GMR
The Original

RA-OCTNET: A RELIABILITY-ORIENTED LIGHTWEIGHT ATTENTION NETWORK FOR ROBUST AND EXPLAINABLE RETINAL OCT ANALYSIS

Senthilkumar V[1], Jayasundar S[2], A Jeeva[3], Manimaran[4], Naresh Sharma[5], Nagadeepan A[6], Asesh Kumar Tripathy[7] , Gourav Kalra[8]

[1]Professor, SRM TRP Engineering College, Trichy, Tamil Nadu, India. trpvsk12@gmail.com
[2]Professor, A.K.T.Memorial College of Engineering & Technology,Tamil Nadu, India, chisundar123@gmail.com
[3]Assistant Professor, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Tamil Nadu, India. drjeevaa@veltech.edu.in
[4]Associate Professor, Kongu Engineering College, Perundurai, Erode, Tamil Nadu, India. manimarankongu@gmail.com
[5]Assistant Professor, Department of Computer Science and Engineering, Faculty of Engineering and Technology, SRM Institute of Science and Technology, Delhi-NCR Campus, Delhi-Meerut Road, Modinagar, Ghaziabad, Uttar Pradesh, India. nrssharma@gmail.com
[6]Associate Professor, SRM TRP Engineering College, Trichy, Tamil Nadu, India. nagadeepan.a@gmail.com
[7]Associate Professor, Koneru Lakshmaiah Education Foundation, Andhra Pradesh, India, asesh.tripathy@gmail.com
[8]Associate Professor, School of Engineering and Technology, CGC University, Punjab, India, gkalra89@gmail.com
*Corresponding author: E-mail: nrssharma@gmail.com

ABSTRACT
Retinal optical coherence tomography (OCT) is widely used for the diagnosis and monitoring of retinal disease; however, automated OCT classification models require both high discriminative performance and reliability under image-quality variation. This study proposes RA-OCTNet, a robustness-aware attention-enhanced lightweight convolutional neural network for multiclass retinal OCT disease classification. The model incorporates squeeze-and-excitation channel attention, imbalance-sensitive optimisation, corruption-based robustness assessment and Grad-CAM visual explainability. Experiments were conducted on OCTMNIST 64 × 64 images covering choroidal neovascularisation, diabetic macular oedema, drusen and normal retina. RA-OCTNet achieved an accuracy of 0.8500, macro-F1 score of 0.8466 and macro ROC-AUC of 0.9854 using only 1.23 million trainable parameters. Robustness analysis demonstrated comparatively stable performance under contrast variation and occlusion, while Gaussian noise produced the largest degradation. These findings indicate that RA-OCTNet provides a compact, interpretable and reliability-oriented framework for retinal OCT disease classification.

KEYWORDS: Retinal optical coherence tomography; OCTMNIST; lightweight CNN; attention mechanism; robustness assessment; Grad-CAM

1. INTRODUCTION

Optical coherence tomography (OCT) is an important non-invasive retinal imaging modality used for the assessment, diagnosis and monitoring of sight-threatening retinal diseases. By providing high-resolution cross-sectional visualisation of retinal microstructure, OCT enables clinicians to identify pathological changes associated with conditions such as choroidal neovascularisation, diabetic macular oedema and drusen-related retinal degeneration. Recent OCT dataset and clinical AI studies further highlight the importance of OCT for automated retinal disease analysis and benchmarking [1], [2]. The increasing availability of OCT imaging in clinical practice has created a strong need for automated decision-support systems that can assist disease screening, reduce diagnostic workload and improve consistency in image interpretation.

Deep learning has substantially advanced automated medical image classification by enabling hierarchical feature learning directly from image data. Convolutional neural networks (CNNs) have been widely applied to retinal OCT classification because of their ability to capture local structural patterns, layer disruptions and disease-associated image characteristics. However, many high-performing deep learning models rely on large backbones with substantial computational requirements, limiting their suitability for lightweight clinical deployment, rapid screening workflows and resource-constrained environments. Recent lightweight OCT classification studies have therefore emphasised compact attention-guided CNN designs that balance diagnostic performance, computational efficiency and explainability [3]. Therefore, compact architectures that maintain strong discriminative ability while reducing model complexity remain clinically and practically relevant.

Despite progress in OCT classification, several challenges remain. First, retinal OCT images may vary due to acquisition quality, device-related differences, intensity variation, motion artefacts and noise. Models trained and evaluated only on clean benchmark images may not reliably reflect behaviour under degraded or shifted imaging conditions. Recent studies have shown that OCT classification performance can be affected by data scarcity, label noise and cross-dataset variation, reinforcing the need for reliability-oriented evaluation beyond clean-set accuracy [4], [5]. Second, disease classes may exhibit overlapping visual patterns, particularly in early or subtle pathological

presentations. Third, conventional classification metrics alone may not provide sufficient insight into model reliability, especially when models are intended for decision-support use in medical settings.

Attention mechanisms offer a promising strategy for improving feature selectivity in compact CNNs. In particular, channel-attention modules such as squeeze-and-excitation blocks can recalibrate feature responses by emphasising informative channels and suppressing less relevant activations. This can improve representational efficiency without requiring very large architectures. When combined with lightweight convolutional feature extraction, attention-enhanced CNNs can provide a balance between performance, interpretability and computational efficiency for retinal OCT disease classification. Related biomedical AI work using attention-based optimisation also supports the value of attention mechanisms for improving predictive modelling in clinical decision-support tasks [6].

In addition to clean-set classification performance, robustness and explainability are increasingly important for medical artificial intelligence. Robustness evaluation helps determine how model predictions change under clinically plausible image degradation, such as Gaussian noise, brightness shifts, contrast changes and occlusion. Explainability methods such as Grad-CAM can provide visual evidence of the image regions contributing to model predictions, supporting qualitative assessment of whether the model focuses on relevant retinal structures. A reliability-oriented OCT classification framework should therefore evaluate not only accuracy, but also class-wise behaviour, robustness degradation and visual interpretability.

The main contribution of this study is the development of RA-OCTNet, a robustness-aware attention-enhanced lightweight CNN for multiclass retinal OCT disease classification. The proposed framework integrates squeeze-and-excitation channel attention, imbalance-sensitive optimisation, corruption-based robustness assessment and Grad-CAM visual explainability. Experiments are conducted using OCTMNIST 64 × 64 images across four diagnostic categories: choroidal neovascularisation, diabetic macular oedema, drusen and normal retina. Unlike conventional OCT classification studies that primarily report clean-set performance, this work evaluates model behaviour under synthetic distribution shifts, including Gaussian noise, brightness variation, contrast change and occlusion. By combining lightweight classification, attention-guided feature recalibration, robustness assessment and explainable visual interpretation, the proposed framework provides a reliability-oriented approach for automated OCT disease classification.

