 0.9331 |
|          |            | Precision | 0.9492        | 0.9892              | 1      | 0.75           | 0.9600            | 0.9409 |
|          |            | F1-score  | 0.9573        | 0.9946              | 0.9787 | 0.7692         | 0.9730            | 0.9421 |
| RScfA    | 78         | Accuracy  | 1             | 1                   | 0.9583 | 0.9091         | 0.9459            | 0.9754 |
|          |            | Recall    | 1             | 1                   | 0.9583 | 0.9091         | 0.9459            | 0.9627 |
|          |            | Precision | 0.9355        | 1                   | 1      | 0.8333         | 0.9859            | 0.9509 |
|          |            | F1-score  | 0.9667        | 1                   | 0.9787 | 0.8696         | 0.9655            | 0.9561 |

The time column in [Table 5](#) denotes the five-fold average final training time rather than the total hyperparameter-search cost. For the proposed RScfA framework, the reported 78 min corresponds to the final training stage after ACLSCSO determines the branch-wise initial learning rates. The ACLSCSO search required approximately 183.6 min as a one-time offline optimization step in our implementation. This additional cost is incurred only during model development and does not affect the deployed inference latency, which remains 9.32 ms per PRPD map.

The empirical findings demonstrate certain architectural limitations when standard off-the-shelf attention layers are directly deployed for partial discharge diagnostics. The pure vision transformer ViT-B/16 exhibits difficulties with the domain-specific spatial-phase topologies of Phase-Resolved Partial Discharge maps, especially under data-starved scenarios. Consequently, it yields a lower total classification accuracy of 82.27 percent and undergoes a significant performance reduction on the scarce Void Discharge category, which is reflected by an F1-score of 0.1667. Although DenseNet121 and MaxViT-Tiny achieve favorable generalization capability with total accuracies reaching 97.16 percent and 95.74 percent, respectively, they still exhibit minor constraints in maintaining an optimal balance between precision and recall across morphologically resembling discharge patterns.

Most significantly, the proposed RScfA framework achieves competitive comprehensive performance among the evaluated networks, showing balanced improvements across the evaluation metrics. By decoupling local convolutional feature extraction via the ResNet branch from long-range global dependency modeling within the Swin Transformer branch, the RScfA architecture helps alleviate the overfitting tendencies often observed in pure attention mechanisms when training data is restricted. The optimized RScfA model achieves a total accuracy of 97.54 percent and a macro F1-score of 95.61 percent. Furthermore, the proposed architecture exhibits reasonable computational efficiency during the optimization phase, requiring 78 min of training time on the target hardware. This performance compares favorably to the training overhead of alternatives such as ViT-B/16 and MaxViT-Tiny, which demand 110 and 105 min, respectively. These quantitative comparisons indicate that the physical-informed feature decoupling mechanism and the automated swarm optimization within the RScfA model provide a viable option for insulation condition monitoring.

### 4.3 Feature Visualization and Interpretability Analysis

Following recent Explainable Artificial Intelligence (XAI) studies in data-driven process monitoring and PRPD pattern recognition [34], Grad-CAM and t-SNE are used in this section as qualitative tools to examine the decision regions and feature separability of the trained model. In engineering diagnosis, predictive accuracy alone is insufficient for deployment because the model decision should also be examined against process knowledge and known physical mechanisms. Inspired by studies that combine data-driven prediction with global and local XAI analyses, such as Shapley Additive Explanations (SHAP) and Grad-CAM, this section interprets the visualized response regions with respect to phase-resolved discharge characteristics rather than treating them as purely image-level saliency maps.

#### 4.3.1 Visualization of High-Dimensional Feature Distributions

To assess the discriminative power of the representations learned by the improved Res-Swin fusion model, the t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm was used to project the 1280-dimensional feature vectors from the validation samples in cross-validation into a two-dimensional manifold. Fig. 10 illustrates the resulting feature distribution, where the five insulation conditions show generally separable cluster distributions with limited overlap.

The visualization reveals that the model achieves high intra-class compactness and inter-class separability. Specifically, the Suspended Discharge samples (yellow crosses) form the most isolated cluster, corroborating the distinctiveness of their phase-amplitude signatures. A critical observation lies in the clear separation between Void Discharge (blue pluses) and Surface Discharge (purple diamonds). Despite the potential morphological similarities between these two patterns in raw PRPD spectra, the fusion network successfully disentangles them in the latent feature space. This indicates that the synergistic combination

of ResNet's local texture extraction and Swin Transformer's global context modeling effectively captures the subtle nuances required to distinguish these challenging categories.

