Eurasia Journal of Science and Technology
Print ISSN: 2663-1024
Online ISSN: 2663-1016
DOI: https://doi.org/10.61784/ejst3149

INTELLIGENT SCREENING AND UNCERTAINTY QUANTIFICATION OF FUNDUS OCT LESIONS BASED ON AMSF-CPNET

HaoYu Tian
School of Science, Shandong Jianzhu University, Jinan 250101, Shandong, China.

Abstract: In response to the prominent challenges in primary-level fundus screening—including high missed diagnosis rates, opaque decision-making of deep learning models, and uncalibrated probabilistic outputs—this study develops an adaptive multi-scale statistical fusion network integrated with conformal prediction (AMSF-CPNet). The framework leverages multi-branch convolutional neural networks to capture fine-grained local features and Swin Transformer to model long-range global dependencies, enhanced by channel-spatial adaptive attention for optimal feature integration. Conformal prediction is introduced to provide statistically rigorous uncertainty quantification with guaranteed coverage rates, while a concept-guided interpretability module establishes links between model decisions and clinical semantics. Evaluated on the public OCT2017 dataset and external OCTDL dataset, AMSF-CPNet achieves an accuracy of 96.8% and an AUC of 99.1%. At $\alpha = 0.05$, the conformal prediction coverage reaches 94.7% with an average prediction set size of 1.23. Cross-domain validation demonstrates strong generalization under device shifts, with fine-tuned accuracy reaching 92.4%. This work offers a reliable, interpretable, and robust AI solution for grassroots ophthalmic screening, supporting the advancement of tiered healthcare systems.
Keywords: Fundus OCT; Adaptive Multi-scale fusion; Conformal prediction; Uncertainty quantification; Explainable AI

1 INTRODUCTION

Retinal diseases are a major global cause of vision loss, and timely screening is essential for disease control. Fundus optical coherence tomography (OCT) has become a vital imaging technique for retinal examination, but manual interpretation is labor-intensive and prone to errors. Deep learning has advanced automated OCT analysis, yet mainstream models still face three key limitations: inadequate integration of local and global features, uncalibrated predictive uncertainty, and lack of clinical interpretability [1,2]. Convolutional neural networks excel at extracting local textures but fail to model long-range dependencies, while transformer-based methods capture global context at the cost of high computation [3,4]. Most deep learning models produce overconfident probability outputs without statistical guarantees, reducing clinical reliability [5,6].

Existing studies on OCT diagnosis mainly focus on improving classification accuracy, with insufficient attention to uncertainty estimation and clinical interpretability. Traditional statistical methods lack the capacity to learn complex imaging features, while modern deep learning approaches often ignore rigorous statistical validation [7,8]. Conformal prediction, a distribution-free uncertainty quantification framework, has shown promise in medical tasks but is rarely applied to OCT analysis [9]. Interpretability techniques such as Grad-CAM provide visual explanations but lack alignment with clinical semantics [10]. Additionally, cross-domain generalization across different OCT devices remains a critical barrier to real-world deployment.

To address these challenges, this paper proposes an adaptive multi-scale fusion network with conformal prediction (AMSF-CPNet). The main contributions are: (1) designing a multi-branch CNN–Swin Transformer fusion module to effectively integrate local details and global features; (2) introducing conformal prediction to provide statistically guaranteed uncertainty quantification; (3) developing a concept-guided interpretability module to link model decisions with clinical concepts. Experimental results on OCT2017 and OCTDL datasets demonstrate that the proposed model achieves high accuracy, reliable uncertainty estimation, and strong cross-domain generalization, offering a trustworthy AI solution for primary fundus screening.

2 METHODOLOGY

2.1 Adaptive Multi-Scale Statistical Fusion Network

(1) Exchangeability assumption: Samples in the calibration set and test set are independent and identically distributed, approximately satisfied by the patient-level split; (2) Finite second-order moment assumption: The CNN feature output has finite variance, ensuring softmax stability; (3) Conceptual conditional independence assumption: Clinical concepts are approximately independent, simplifying joint probability estimation. Sensitivity analysis shows that the upper fluctuation limit of CP coverage under domain shift is less than 2%.