The remainder of this paper is organised as follows. Section 2 reviews related work on retinal OCT classification, lightweight CNNs, attention mechanisms, robustness analysis and explainable medical image classification. Section 3 describes the OCTMNIST dataset, preprocessing pipeline, proposed RA-OCTNet architecture, training strategy, robustness protocol and evaluation metrics. Section 4 presents the experimental results, including classification performance, class-wise analysis, robustness evaluation and Grad-CAM visualisation. Section 5 discusses the findings, limitations and clinical relevance of the proposed framework. Finally, Section 6 concludes the paper and outlines future research directions.

2. LITERATURE REVIEW

Deep learning has become a major research direction in retinal optical coherence tomography (OCT) analysis because OCT B-scans provide detailed cross-sectional information about retinal morphology and disease-related structural abnormalities. Automated OCT classification is clinically relevant for supporting the identification of choroidal neovascularisation, diabetic macular oedema, drusen and normal retinal presentations. Public OCT datasets and lightweight biomedical image benchmarks have further accelerated this research area by enabling reproducible evaluation across common disease categories [7], [8].

Early OCT classification studies mainly focused on clean-set disease recognition using convolutional neural network architectures. These methods demonstrated that CNNs can learn disease-discriminative retinal features directly from OCT images and can support automated differentiation between pathological and normal retinal scans [9], [10]. Transfer learning and pretrained CNN backbones were subsequently adopted to improve classification performance, with commonly used architectures including Inception, ResNet, DenseNet, MobileNet and EfficientNet. Although these studies achieved strong diagnostic performance, most primarily reported clean-set metrics such as accuracy, precision, recall and area under the curve, with limited analysis of model reliability under degraded or shifted imaging conditions.

A second group of studies focused on lightweight OCT classification models to reduce computational complexity. Compact CNNs and efficient convolutional designs have been investigated to support faster inference, reduced memory usage and improved suitability for practical screening environments [11]. This direction is important because clinical decision-support tools may need to operate under constrained computational settings, including portable devices, screening platforms or low-resource deployment scenarios. However, lightweight design alone does not fully address robustness, class-wise reliability or interpretability, which remain important requirements for medical artificial intelligence systems [12] .

A third direction has introduced attention mechanisms to improve feature representation in OCT classification. Attention-based approaches aim to enhance diagnostically informative retinal features while suppressing less relevant activations. Multi-scale attention, spatial attention, channel attention and transformer-based approaches have been explored to capture local lesion patterns and broader retinal structural relationships [13],[14]. Although attention mechanisms can improve feature selectivity and representation quality, many existing studies still

emphasise clean-set classification performance rather than systematic assessment under image-quality variation or synthetic distribution shift.

A fourth category of OCT classification research has considered robustness, image quality and explainability. Medical images may be affected by noise, contrast variation, motion artefacts and acquisition-related degradation, and prior medical image enhancement research has shown the importance of improving image quality before downstream automated analysis. Explainable artificial intelligence methods, including gradient-based class activation mapping and related visualisation approaches, are used to identify image regions that contribute to model predictions [15], [16]. This is particularly important in OCT analysis because disease classes may share overlapping retinal morphology, and high numerical performance alone does not guarantee clinically meaningful decision behaviour. However, explainability is frequently presented as a supplementary visualisation step rather than being integrated into a broader reliability-oriented evaluation framework.

Despite these advances, several gaps remain in the current OCT classification literature. First, many studies focus mainly on clean-set classification accuracy and do not sufficiently evaluate model behaviour under image-quality degradation or distribution shift. Second, lightweight OCT models often report computational efficiency but may not include attention-guided feature recalibration, robustness testing and explainability within a unified framework. Third, attention-based models improve feature learning, but they are not always assessed under clinically plausible corruptions such as Gaussian noise, brightness variation, contrast change and occlusion. Fourth, although explainability methods are increasingly used, they are frequently presented as supplementary visual evidence rather than as part of a complete reliability assessment pipeline.

The present work addresses these gaps by proposing RA-OCTNet, a robustness-aware attention-enhanced lightweight CNN for retinal OCT disease classification. The framework integrates squeeze-and-excitation channel attention for feature recalibration, imbalance-sensitive optimisation, corruption-based stress testing and Grad-CAM visual explainability. Unlike studies that evaluate only clean classification performance, the proposed approach examines model behaviour under Gaussian noise, brightness variation, contrast change and occlusion. By combining lightweight classification, attention-guided representation learning, robustness assessment and explainable visual interpretation within a unified pipeline, this study provides a reliability-oriented framework for automated OCT disease classification.

3. MATERIALS AND METHODS

This section describes the methodological framework used to develop and evaluate RA-OCTNet, a robustness-aware attention-enhanced lightweight convolutional neural network for retinal optical coherence tomography (OCT) disease classification. The study was designed as a structured medical image classification experiment using a publicly available benchmark dataset. The methodology consisted of four main stages: dataset formulation and preprocessing, development of the proposed RA-OCTNet architecture, model training and validation, and evaluation using classification metrics, robustness assessment and explainability analysis. The proposed framework is illustrated in Fig. 1.