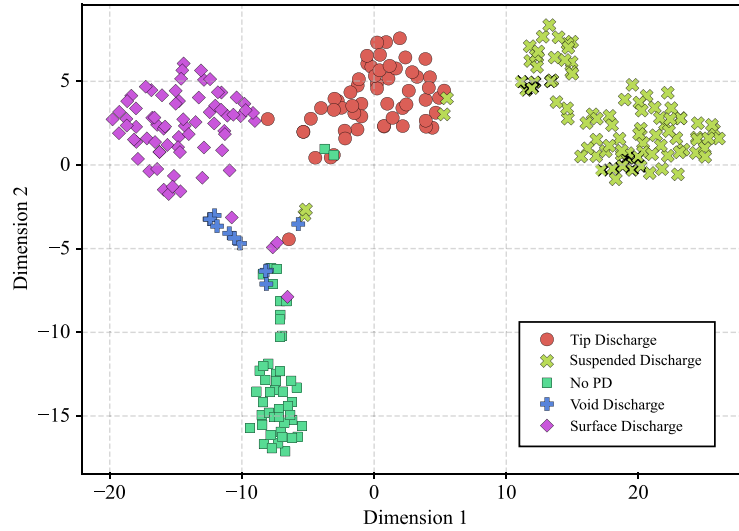


**Figure 10:** t-SNE visualization of feature vectors extracted by the improved Res-Swin model.

#### 4.3.2 Model Decision Logic Interpretation via Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) was introduced to provide qualitative explanatory evidence for the decision regions learned by the proposed fusion model. In this study, Grad-CAM was generated from the last residual module of the ResNet18 branch before feature aggregation. Therefore, the heatmaps mainly reflect the convolutional-branch contribution to the final fused prediction, rather than a direct visualization of the Swin Transformer branch or the final concatenated feature vector. The target class for each heatmap was set as the predicted class of the trained model, and all input images followed the same deterministic validation preprocessing pipeline, including resizing, center cropping, and ImageNet normalization.

A preliminary occlusion check was further conducted to examine whether the highlighted regions were relevant to the model prediction. For 282 correctly classified validation samples, the top 20% high-response Grad-CAM regions and the bottom 20% low-response regions were separately masked. The decrease in prediction confidence for the original predicted class was then recorded. Masking the high-response regions caused an average confidence decrease of 0.4240, whereas masking the low-response regions caused a smaller decrease of 0.1243. This result suggests that the highlighted regions were more influential for the model prediction. Nevertheless, Grad-CAM is interpreted here as qualitative explanatory evidence rather than definitive proof of physical causality.

As shown in Fig. 11, the attention distributions are broadly consistent with the characteristic structures of different PRPD patterns. For Tip Discharge, the model mainly focuses on high-amplitude triangular regions within specific phase intervals. For Suspended Discharge, the highlighted areas appear across both half-cycles, corresponding to the relatively symmetric discharge distribution. For the No PD category, the response is concentrated near the low-amplitude baseline region, which is consistent with the absence of strong discharge activity. For Surface Discharge and Void Discharge, the heatmaps show different response patterns: Surface Discharge presents broader phase-spanning regions, whereas Void Discharge exhibits more localized hotspots.