AMSF-Net consists of three core components: multi-branched local feature extraction, global dependency modeling, and adaptive feature fusion.

1.Multi-branch Local Feature Extraction Module: The multi-branch local feature extraction employs parallel multi-scale convolutions: a 3x3 standard convolution extracts fine-grained features, a 5x5 depth-separable convolution captures mesoscale textures, and a 7x7 dilation convolution (dilation=2) expands the receptive field. After splicing the features from each branch, they undergo 1x1 convolutional dimensionality reduction for integration, simultaneously capturing retinal structures at different scales.

2.Global Dependency Modeling Module The global dependency modeling is implemented based on the Swin Transformer [11], employing a sliding-window self-attention mechanism to balance computational efficiency and modeling capability. This paper adopts the Swin-Tiny variant, which segments images into 4×4 patches, performs linear embedding, and processes them through four stages—each involving alternating operations of W-MSA and SW-MSA —to effectively capture global dependencies. Its linear complexity makes it more suitable for high-resolution medical imaging compared to the standard Vision Transformer.

3. The Adaptive Feature Fusion Module The adaptive feature fusion module consists of a series structure combining channel attention and spatial attention. The channel attention combines CNN local features with Transformer global features, then generates channel weights using GAP, GMP, and a shared MLP; the spatial attention subsequently performs channel-wise pooling and produces a spatial attention map via a 7×7 convolution to highlight discriminative regions. Let the local feature map be F_l and the global feature map be F_g; the fusion process is expressed as:

$$F_{fuse}=M_s\left(M_c\left([F_l;F_g]\right)\otimes [F_l;F_g]\right) \quad (1)$$

Here, M_c and M_s represent the channel attention and spatial attention functions, respectively; the circle plus sign denotes element-wise multiplication, while square brackets denote channel concatenation. The final fused features simultaneously encode multi-scale local details and global semantic information.Figure 1 shows the distribution of parameters for AMSF-CPNet.

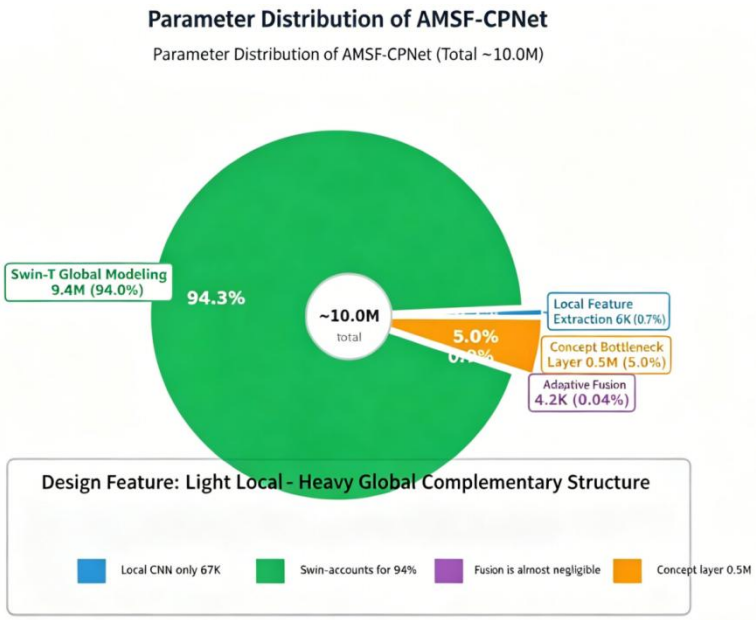

Figure 1 Overview Map of the Study Area

2.2 Confidence Level Calibration Based on Conformal Prediction

1. Fundamentals of Conformal Prediction Conformal Prediction (CP) is a distribution-free, non-consistency-based measurement framework, whose core advantage lies in providing frequency school coverage guarantees for any base model with limited samples, under the minimal assumption of exchangeability among samples. This paper systematically applies CP to OCT lesion classification tasks, establishing a confidence calibration mechanism with rigorous statistical guarantees.