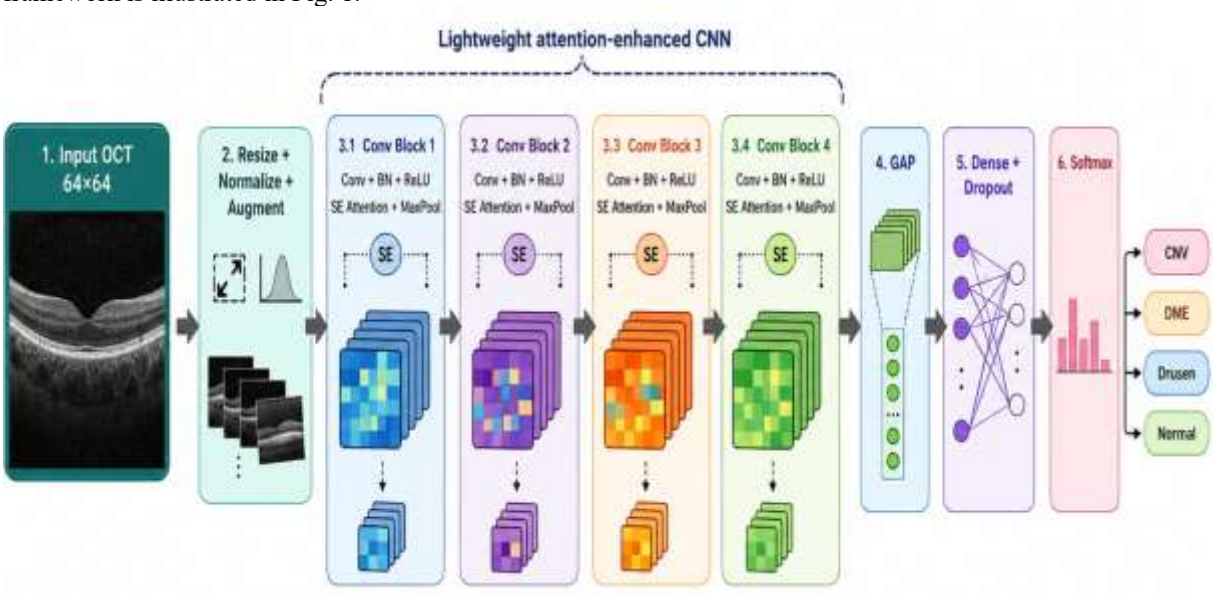

Fig. 1. Proposed RA-OCTNet architecture for retinal OCT disease classification.

The model processes 64 × 64 OCT images using preprocessing and augmentation, followed by four attention-enhanced convolutional blocks with squeeze-and-excitation modules, global average pooling, dense classification and softmax-based prediction into CNV, DME, drusen and normal classes.

3.1 Dataset and Study Design

The OCTMNIST dataset was used as the benchmark dataset for retinal OCT disease classification. The dataset contains grayscale OCT images organised into four diagnostic categories: choroidal neovascularisation, diabetic

macular oedema, drusen and normal retina. The task was defined as multiclass retinal disease classification, where each input OCT image was assigned to one of the four target categories.
The dataset was represented as:
$D = \{(x_i, y_i)\}_{i=1}^{N}$ (1)
where x_i denotes the OCT image for sample i, $y_i \in \{0, 1, 2, 3\}$ denotes the corresponding diagnostic label, and N denotes the total number of images. The model estimated the class probability distribution as:
$\hat{y}_i = f\theta(x_i)$ (2)
where $f\theta$ denotes the proposed neural network with trainable parameters θ, and $\hat{y}_i$ denotes the predicted probability vector across the four diagnostic categories.
The dataset was used in its 64 × 64 image format to support lightweight model development and efficient experimentation. Each image was treated as an independent sample with a corresponding diagnostic label. The objective of the model was to learn disease-discriminative retinal features and predict the diagnostic category of each OCT image. Representative image samples and class distribution are presented in Fig. 2.

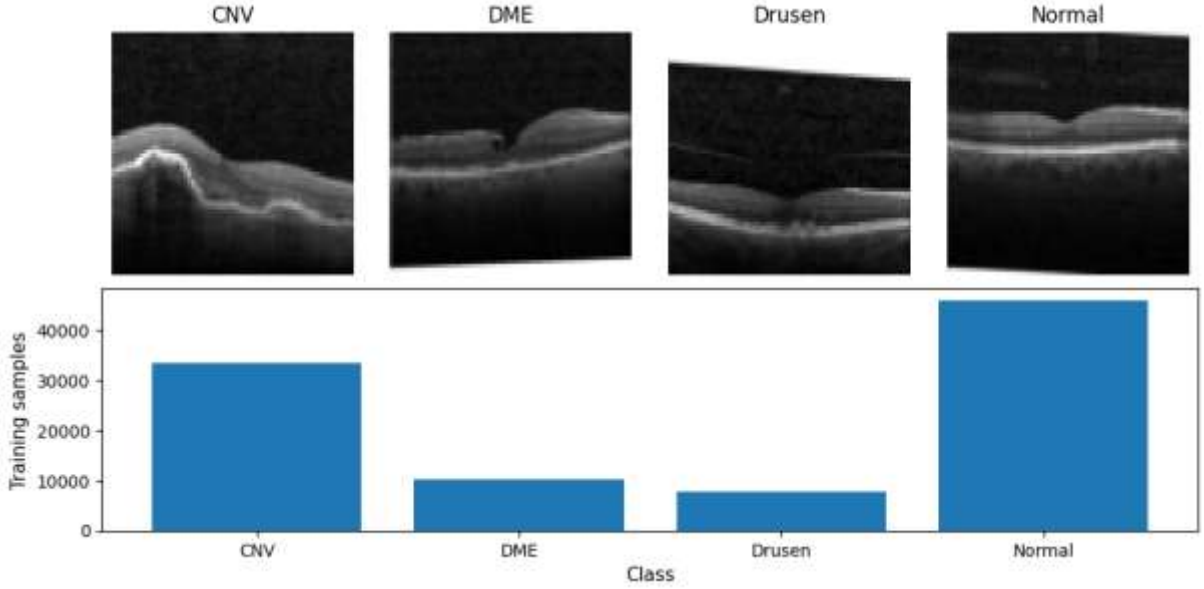

Fig. 2. OCTMNIST dataset overview showing representative OCT images from the four diagnostic categories and the corresponding training class distribution used for model development.

All images were resized and normalised before being passed to the model. Data augmentation was applied only to the training set to improve generalisation and reduce sensitivity to minor acquisition-related variations. The augmentation operations included horizontal flipping, small-angle rotation, affine transformation, mild scaling, brightness adjustment and contrast adjustment. Validation and test images were processed using deterministic resizing and normalisation without random augmentation.

3.2 Proposed RA-OCTNet Framework

RA-OCTNet was developed as a compact attention-enhanced CNN for retinal OCT classification. The architecture consisted of four convolutional feature extraction blocks followed by global average pooling and a dense classification head. Each convolutional block included convolution, batch normalisation, rectified linear unit activation, max-pooling and squeeze-and-excitation channel attention.
For an intermediate feature map, the convolutional transformation was expressed as:
$h^l = \phi(BN(W^l * h^{l-1} + b^l))$ (3)
where h^l denotes the feature representation at layer l, W^l and b^l denote the convolutional weights and bias terms, BN denotes batch normalisation, * denotes convolution, and $\phi(.)$ denotes ReLU activation.
The squeeze-and-excitation attention modules were incorporated to recalibrate channel-wise feature responses. This allowed the model to emphasise informative retinal image features while suppressing less relevant activations. The channel attention operation was represented as:
$\tilde{U} = s \cdot U$ (4)
where U denotes the input feature map, s denotes the learned channel-attention weights, and $\tilde{U}$ denotes the recalibrated feature representation. After feature extraction, global average pooling was used to generate a compact image representation, which was passed through fully connected layers to produce four-class prediction probabilities using softmax activation.
The framework was designed to provide a balance between model compactness, feature selectivity and interpretability. Instead of using a large pretrained backbone, the proposed model used a lightweight custom CNN with integrated attention modules. This design supported efficient training and enabled visual explanation through Grad-CAM.