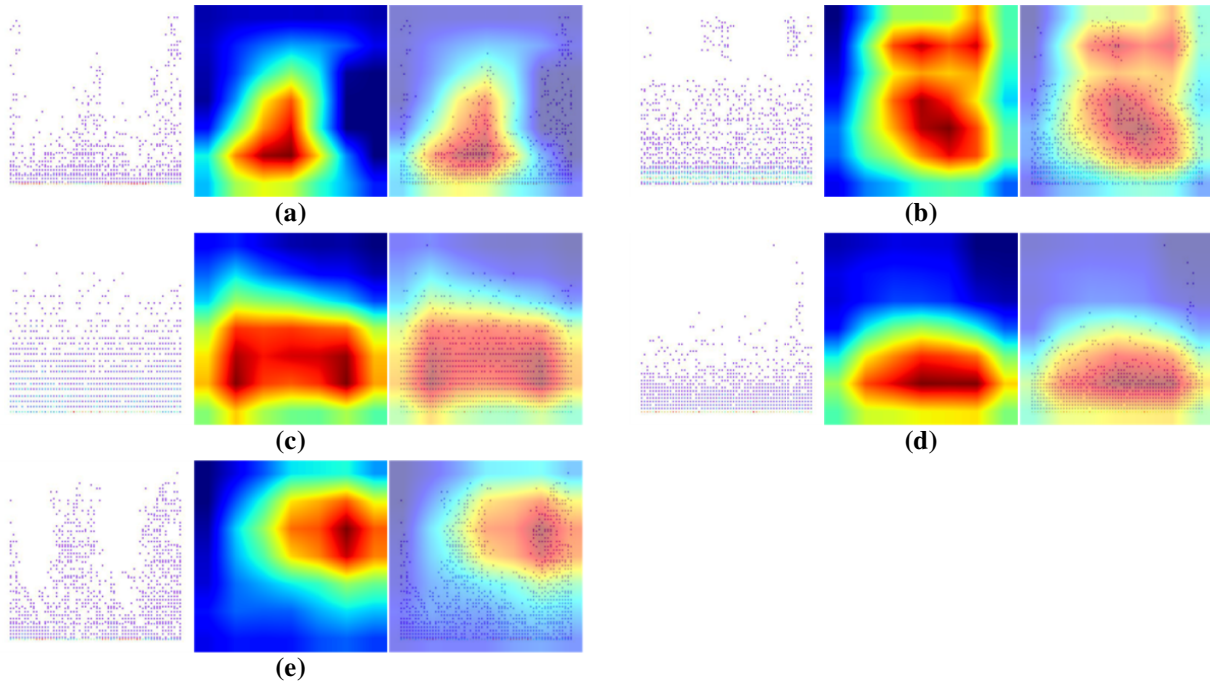


**Figure 11:** Grad-CAM visualization of decision logic for the improved Res-Swin fusion model. (a) Tip discharge (b) suspended discharge (c) no PD (d) void discharge (e) surface discharge.

Overall, the Grad-CAM results, together with the occlusion check, provide qualitative support that the proposed model relies on visually relevant PRPD regions for classification. These visual observations are consistent with the learned feature separability shown by t-SNE, but they should be regarded as supportive interpretability evidence rather than conclusive proof of the underlying discharge mechanisms.

## 5 Conclusion

This study presents an improved Res-Swin fusion framework for partial discharge pattern recognition under data-constrained conditions. By integrating convolutional local feature extraction with transformer-based global dependency modeling, the proposed architecture learns complementary PRPD representations from local discharge textures and global phase-amplitude distributions. Supported by staged parameter freezing and adaptive swarm-based optimization, the model achieves an average diagnostic accuracy of 97.54% under stratified five-fold cross-validation. Comparative benchmarking indicates that the proposed network provides competitive performance against representative deep architectures, while the staged freezing strategy helps alleviate overfitting risk in the limited-data setting. The proposed framework also improves the discrimination between morphologically similar patterns, reducing the misclassification rate of No PD as Void Discharge from 6.2% to 2.1%. These findings suggest that carefully structured hybrid architectures can improve PRPD-based insulation condition assessment by combining data-driven feature learning with the characteristic phase-amplitude structure of discharge patterns.

Nevertheless, the current physical dataset remains limited in scale and imbalanced across classes. In particular, the Void Discharge category contains fewer PRPD samples than the other classes, and its class-wise performance should therefore be interpreted with appropriate caution. Although training-time augmentation, ImageNet-based transfer learning, multi-stage freezing, and stratified five-fold cross-validation were adopted to reduce overfitting risk, the reported results should be regarded as record-level validation on the

present physical PRPD dataset rather than exhaustive evidence of universal generalization. Future work will focus on validating the proposed framework under authentic field conditions, including real cable joints, terminations, and transformer insulation with multi-source interference. In addition, physically constrained data augmentation and advanced generative frameworks, such as Denoising Diffusion Probabilistic Models and Conditional Flow Matching, will be explored to expand minority-class PRPD samples and further improve diagnostic robustness.

**Acknowledgement:** Not applicable.

**Funding Statement:** This work was supported by the National Natural Science Foundation of China (62173284), the Key Industrial Guidance Projects of Fujian Province (2024H0003) and the Natural Science Foundation of Guangdong Province (2024A151010544).