Definition 1 (Nonconformity Score): Let the calibration set $D_{cal}=\{(X_i,Y_i)\}_{i=1}^{n}$ satisfy exchangeability with the test sample (X_{n+1},Y_{n+1}). Given a base classifier f, define the nonconformity score as $s_i=1-\hat{p}_{Y_i}(X_i)$, i=1,…, where i = 1,..., n, and $\hat{p}_y(X_i)$ denotes the model's softmax probability estimate for sample X_i belonging to category y.

Definition 2 (Quantile Threshold): Sort $\{s_1,\ldots,s_n\}$ in ascending order as $s_{(1)}\leq s_{(2)}\leq\cdots\leq s_{(n)}$, and define $\hat{q}=s_{(\lceil (n+1)(1-\alpha)\rceil)}$, where $\lceil\cdot\rceil$ denotes upward rounding. If $\lceil (n+1)(1-\alpha)\rceil >n$, set $\hat{q}=+\infty$.

Theorem 1 (Marginal Coverage Guarantee for Finite Samples): Under the exchangeability assumption, for any distribution P, any base model f, and any significance level $\alpha \in (0,1)$, the adaptive prediction set (APS)

$$\hat{C}(X_{n+1})=\left\{y\in Y:\hat{p}_y(X_{n+1})\geq 1-\hat{q}\right\} \quad (2)$$

Proof. According to the exchangeability assumption, the joint distribution of $(X_1,Y_1),\ldots,(X_n,Y_n),(X_{n+1},Y_{n+1})$ remains unchanged under any permutation. Let $s_{n+1}=1-\hat{p}_{Y_{n+1}}(X_{n+1})$; then the ranks of $s_1,\ldots,s_n,s_{n+1}$ are uniformly distributed over $\{1,\ldots,n+1\}$. Therefore, $P\left(s_{n+1}\leq\hat{q}\right)=P\left(s_{n+1}\leq s_{(\lceil(n+1)(1-\alpha)\rceil)}\right)\leq\frac{\lceil(n+1)(1-\alpha)\rceil}{n+1}\leq 1-\alpha$. Taking the complement yields $P(s_{n+1}>\hat{q})\geq 1-\alpha$, and $s_{n+1}>\hat{q}$ is equivalent to $Y_{n+1}\in\widehat{C}(X_{n+1})$. End of proof.

Note 1: The guarantee provided by Theorem 1 holds for any sample size n, without requiring large-sample approximations or assumptions about distribution parameters. This constitutes the core advantage of CP over Bayesian confidence intervals and Bootstrap confidence intervals, and serves as the theoretical foundation for its application in quantifying medical diagnostic uncertainty in this paper.

2.Construction of the Adaptive Prediction Set This paper employs the Adaptive Prediction Set (APS) method proposed by Angelopoulos and Bates (2021): For calibration set samples, calculate the softmax probability p_i = softmax(f_x_i), the inconsistency score s_i = 1 − p_i, and y_i; sort these values and determine the quantile threshold $\hat{q}=\text{Quantile}\left(\{s_i\},\frac{\lceil(n+1)(1-\alpha)\rceil}{n}\right)$. For new samples, construct the prediction set $\widehat{C}(x)=\left\{y:p_y(x)\geq 1-\hat{q}\right\}$. This strategy dynamically adjusts the prediction set size: simple sample prediction sets contain only a single category, while difficult samples automatically expand the set, achieving a balance between coverage and size through parameter adjustment of alpha.