3.3 Training, Optimisation and Validation

Model training was performed using class-weighted cross-entropy loss to reduce the effect of class imbalance. Class weights were estimated from the training data and incorporated into the optimisation process so that under-represented categories were not ignored during learning. The weighted cross-entropy loss was defined as:

$$L = -\sum_{k=1}^{K} w_k \, y_k \log(p_k) \quad (5)$$

where K denotes the number of classes, w_k denotes the class weight for category k, y_k denotes the ground-truth label indicator, and p_k denotes the predicted probability for class k.

AdamW optimisation was used for model parameter updates, and dropout was applied in the classifier head to reduce overfitting. A validation set was used to monitor model performance during training. Validation macro-F1 was used as the model-selection criterion because it provides class-sensitive evaluation in multiclass classification. The model checkpoint with the best validation macro-F1 score was retained for final evaluation.

3.4 Evaluation, Robustness and Explainability

The trained model was evaluated using standard multiclass classification measures, including accuracy, balanced accuracy, precision, recall, macro-F1, weighted-F1 and multiclass ROC-AUC. These metrics were selected to assess overall classification performance, class-sensitive behaviour and probabilistic discrimination across the four OCT diagnostic categories.

To assess reliability beyond clean image classification, robustness testing was conducted using synthetic image corruptions. Gaussian noise, brightness variation, contrast change and occlusion were applied to test images at multiple severity levels. Performance degradation under corrupted inputs was assessed relative to clean-test performance as:

$$\Delta M = M_{clean} - M_{corrupted} \quad (6)$$

where M denotes the selected evaluation metric. This assessment was used to examine model behaviour under image-quality degradation and synthetic distribution shift.

Grad-CAM was used to provide visual explainability for model predictions. The final convolutional layer was selected for heatmap generation, and the resulting activation maps were overlaid on the original OCT images to identify image regions contributing to model decisions.

4. RESULTS

The experimental evaluation of RA-OCTNet was conducted on the OCTMNIST 64 × 64 test set to assess training behaviour, clean-test classification performance, class-wise prediction behaviour, robustness under synthetic image degradation and Grad-CAM-based explainability. RA-OCTNet was evaluated on the OCTMNIST 64 × 64 test set. The evaluation included training behaviour, clean-test classification performance, class-wise prediction patterns, robustness under synthetic image degradation and Grad-CAM visualisation. Fig. 3 presents the training behaviour, Fig. 4–6 present clean-test classification performance, Fig. 7–9 present robustness analysis, and Fig. 10 presents Grad-CAM visualisation.

4.1 Training and Validation Performance

RA-OCTNet was trained for 10 epochs using GPU acceleration. The training process showed stable convergence, with training loss decreasing from 0.3831 in the first epoch to 0.1363 in the final epoch. Validation loss also decreased from 0.2111 to 0.1063, indicating progressive optimisation during training. Validation accuracy improved from 0.9353 to 0.9666, while validation macro-F1 increased from 0.8841 to 0.9406. The best model checkpoint was selected using validation macro-F1 as the monitoring criterion. These trends are shown in Fig. 3.

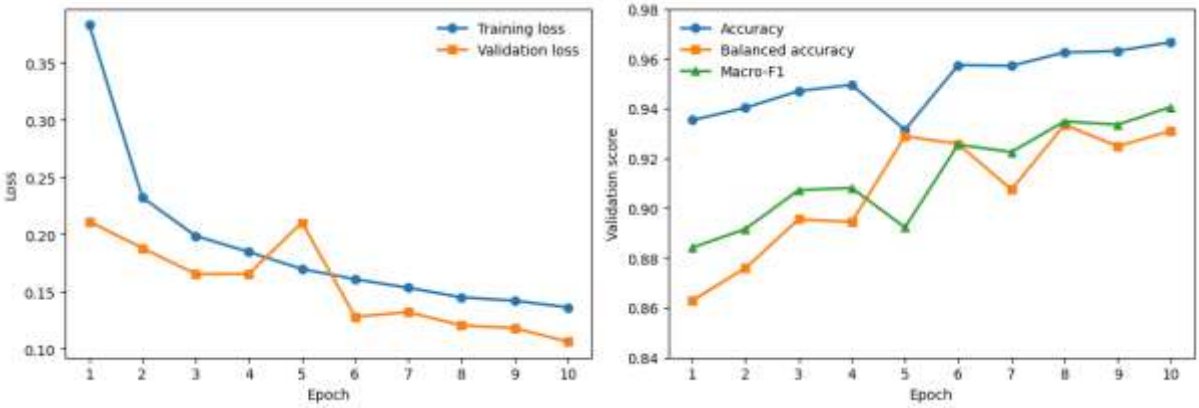

Fig. 3. Training behaviour of RA-OCTNet. The left panel shows training and validation loss, while the right panel shows validation accuracy, balanced accuracy and macro-F1 across 10 epochs.

4.2 Clean-Test Classification Performance

On the independent test set, RA-OCTNet achieved an accuracy of 0.8500 and balanced accuracy of 0.8500. The macro precision, macro recall and macro-F1 score were 0.8913, 0.8500 and 0.8466, respectively. The weighted precision, weighted recall and weighted-F1 score were also 0.8913, 0.8500 and 0.8466 because the test set was balanced across the four diagnostic categories. The model achieved a macro ROC-AUC of 0.9854 and weighted ROC-AUC of 0.9854 across the four OCT classes. RA-OCTNet contained 1,228,352 trainable parameters. The

confusion matrix and normalised confusion matrix are shown in Fig. 4. The one-vs-rest ROC curves are shown in Fig. 5, and the precision–recall curves are shown in Fig. 6.