**Author Contributions:** The authors confirm contribution to the paper as follows: conceptualization, Jingrui Zhang; methodology, Zihao Jia and Jingrui Zhang; software, Zihao Jia; validation, Zihao Jia and Jingrui Zhang; formal analysis, Zihao Jia; investigation, Zihao Jia; resources, Jingrui Zhang; data curation, Zihao Jia; writing—original draft preparation, Zihao Jia; writing—review and editing, Jingrui Zhang; visualization, Zihao Jia; supervision, Jingrui Zhang; project administration, Jingrui Zhang; funding acquisition, Jingrui Zhang. All authors reviewed and approved the final version of the manuscript.

**Availability of Data and Materials:** The data that support the findings of this study are available from the corresponding author, Jingrui Zhang, upon reasonable request.

**Ethics Approval:** Not applicable.

**Conflicts of Interest:** Given his role as an Editorial Board Member of this journal, Jingrui Zhang had no involvement in the peer review of this article and had no access to information regarding its peer review. Full responsibility for the editorial process for this article was delegated to another journal editor. The authors declare no conflicts of interest.

## Appendix A

**Table A1:** Symbol description.

| Abbreviations/Symbols              | Description                                                                                                               |
|------------------------------------|---------------------------------------------------------------------------------------------------------------------------|
| $\eta_t, \eta_{\max}, \eta_{\min}$ | Current, maximum, and minimum learning rates                                                                              |
| $search\_space_j^{\max/\min}$      | Upper/Lower bounds of search space (Res: $[10^{-5}, 10^{-2}]$ , Swin: $[10^{-6}, 10^{-3}]$ , Head: $[10^{-5}, 10^{-2}]$ ) |
| $X_{\max/\min}$                    |                                                                                                                           |
| $T_{cur}, T_{\max}$                | Current and maximum training epochs (Set to 100)                                                                          |
| $N_{batch}, N_{class}$             | Batch size (Set to 8) and Number of PD categories (Set to 5)                                                              |
| $X_b, X_c, X_{rnd}$                | Best, current, and random position of the sand cat agent                                                                  |
| $\vec{r}, \theta$                  | Sensitivity range vector and random angle for position update                                                             |
| $R, r_G, s_M$                      | Mode control parameter ( $R \in [0.1, 2.0]$ ), sensitivity range variable, and max sensitivity                            |
| $iter_c, iter_{\max}$              | Current and maximum iterations for optimization (Max = 80)                                                                |
| $T(x), x_j$                        | Tent chaotic map function and chaotic sequence value                                                                      |
| $X_i^{new}, X_i, step$             | New position at $t + 1$ , position at $t$ , and Lévy flight step size                                                     |
| $p_{switch}, p_{base}$             | Adaptive switching probability and base probability (Set to 0.5)                                                          |
| $\alpha$                           | Adjustment coefficient for mode switching (Set to 0.2)                                                                    |

(Continued)

**Table A1 (continued)**

| Abbreviations/Symbols  | Description                                            |
|------------------------|--------------------------------------------------------|
| $L_{avg}, L_{current}$ | Average loss of preceding generations and current loss |
| $N_{pop}$              | Population size of the swarm (Set to 30)               |
| $w_{decay}$            | Weight decay coefficient (Set to $5 \times 10^{-2}$ )  |

**Table A2:** Training configurations for the proposed RScfA framework and baseline architectures.

| Model       | Pretrained Weights                                  | Input Size       | Optimizer | Initial Learning Rate                           | Epochs | Weight Decay       | LR Scheduler | Preprocessing/ Augmentation                                                                                                                  | Special Strategy                                                                     |
|-------------|-----------------------------------------------------|------------------|-----------|-------------------------------------------------|--------|--------------------|--------------|----------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------|
| VGG16       | ImageNet                                            | $224 \times 224$ | AdamW     | $1 \times 10^{-4}$                              | 100    | $1 \times 10^{-4}$ | Cosine decay | Training-time random resized cropping and horizontal flipping; deterministic validation resizing and center cropping; ImageNet normalization | Standard fine-tuning                                                                 |
| DenseNet121 | ImageNet                                            | $224 \times 224$ | AdamW     | $1 \times 10^{-4}$                              | 100    | $1 \times 10^{-4}$ | Cosine decay | Same as VGG16                                                                                                                                | Standard fine-tuning                                                                 |
| ViT-B/16    | ImageNet                                            | $224 \times 224$ | AdamW     | $5 \times 10^{-5}$                              | 100    | $1 \times 10^{-4}$ | Cosine decay | Same as VGG16                                                                                                                                | Standard fine-tuning                                                                 |
| MaxViT-Tiny | ImageNet                                            | $224 \times 224$ | AdamW     | $5 \times 10^{-5}$                              | 100    | $1 \times 10^{-4}$ | Cosine decay | Same as VGG16                                                                                                                                | Standard fine-tuning                                                                 |
| RScfA       | ImageNet for ResNet18 and Swin Transformer branches | $224 \times 224$ | AdamW     | Branch-wise learning rates optimized by ACLSCSO | 100    | $1 \times 10^{-4}$ | Cosine decay | Same as VGG16                                                                                                                                | Multi-stage freezing + cosine decay + ACLSCSO branch-wise learning-rate optimization |