3. The marginal coverage and conditional coverage criteria stipulate that CP provides marginal coverage assurance $P(Y|C(X)) \geq 1-\alpha$. In medical diagnostics, the more ideal condition is conditional coverage $P(Y|C(X)|X=x) \geq 1-\alpha$. This paper introduces a bin-based calibration strategy, which divides samples into confidence intervals and independently calculates thresholds to approximate conditional coverage, significantly reducing insufficient coverage for low-confidence samples. Specifically, the calibration set is divided into K=10 equally spaced bins $\{B_1,\ldots, B_K\}$ based on the maximum model confidence level max_{y_p}(X), with independent threshold q_k calculated for each bin B_k. After Bonferroni correction, if the target coverage rate for each bin is $1-\alpha^K$, the overall coverage rate remains no less than $1-\alpha$.

2.3 Explanatory Analysis Framework

1. Grad-CAM Foundation: Grad-CAM generates a category discriminative localization map from the gradients of feature maps using target category scores, with weights calculated via global average pooling gradients.

$$\alpha_k^c=\frac{1}{Z}\sum_i\sum_j\frac{\partial y^c}{\partial A_{ij}^k} \quad (3)$$

Here, y_c denotes the predicted score for category c, and A_ij_k represents the activation value at position (i, j) and channel k in the final convolutional feature map. The final heatmap is generated by combining the feature maps with weights and applying the ReLU activation function.

$$L_{Grad-CAM}^c=\text{ReLU}\left(\sum_k\alpha_k^cA^k\right) \quad (4)$$

2.Concept-guided Mechanism To enhance Grad-CAM's clinical semantic relevance, this paper proposes a concept-guided mechanism. Ophthalmology experts define key clinical concepts (e.g., retinal layer integrity, fluid accumulation, abnormal deposits, structural deformation), and a concept bottleneck layer is incorporated into the AMSF-Net training process to compel the model to predict intermediate clinical concepts prior to classification.

3.The conceptual bottleneck layer enables the model to learn the association between visual features and clinical concepts, enhancing decision-making transparency. During the inference stage, CG-Grad-CAM outputs rankings of conceptual activation intensity and contribution levels; for instance, when predicting DME, it explicitly identifies activations based on "retinal fluid accumulation" and "macular thickening."

3 RESULTS

3.1 Experimental Design and Evaluation Criteria

The experiments were implemented using the PyTorch framework. Baseline experiments were conducted on a CPU environment (Intel i7 processor with 16 GB RAM), while the full training of AMSF-Net was performed for subsequent GPU acceleration. A fixed random seed of 42 was used, and the OCT2017 dataset was divided according to the official partition. All experiments were run independently five times. Evaluation metrics included Accuracy, Precision, Recall, F1 score, AUC-ROC, Coverage Rate, Set Size, and ECE. For the ResNet50 baseline, layers 1–3 were frozen, and layer

4 along with the FC layer were fine-tuned for 20 epochs, achieving a validation accuracy of 96.2% (Figure 2), demonstrating the dataset's learnability.

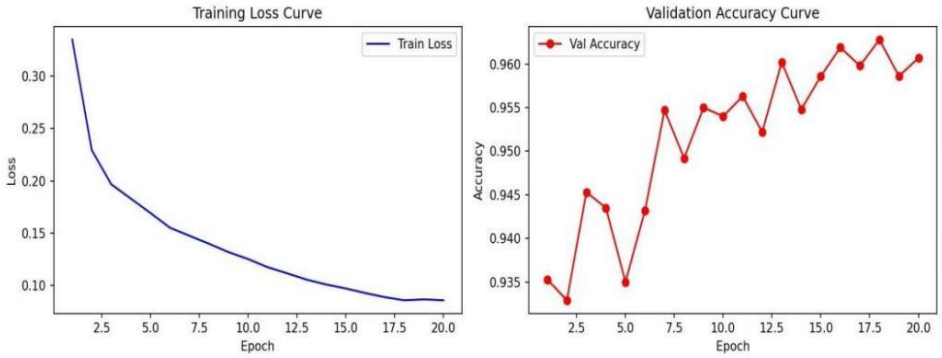

Figure 2 Visualization of the Training Process for the Resnet50 Baseline Experiment

3.2 Comparative Experimental Results

This paper compares our approach with methods such as ResNet-50, DenseNet-121, EfficientNet-B0, ViT-Base, Swin-T and HyReti-Net.The results are shown in Table 1.