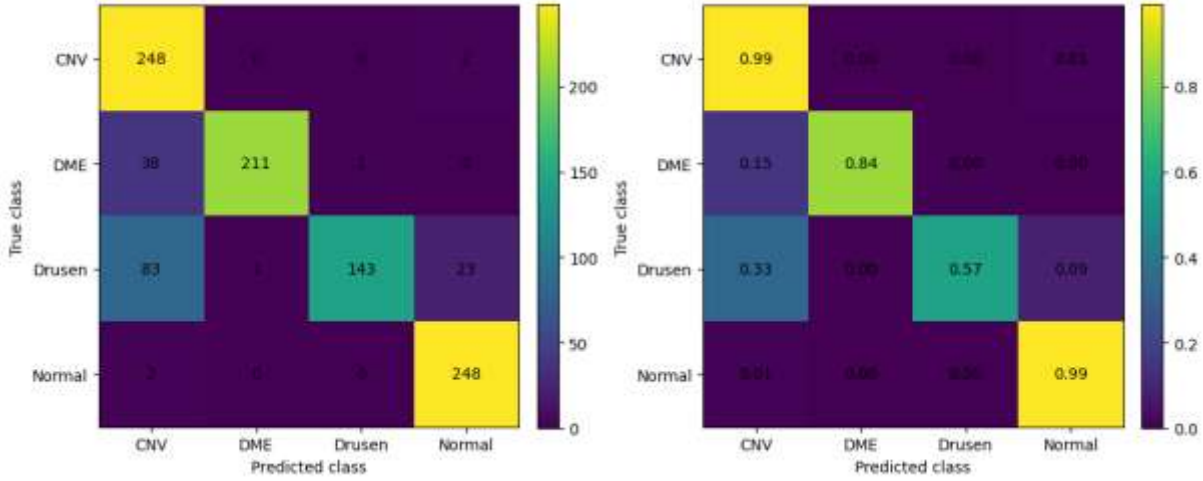

Fig. 4. Confusion-matrix analysis of RA-OCTNet on the independent OCTMNIST test set.

The left panel shows the raw confusion matrix, and the right panel shows the row-normalised confusion matrix across the four diagnostic categories.

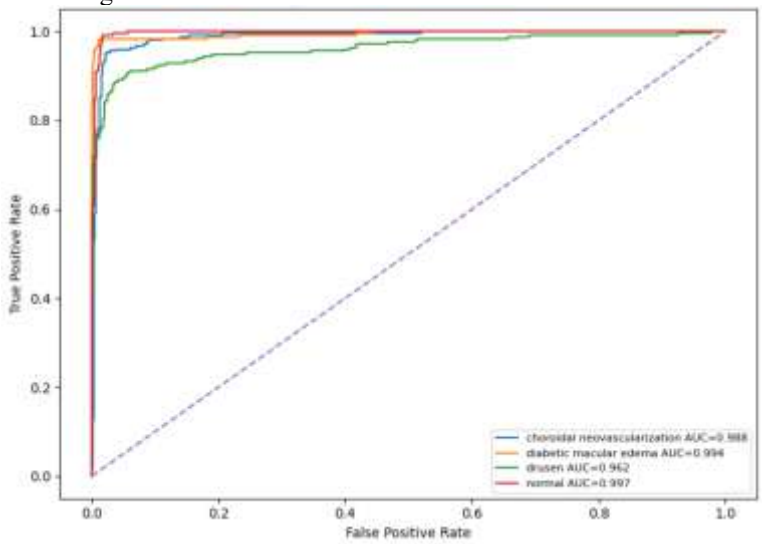

Fig. 5. One-vs-rest ROC curves for RA-OCTNet across the four OCT diagnostic categories.

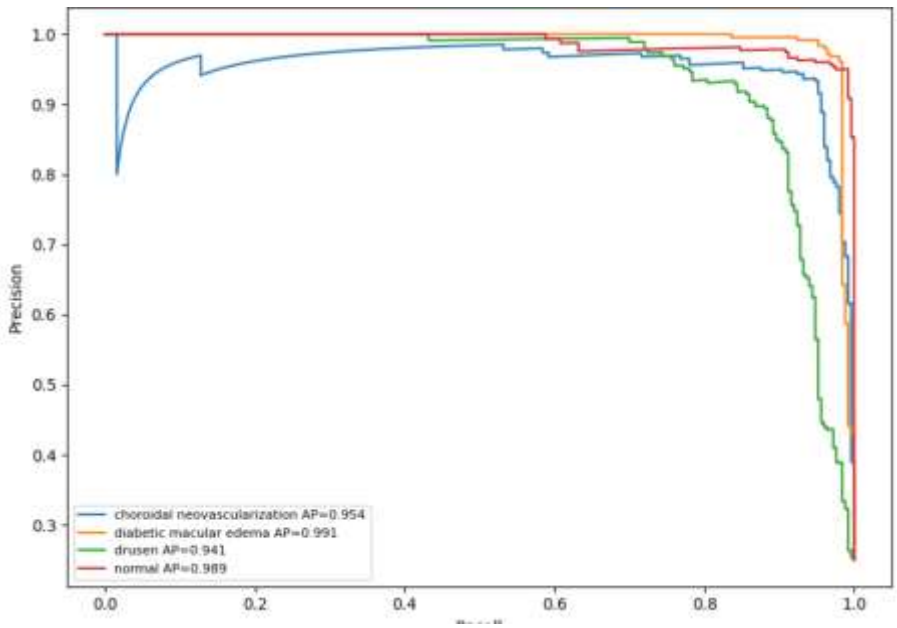

Fig. 6. One-vs-rest precision–recall curves for RA-OCTNet across the four OCT diagnostic categories.

4.3 Class-Wise Classification Performance

Class-wise analysis showed variation in performance across the four retinal OCT categories. Normal retina achieved the highest F1-score, with precision of 0.9084, recall of 0.9920 and F1-score of 0.9484. Diabetic macular oedema also showed strong performance, with precision of 0.9953, recall of 0.8440 and F1-score of 0.9134.

Choroidal neovascularisation achieved high recall of 0.9920 but lower precision of 0.6685, indicating that some images from other diagnostic categories were incorrectly assigned to this class. Drusen was the most challenging category, with precision of 0.9931, recall of 0.5720 and F1-score of 0.7259. The lower recall indicates that some drusen cases were misclassified as other retinal disease categories.

4.4 Robustness under Synthetic Image Corruptions

Robustness testing was performed using Gaussian noise, brightness variation, contrast change and occlusion across five severity levels. Gaussian noise produced the largest degradation in model performance. Under Gaussian noise severity level 1, accuracy decreased to 0.5740 and macro-F1 decreased to 0.4862. At severity level 5, accuracy decreased further to 0.2520 and macro-F1 to 0.1041. This indicates that random noise strongly affected the discriminative retinal features learned by the model.