## References

- Li S, Cao B, Li J, Cui Y, Kang Y, Wu G. Review of condition monitoring and defect inspection methods for composited cable terminals. *High Volt.* 2023;8(3):431–44. doi:10.1049/hve2.12318.
- CIGRE Working Group D1.11. Knowledge rules for partial discharge diagnosis in service. Paris, France: CIGRE; 2003.
- Annuar Isa M, Nur Khairul Hafizi Rohani M, Ismail B, Kamarol Jamil M, Isa M, Shuhada Rosmi A, et al. Electrical tree image segmentation using hybrid multi scale line tracking algorithm. *Comput Mater Contin.* 2023;75(1):741–60. doi:10.32604/cmc.2023.036077.
- Prabhu DK, Maheswari RV, Vigneshwaran B. Deep learning based power transformer monitoring using partial discharge patterns. *Intell Autom Soft Comput.* 2022;34(3):1441–54. doi:10.32604/iasc.2022.024128.
- Choudhary M, Shafiq M, Kiitam I, Palu I, Hassan W, Singh PP. Investigation of partial discharge characteristics in XLPE cable insulation under increasing electrical stress. *Eng Fail Anal.* 2024;158(3):108006. doi:10.1016/j.engfailanal.2024.108006.
- Liu K, Nie G, Fan L, Yi R, Gao B, Fu J, et al. Recognition of partial discharge in high-speed train cable terminals based on few-shot learning: from PD signals to PD time graph. *IEEE Trans Instrum Meas.* 2025;74:2544515. doi:10.1109/TIM.2025.3600727.