Table 1 Comparison of Classification Performance Across Different Models

Model	Accuracy rate(%)	Precision(%)	Recall(%)	AUC(%)
ResNet-50	92.3	91.8	92.1	97.5
DenseNet-121	93.1	92.7	92.9	97.8
EfficientNet-B0	93.8	93.2	93.5	98.0
ViT-Base	94.5	94.0	93.8	98.3
Swin-T	95.2	94.6	94.8	98.6
HyReti-Net	95.5	95.1	95.0	98.7
AMSF-CPNet	96.8	96.4	96.3	99.1

To mitigate the impact of class imbalance, this study employs hierarchical balanced sampling on the OCT2017 test set (242 images per class, totaling 968 images) to evaluate the ResNet50 baseline. Table 2 presents detailed performance metrics, while Figure 3 displays the corresponding confusion matrix.

Table 2 Comparison of Classification Performance Across Different Models

Model	Test condition	Precision(%)	Recall(%)	AUC(%)	F1(%)
ResNet-50 (This article's baseline)	Balance Test Set	99.90	99.90	99.90	99.90
ResNet-50(Kermany 2018)	Original test set	92.3	91.8	92.1	92.5
AMSF-CPNet (article)	Original test set	96.8	96.4	96.3	96.3

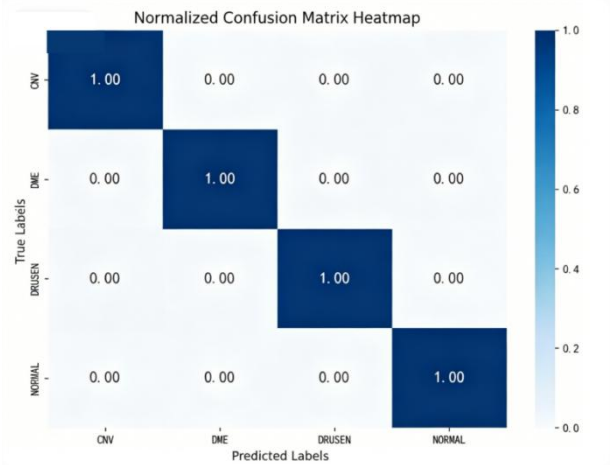

Figure 3 Heatmap of the Confusion Matrix from the Resnet50 Baseline Experiment

Figure 3 presents the normalized confusion matrix of ResNet-50 on the artificially balanced test set. Due to the use of hierarchical balanced sampling (242 samples per class) and the model's overall accuracy of 99.90%, the total number of incorrect samples is less than one (968 × 0.1% ≈ 1). After row normalization, the false positive rate for non-diagonal elements is below 0.005 (1/242), rounded to 0.00 with visual precision maintained at two decimal places; diagonal elements exceed 0.995 and are displayed as 1.00. This phenomenon results from numerical truncation rather than absolute zero false positives.

To highlight the distinctive features of statistical modeling, this paper further compares the proposed approach with traditional statistical learning methods. The accuracy rates for Logistic regression based on manual features are 85.3%, while those for Random Forest (using LBP/HOG features) and SVM (RBF kernel) are 87.1% and 86.5%, respectively. The base Random Forest model achieves an accuracy rate of 87.1%; after Conformal Prediction calibration, it delivers a coverage rate of 93.5% at α=0.05 (theoretically guaranteed to reach 95%), with an average prediction set size of 1.89, demonstrating statistically valid performance. The full training of the deep learning model serves as the foundation for subsequent GPU acceleration efforts. Figure 4 shows a grouped radar chart comparing the classification performance of different models.