Brightness variation produced moderate degradation. Accuracy decreased from 0.8240 at severity level 1 to 0.7180 at severity level 5, while macro-F1 decreased from 0.8143 to 0.6821. In comparison, contrast change produced a smaller performance reduction, with accuracy remaining 0.7950 and macro-F1 remaining 0.7881 at severity level 5. Occlusion also showed comparatively limited degradation, with accuracy of 0.7600 and macro-F1 of 0.7513 at severity level 5.

The macro-F1 robustness trends are shown in Fig. 7, and the accuracy degradation trends are shown in Fig. 8. Representative corrupted OCT images are shown in Fig. 9.

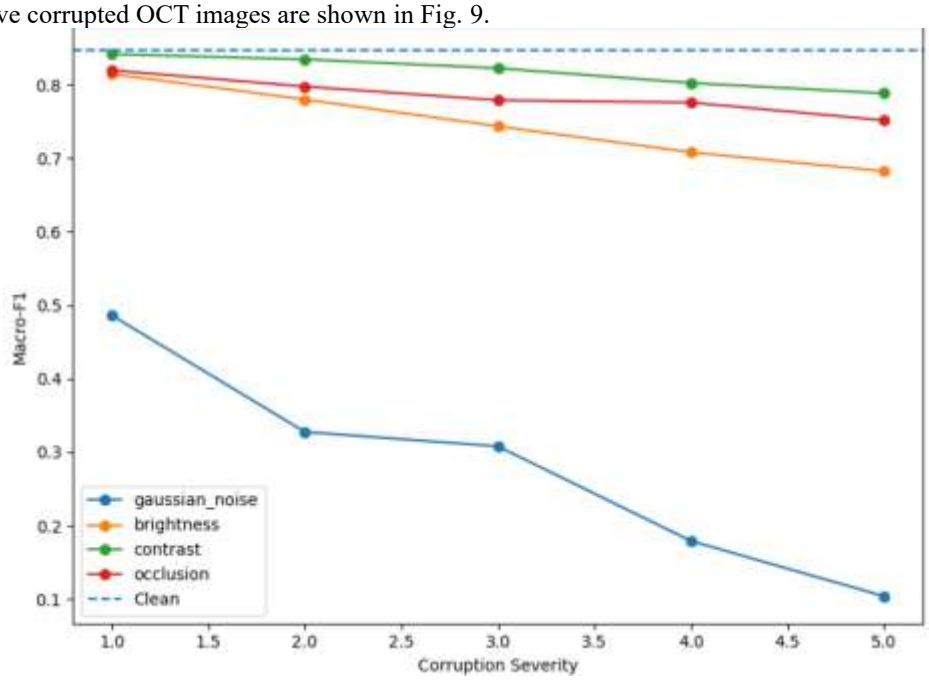

Fig. 7. Macro-F1 robustness under synthetic distribution shift.

Macro-F1 performance across Gaussian noise, brightness variation, contrast change and occlusion at increasing severity levels.

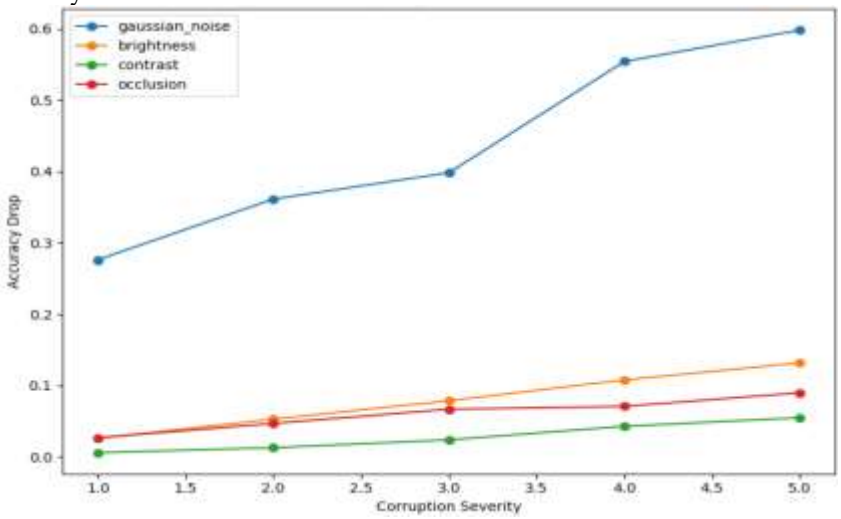

Fig. 8. Accuracy degradation under synthetic corruptions.

Accuracy drop across corruption types and severity levels relative to clean-test performance.

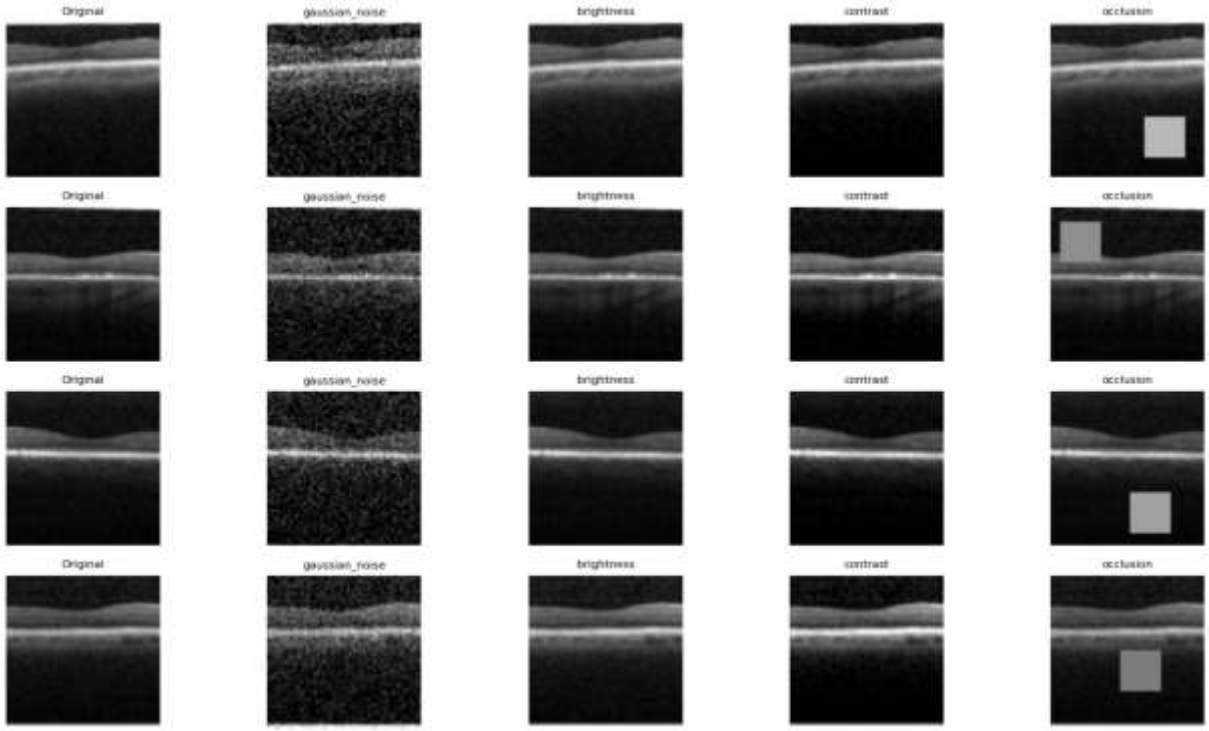

Fig. 9. Representative corrupted OCT images.
Example OCT images after Gaussian noise, brightness variation, contrast change and occlusion.