7. Rodrigo Mor A, Castro Heredia LC, Harmsen DA, Muñoz FA. A new design of a test platform for testing multiple partial discharge sources. *Int J Electr Power Energy Syst.* 2018;94(6):374–84. doi:10.1016/j.ijepes.2017.07.013.
8. Ji J, Shu Z, Li H, Lai KX, Lu M, Jiang G, et al. Edge-computing-based knowledge distillation and multitask learning for partial discharge recognition. *IEEE Trans Instrum Meas.* 2024;73:5008011. doi:10.1109/TIM.2024.3351239.
9. Li Z, Zhou K, Xu X, Meng P, Fu Y, Zeng Z. A thermal excitation based partial discharge detection method for cable accessory. *IEEE Trans Power Deliv.* 2023;38(4):2703–10. doi:10.1109/TPWRD.2023.3254907.
10. Berkemeier R, Bach R, Tenbohlen S. UHF partial discharge detection and localization: on-site experiences at 110-kV-cable terminations. *IEEE Trans Power Deliv.* 2025;40(2):1169–79. doi:10.1109/TPWRD.2025.3542846.
11. Sozzi M, Chiarello S, Pirovano G, Borghetto J, Contin A. A new integrated instrument for partial discharge monitoring and diagnostics of medium voltage cables. In: *Proceedings of the 2024 AEIT International Annual Conference (AEIT)*; 2024 Sep 25–27; Trento, Italy.
12. Liu Z, Zhu H. RC2DNet: real-time cable defect detection network based on small object feature extraction. *Comput Mater Contin.* 2025;85(1):681–94. doi:10.32604/cmc.2025.064191.
13. Seitz S, Götz T, Lindenberg C, Tetzlaff R, Schlegel S. Towards generalizable classification of partial discharges in gas-insulated HVDC systems using neural networks: protrusions and particles. *IEEE Trans Power Deliv.* 2024;39(3):1491–9. doi:10.1109/TPWRD.2024.3369872.
14. Lu S, Chai H, Sahoo A, Phung BT. Condition monitoring based on partial discharge diagnostics using machine learning methods: a comprehensive state-of-the-art review. *IEEE Trans Dielectr Electr Insul.* 2020;27(6):1861–88. doi:10.1109/TDEI.2020.009070.
15. Sahoo R, Karmakar S. Comparative analysis of machine learning and deep learning techniques on classification of artificially created partial discharge signal. *Measurement.* 2024;235(4):114947. doi:10.1016/j.measurement.2024.114947.
16. Peng X, Yang F, Wang G, Wu Y, Li L, Li Z, et al. A convolutional neural network-based deep learning methodology for recognition of partial discharge patterns from high-voltage cables. *IEEE Trans Power Deliv.* 2019;34(4):1460–9. doi:10.1109/TPWRD.2019.2906086.
17. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556.* 2014. doi:10.48550/arxiv.1409.1556.
18. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*; 2016 Jun 27–30; Las Vegas, NV, USA. doi:10.1109/CVPR.2016.90.
19. Zhou X, Wu X, Ding P, Li X, He N, Zhang G, et al. Research on transformer partial discharge UHF pattern recognition based on CNN-LSTM. *Energies.* 2020;13(1):61. doi:10.3390/en13010061.
20. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *Adv Neural Inf Process Syst.* 2017;30:5998–6008. doi:10.65215/2q58a426.
21. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*; 2021 Oct 10–17; Montreal, QC, Canada. doi:10.1109/iccv48922.2021.00986.
22. Deng Y, Zhu K, Liu J, Liu H. MRATNet: learning discriminative features for partial discharge pattern recognition via transformers. *IEEE Trans Dielectr Electr Insul.* 2024;31(4):2198–207. doi:10.1109/TDEI.2024.3399171.
23. Zheng S, Liu J, Zeng J. MDTCNet: a novel multiscale denoising transformer convolutional network for fault diagnosis of partial discharge. *IEEE Trans Dielectr Electr Insul.* 2025;32(5):2938–47. doi:10.1109/TDEI.2025.3533472.
24. Li J, Ma S, Jin F, Zhao R, Zhang Q, Xie J. A study on partial discharge fault identification in GIS based on swin transformer-AFPN-LSTM architecture. *Information.* 2025;16(2):110. doi:10.3390/info16020110.
25. Rauscher A, Kaiser J, Devaraju M, Endisch C. Deep learning and data augmentation for partial discharge detection in electrical machines. *Eng Appl Artif Intell.* 2024;133(1):108074. doi:10.1016/j.engappai.2024.108074.
26. Zhang C, Fulneček J, Yang L, Zhang Y, Zheng J. Combining multi-level feature extraction algorithm with residual graph convolutional neural network for partial discharge detection. *Measurement.* 2025;242:116151. doi:10.1016/j.measurement.2024.116151.

27. Zhang Y, Xie Q, Xie J, Wang C, Zheng Y, Xie C. An automatic annotation and pattern recognition method based on semisupervised ACGAN for multisource partial discharge diagnosis. *IEEE Trans Dielectr Electr Insul.* 2024;31(6):3001–10. doi:10.1109/TDEI.2024.3403798.
28. Li Y, Gong C, Deng T, Jia Z, Wang F, Zhao Q, et al. Partial discharge pattern recognition based on swin transformer for power cable fault diagnosis in modern distribution systems. *Processes.* 2025;13(3):852. doi:10.3390/pr13030852.
29. Kirkpatrick J, Pascanu R, Rabinowitz N, Veness J, Desjardins G, Rusu AA, et al. Overcoming catastrophic forgetting in neural networks. *Proc Natl Acad Sci U S A.* 2017;114(13):3521–6. doi:10.1073/pnas.1611835114.
30. Seyyedabbasi A, Kiani F. Sand Cat swarm optimization: a nature-inspired algorithm to solve global optimization problems. *Eng Comput.* 2023;39(4):2627–51. doi:10.1007/s00366-022-01604-x.
31. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; 2017 Jul 21–26; Honolulu, HI, USA. doi:10.1109/cvpr.2017.243.
32. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. *arXiv:2010.11929.* 2020.
33. Tu Z, Talebi H, Zhang H, Yang F, Milanfar P, Bovik A, et al. MaxViT: multi-axis vision transformer. In: *Proceedings of the Computer Vision–ECCV 2022*; 2022 Oct 23–27; Tel Aviv, Israel. doi:10.1007/978-3-031-20053-3\_27.
34. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int J Comput Vis.* 2020;128(2):336–59. doi:10.1007/s11263-019-01228-7.