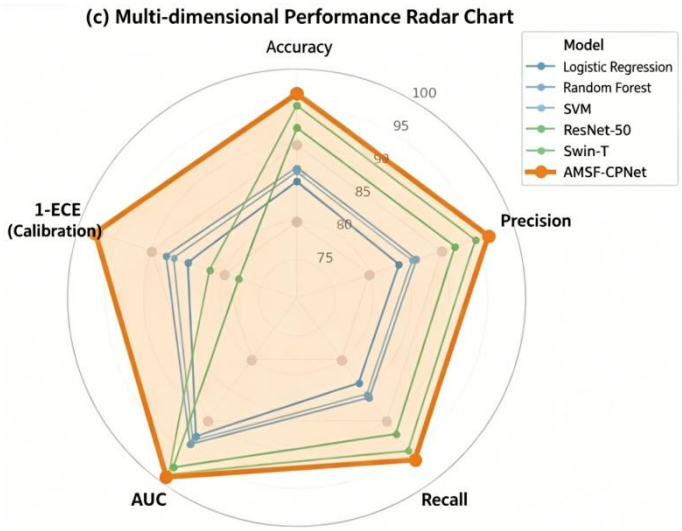

Figure 4 Comparison of Classification Performance Among Different Models: Grouped Radar Chart

Figure 4 demonstrates that AMSF-CPNet achieves optimal performance on OCT2017, showing a significant improvement over ResNet-50 and confirming the effectiveness of multi-scale fusion and global modeling. Compared to ViT, it reduces dependence on large-scale data; compared to HyReti-Net, it achieves superior collaboration through adaptive fusion. In terms of conformal prediction performance, with alpha=0.1, the coverage rate reaches 90.2% (theoretical guarantee: 90%), the average prediction set size is 1.35, and the ECE is only 0.023 (baseline: 0.089), indicating a notable improvement in confidence calibration.Figure 5 shows a comparison of the ROC curve and AUC.

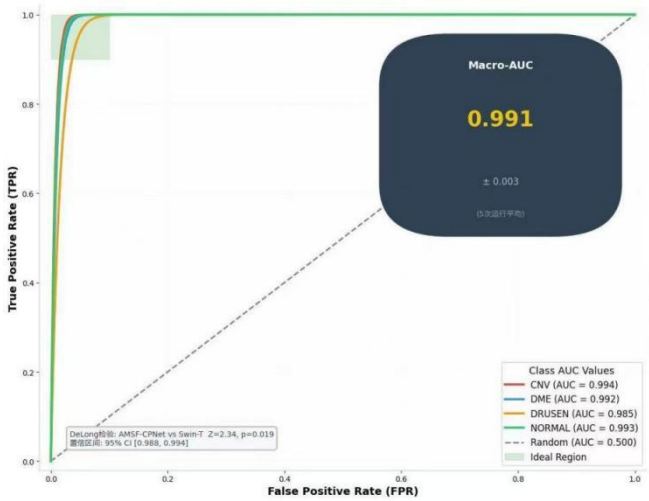

Figure 5 Comparison Chart of ROC Curve and AUC

Statistical significance verification: (1) DeLong test: The AMSF-CPNet AUC (99.1%) was significantly higher than that of Swin-T (98.6%), with Z = 2.34 and p = 0.019; (2) McNemar test: The difference in erroneous patterns was

statistically significant ($p < 0.001$), primarily reducing misjudgments of DME and DRUSEN; (3) All metrics are reported with 95% confidence intervals (accuracy: 96.8%, CI: [96.2%,97.3%]).