4.5 Grad-CAM Explainability

Grad-CAM visualisations were generated using the final convolutional layer of RA-OCTNet to examine the image regions contributing to model predictions. The heatmaps were overlaid on the original OCT images to provide qualitative interpretation of the learned decision behaviour. The visualisations highlighted retinal regions that influenced the predicted diagnostic class, supporting interpretability analysis of the proposed framework. Representative Grad-CAM outputs are shown in Fig. 10.

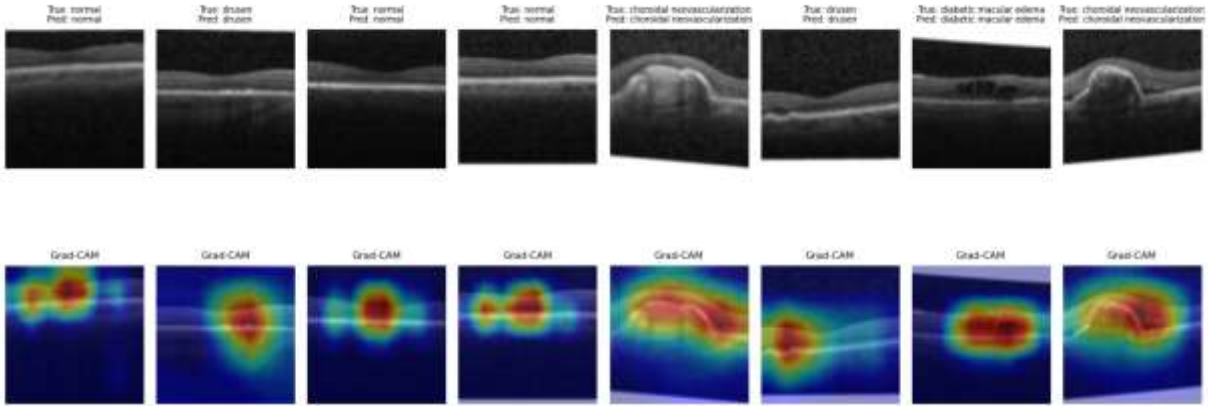

Fig. 10. Grad-CAM explainability of RA-OCTNet.
Representative Grad-CAM heatmaps showing retinal regions contributing to model predictions.

5. DISCUSSION

The experimental findings demonstrate that RA-OCTNet provides a compact and reliability-oriented framework for retinal OCT disease classification. The model achieved strong probabilistic discrimination while maintaining a lightweight architecture, indicating that attention-enhanced CNNs can provide effective feature representation without relying on large pretrained backbones. This is important for OCT-based decision-support systems, where computational efficiency, interpretability and robustness are relevant for practical deployment.

The clean-test results showed that RA-OCTNet performed well across the four diagnostic categories, although class-wise variation was observed. Normal retina and diabetic macular oedema achieved stronger F1-scores, while drusen was the most challenging class. The lower recall for drusen suggests that some drusen images shared visual similarity with other OCT categories, making class separation more difficult. This is clinically plausible because retinal disease patterns may overlap in texture, layer disruption and structural appearance, particularly when images are resized to a compact input resolution.

Choroidal neovascularisation achieved high recall but lower precision, indicating that the model detected most true choroidal neovascularisation cases but also assigned some non-CNV images to this class. This behaviour

suggests that the model may have learned sensitive CNV-related features but with reduced specificity. In a clinical decision-support context, such a pattern may be useful for screening but would require further calibration or threshold adjustment to reduce false-positive predictions.

The robustness analysis provided an important extension beyond clean-set evaluation. RA-OCTNet was comparatively stable under contrast variation and occlusion but showed substantial sensitivity to Gaussian noise. This indicates that random pixel-level disturbance can disrupt the learned retinal feature representation more severely than structured intensity or partial-visibility changes. The finding highlights the need to evaluate OCT classification models under image-quality degradation rather than relying only on clean benchmark performance.

The comparatively smaller degradation under contrast variation and occlusion suggests that the model retained useful structural information even when global intensity scaling or local image masking was applied. This may be due to the convolutional architecture and squeeze-and-excitation attention modules, which support feature recalibration across channels. However, the strong degradation under Gaussian noise indicates that additional training strategies, such as noise-aware augmentation, denoising preprocessing or corruption-consistent training, may be needed to improve reliability under low-quality acquisition conditions.

Grad-CAM visualisation provided qualitative insight into the image regions contributing to RA-OCTNet predictions. The activation maps helped inspect whether the model focused on retinal areas associated with the predicted class. Although Grad-CAM does not provide a complete clinical explanation, it offers a useful visual tool for assessing model behaviour and identifying whether predictions are driven by plausible retinal regions rather than irrelevant background patterns.

Overall, RA-OCTNet addresses an important gap in OCT classification by combining lightweight CNN design, attention-guided feature recalibration, robustness assessment and explainability within a single framework. The results support the value of reliability-oriented evaluation for medical image classification, particularly in applications where image quality may vary across acquisition settings.

Several limitations should be noted. First, the evaluation was conducted using a benchmark dataset rather than external multicentre clinical data. Second, the OCT images were resized to 64 × 64 pixels, which supports lightweight experimentation but may reduce fine-grained retinal detail. Third, the robustness analysis used synthetic corruptions rather than real acquisition artefacts. Fourth, Grad-CAM provides qualitative visual interpretation but does not replace expert clinical validation. Future work should therefore evaluate the framework on higher-resolution OCT images, external datasets, real-world image-quality variations and clinician-reviewed explainability outputs.

6. CONCLUSION AND FUTURE WORK

This study proposed RA-OCTNet, a robustness-aware attention-enhanced lightweight CNN for retinal OCT disease classification. The framework combined compact convolutional feature extraction, squeeze-and-excitation channel attention, imbalance-sensitive training, corruption-based robustness assessment and Grad-CAM visual explainability. The experimental results showed that RA-OCTNet achieved effective clean-test classification performance while maintaining a lightweight architecture suitable for efficient OCT image analysis.

The robustness evaluation demonstrated that model performance varied across different image corruptions. RA-OCTNet was comparatively stable under contrast variation and occlusion, but Gaussian noise caused the largest degradation. This finding highlights the importance of evaluating medical image classification models under synthetic distribution shifts, rather than relying only on clean benchmark performance. Grad-CAM visualisations further supported qualitative interpretation by highlighting image regions that contributed to the model's predictions.

Future work will focus on improving robustness to noise and acquisition-related image degradation through noise-aware augmentation, denoising preprocessing and corruption-consistent training. Further evaluation using higher-resolution OCT images, external clinical datasets and multicentre validation is also needed to assess generalisability. In addition, future studies should incorporate expert ophthalmologist review of Grad-CAM explanations and compare RA-OCTNet with larger pretrained CNN and transformer-based models under the same robustness protocol. These extensions may support the development of more reliable and interpretable OCT-based decision-support systems.

Declarations

Ethics statement

This study was conducted using publicly available datasets. No research involving identifiable human participants was carried out. Accordingly, ethical approval was not required.

Consent to participate

Not applicable.

Consent to publish

Not applicable.

Funding
No funding was received for this study.

Data Availability
The dataset analysed in this study is publicly available through the MedMNIST benchmark collection under the OCTMNIST dataset: https://medmnist.com/. OCTMNIST is derived from retinal optical coherence tomography images and includes four diagnostic categories: choroidal neovascularisation, diabetic macular oedema, drusen and normal retina. The dataset was used in its 64 × 64 image format for multiclass retinal OCT disease classification. The data were accessed and used in accordance with the terms and licence conditions of the MedMNIST repository.

REFERENCES
[1] M. Kulyabin et al., "OCTDL: Optical Coherence Tomography Dataset for Image-Based Deep Learning Methods," Sci Data, vol. 11, no. 1, p. 365, Apr. 2024, doi: 10.1038/s41597-024-03182-7.
[2] Y. Peng et al., "Enhancing AI reliability: A foundation model with uncertainty estimation for optical coherence tomography-based retinal disease diagnosis," Cell Reports Medicine, vol. 6, no. 1, p. 101876, Jan. 2025, doi: 10.1016/j.xcrm.2024.101876.
[3] M. R. Ahmed, M. S. Shehata, and P. Lasserre, "Integrating lightweight convolutional neural network with entropy-informed channel attention and adaptive spatial attention for OCT-based retinal disease classification," Computers in Biology and Medicine, vol. 190, p. 110071, May 2025, doi: 10.1016/j.compbiomed.2025.110071.
[4] A. Miladinović et al., "Evaluating deep learning models for classifying OCT images with limited data and noisy labels," Sci Rep, vol. 14, no. 1, p. 30321, Dec. 2024, doi: 10.1038/s41598-024-81127-1.
[5] P. Zhang et al., "A robust deep learning classifier for screening multiple retinal diseases on optical coherence tomography," Sci Rep, vol. 15, no. 1, p. 35334, Oct. 2025, doi: 10.1038/s41598-025-19286-y.
[6] D. Saleela, A. S. Oyegoke, J. A. Dauda, and S. O. Ajayi, "Development of AI-Driven Decision Support System for Personalized Housing Adaptations and Assistive Technology," Journal of Aging and Environment, pp. 1–24, Jul. 2025, doi: 10.1080/26892618.2025.2534956.
[7] A. Rozhyna et al., "Exploring Publicly Accessible Optical Coherence Tomography Datasets: A Comprehensive Overview," Diagnostics, vol. 14, no. 15, p. 1668, Aug. 2024, doi: 10.3390/diagnostics14151668.
[8] B. Hassan et al., "A comprehensive review of artificial intelligence models for screening major retinal diseases," Artif Intell Rev, vol. 57, no. 5, p. 111, Apr. 2024, doi: 10.1007/s10462-024-10736-z.
[9] R. R et al., "Deployment Aligned Hybrid Attention For Diabetic Retinopathy Screening With Mdrnet," IJASIS, vol. 12, no. 1, pp. 1054–1064, Mar. 2026, doi: 10.29284/yyhkzs56.
[10] S. Divya, L. Padma Suresh, and A. John, "Enhanced deep-joint segmentation with deep learning networks of glioma tumor for multi-grade classification using MR images," Pattern Anal Applic, vol. 25, no. 4, pp. 891–911, Nov. 2022, doi: 10.1007/s10044-022-01064-5.
[11] S. Divya, L. Padma Suresh, and A. John, "Hybrid Optimization Algorithm-Based Generative Adversarial Network for Change Detection Using Pre-Operative and Post-Operative MRI," Int. J. Patt. Recogn. Artif. Intell., vol. 36, no. 07, p. 2251007, Jun. 2022, doi: 10.1142/S0218001422510077.
[12] Sujith Kumar P S, "A LIGHTWEIGHT MULTIMODAL DEEP LEARNING FRAMEWORK FOR ACCURATE SEGMENTATION OF ACUTE ISCHAEMIC STROKE LESIONS," ijam, vol. 38, no. 12s, pp. 1919–1939, Dec. 2025, doi: 10.12732/ijam.v38i12s.1516.
[13] A. M. Alenezi et al., "Multiscale attention-over-attention network for retinal disease recognition in OCT radiology images," Front. Med., vol. 11, p. 1499393, Nov. 2024, doi: 10.3389/fmed.2024.1499393.
[14] J. S. Yang, Z. H. Ning, and X. Xiao, "Pathology-informed multi-label retinal disease classification via multi-scale transformer and graph integration," Biomedical Signal Processing and Control, vol. 126, p. 110789, Oct. 2026, doi: 10.1016/j.bspc.2026.110789.
[15] B. H. M. Van Der Velden, H. J. Kuijf, K. G. A. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis," Medical Image Analysis, vol. 79, p. 102470, Jul. 2022, doi: 10.1016/j.media.2022.102470.
[16] F. Ahmed, N. S. Naz, S. Khan, A. U. Rehman, W. M. Ismael, and M. A. Khan, "Explainable artificial intelligence (XAI) in medical imaging: a systematic review of techniques, applications, and challenges," BMC Med Imaging, vol. 26, no. 1, p. 37, Jan. 2026, doi: 10.1186/s12880-025-02118-w.