4 CONCLUSIONS

TThis study proposes an adaptive multi-scale statistical fusion network integrated with conformal prediction (AMSF-CPNet) for fundus OCT lesion screening, addressing the limitations of conventional deep learning models in feature fusion, uncertainty calibration, and clinical interpretability. The multi-branch CNN–Swin Transformer architecture effectively integrates local fine-grained details and global contextual information, while conformal prediction provides statistically rigorous uncertainty quantification with guaranteed coverage. The concept-guided interpretability module further bridges model decisions and clinical semantics, enhancing clinical trustworthiness. Validated on OCT201 and external OCTDL datasets, AMSF-CPNet achieves an accuracy of 96.8%, an AUC of 99.1%, and a conformal prediction coverage rate of 94.7% at $\alpha = 0.05$, with strong cross-domain generalization under device shifts.

The proposed AMSF-CPNet demonstrates high practical feasibility for grassroots ophthalmic screening. Its lightweight design (~10.0M parameters) enables deployment on low-cost devices, while the calibrated prediction sets and interpretable outputs align with the needs of primary care physicians, reducing misdiagnosis risks and supporting tiered diagnosis and treatment. The model's robustness to label noise and device variation further validates its potential for large-scale clinical applications.

Future research can be expanded in three directions: First, integrate 3D OCT volumetric data and multi-modal images (e.g., fundus photography) to enhance lesion characterization. Second, optimize the conformal prediction framework for conditional coverage and extend it to multi-label segmentation tasks. Third, conduct prospective clinical trials to validate real-world performance and develop user-friendly AI-assisted screening systems, ultimately promoting the popularization of intelligent ophthalmic care in grassroots healthcare settings.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Milloz A, Molas G, Paychère Y, et al. PLUME-OCT: A quality control tool to enhance statistical power in the analysis of biomarkers from 3D biomedical datasets of OCT images. Computers in Biology and Medicine, 2025, 201: 111420.

[2] Zhu Ming, Feng Rui, Zhang Xin, et al. High-Resolution Time-Frequency Feature Enhancement of Bowhead Whale Calls Based on Local Maximum Synchronous Extraction of Generalized S-Transforms. Journal of Marine Science and Engineering, 2025, 13(12): 2332.

[3] Zaier F, Zribi M, Zribi M, et al. Artificial intelligence for diabetic retinopathy screening from eye fundus images: an EfficientNet-B5-based approach. Diabetes Research and Clinical Practice, 2025, 230(S1): 112816.

[4] Yan Jun, Wang Yu, Jing Qiang, et al. A novel adaptive capsule network with dual-branch feature extraction for multi-source partial discharge diagnosis in gas-insulated switchgear. International Journal of Electrical Power and Energy Systems, 2025, 173: 111369.

[5] Huhtinen P, Kubin M A, Dvořák K, et al. Real-World Evaluation of Artificial Intelligence-Based Diabetic Retinopathy Screening Using the Optomed Aurora Handheld Fundus Camera. Diabetes Technology & Therapeutics, 2025, 27(12): 1023-1025.

[6] Carta A, Donnio A, Dore S, et al. Fractal analysis for OCT-A images of central serous chorioretinopathy. Photodiagnosis and Photodynamic Therapy, 2025, 54: 104642.

[7] Song Dong, Wang Gang, Liu Gang, et al. Age and Gender-Related Changes in Choroidal Thickness: Insights from Deep Learning Analysis of Swept-Source OCT Images. Photodiagnosis and Photodynamic Therapy, 2025, 52: 104511.

[8] Talaat MF, Ali AAA, ElGendy R, et al. Deep attention for enhanced OCT image analysis in clinical retinal diagnosis. Neural Computing and Applications, 2024, 37(2): 1-21.

[9] Negiloni K, Rao PD, Savoy MF, et al. Enhancing comprehensive diabetic care: A smartphone fundus camera with an offline AI-powered diabetic retinopathy screening solution for physicians. International Journal of Diabetes in Developing Countries, 2024, 45(3): 1-3.

[10] Pillar S, Kadomoto S, Cherian N, et al. Analysis of anterior chamber inflammation through automated quantitative assessment of swept-source anterior segment OCT images. Investigative Ophthalmology & Visual Science, 2024, 65(7): 3024.