STU
SLOVAK UNIVERSITY OF TECHNOLOGY IN BRATISLAVA

Journal of Electrical Engineering, Vol. 76, No. 5, 2025, pp. 476-484
https://reference-global.com/journal/JEE

From public to clinical data: External validation of an explainable MedViT model for retinal OCT images

Samuel Gibala[1], Veronika Kurilova[1,2], Milos Oravec[1], Jarmila Pavlovicova[1], Jana Stefanickova[3,4]

The performance of machine learning models is often evaluated using accuracy metrics, which provide only a partial view of their capabilities, particularly in medical imaging, where explainability is essential. In this study, we propose to use a hybrid CNN-Transformer model MedViT, for the classification of retinal OCT images. The model achieved 95.95% accuracy on the modified public Retinal OCT C8 dataset and 90% on a private independent external test dataset, correctly classifying 9 out of 10 cases. Explainability analysis with Grad-CAM showed that the model consistently focused on clinically relevant macular regions. These results indicate that the proposed approach can accurately detect pathological changes and focus on diagnostically important regions also in an independent dataset beyond the training data, making it a reliable and helpful tool for ophthalmologists in clinical practice.

Keywords: computer vision, hybrid CNN-Transformer, explainable AI, OCT, diabetic macular oedema

1 Introduction

Optical coherence tomography (OCT) is a non-invasive imaging technique that enables high-resolution visualization of the retinal structure. Its ability to provide cross-sectional images of individual retinal layers makes it an essential tool for diagnosing and monitoring a wide range of ophthalmic diseases [1].

This study focuses on the macular region, the central part of the retina responsible for sharp, detailed vision. The macula contains a high density of cone photoreceptors essential for color perception and fine spatial resolution [2]. Structural changes in this area can lead to severe visual impairment, highlighting the importance of accurate detection and diagnosis using OCT imaging.

Automated analysis of OCT scans using machine learning (ML) and deep learning techniques offers significant potential for improving diagnostic efficiency and accuracy. In particular, hybrid CNN-Transformer models have recently demonstrated strong performance in medical image classification tasks due to their ability to capture both local and long-range feature relationships [3].

However, for AI-based diagnostic tools to be effectively implemented in clinical practice, their decision-making processes must be interpretable and explainable [4-6]. Merely achieving high classification accuracy is not sufficient. It is equally important to understand which image regions contribute to the model's predictions. Such interpretability can support ophthalmologists in clinical decision-making, increase the trust in automated systems, and facilitate their integration into diagnostic workflows.

In this work, we therefore investigate the use of the hybrid model MedViT for multi-class classification of retinal OCT images from Retinal OCT C8 dataset [7], with a particular emphasis on explainability and external clinical validation. To further assess the robustness and real-world applicability of the proposed approach, we also perform independent external validation using a subset of private clinical OCT dataset. Our approach aims to evaluate not only the classification performance of the model but also the clinical relevance of its decision-making, thereby exploring its potential as a reliable decision-support tool in ophthalmology.

2 Related works

Research on automated classification of OCT images using deep neural networks has been rapidly evolving, covering a wide range of architectures – from classical convolutional neural networks (CNNs) and their modifications to modern transformer-based approaches.

[1] Faculty of Electrical Engineering and Information Technology, Slovak University of Technology in Bratislava, Ilkovicova 3, 812 19, Bratislava, Slovak Republic
[2] Faculty of Medicine, Slovak Medical University, Limbova 12, 833 03 Bratislava, Slovak Republic
[3] Department of Ophthalmology, Comenius University, Bratislava, Ruzinovska 6, 826 06 Bratislava, Slovak Republic
[4] Eye Center, Hospital Malacky, Duklianskych hrdinov 34, 901 01 Malacky, Slovak Republic
Corresponding authors: samuel.gibala@stuba.sk, jstefanicka@gmail.com

https://doi.org/10.2478/jee-2025-0050, Print (till 2015) ISSN 1335-3632, On-line ISSN 1339-309X

Subramanian et al. compared several CNN architectures – specifically VGG16, VGG19, DenseNet201, and InceptionV3 – were compared for the classification of eight retinal pathologies. The authors used a combined dataset from public sources (Kaggle and Open-ICPSR), which was augmented to approximately 24,000 images. The best results were achieved with the VGG16 architecture, reaching a validation accuracy of 97.2%, demonstrating that even less complex models can achieve high classification performance [8].

Karthik and Mahadevappa introduced an improved version of the ResNet architecture, in which the EdgeEn block and the Cross Activation function were added to emphasize retinal layer boundaries and reduce background noise. The model, tested on the OCT2017 dataset (4 classes) [10] and the Retinal OCT C8 dataset (8 classes) [7], achieved an accuracy of 99.1% with ResNet50, representing a significant improvement over the original architecture. In addition to classification results, the study also analyzed interpretability using feature map visualizations [9].

Aykat and Senan compared several modern CNN architectures, such as DenseNet121, EfficientNetV2S, InceptionV3, MobileNet, VGG16, and Xception, aiming to identify the most suitable model for classifying OCT images into eight categories. The best results were obtained with DenseNet121, which achieved an accuracy of 97.79%. The study showed that careful optimization of the training process can yield results comparable to those of more complex models [11].

Jingzhen et al. presented a modern transformer-based approach, where the Swin-Poly Transformer was proposed, combining multi-level feature extraction with the optimized PolyLoss function. The model achieved an accuracy of 99.8% on the OCT2017 dataset [10] and also provided interpretability through the Score-CAM method, which visualizes activation maps based on output scores. This work demonstrates a shift from traditional CNNs toward transformer-based solutions, which offer higher accuracy and improved visualization capabilities [12].

Khalil et al. proposed OCTNet, a model based on the InceptionV3 architecture enhanced with a multi-scale attention mechanism that combines the Channel Attention Module (CAM) and Spatial Attention Module (SAM). This approach suppresses redundant information and focuses on clinically significant regions of the images. OCTNet achieved an accuracy of 99.5% on the OCT2017 [10] and OCTID [14] datasets, with ablation studies confirming the contribution of the proposed attention modules [13].

Overall, the reviewed studies [8, 9, 11-13] demonstrate the high potential of both CNN- and transformer-based models for OCT image classification and highlight the importance of explainability. However, most of them evaluate models exclusively on publicly available datasets and focus primarily on internal validation. Independent external validation on clinical data is rarely performed, which limits the clinical applicability of these approaches.

3 Methodology

This section describes the methodological framework of the proposed approach, which consists of dataset preparation, image preprocessing, model design and training, model evaluation using classification metrics, explainability analysis, and validation on independent external clinical dataset. The overall workflow of the methodology is illustrated in Fig. 1.

3.1 Dataset

The Retinal OCT C8 dataset contains 24,000 high-quality OCT images of the macular region, divided into eight classes of retinal pathologies. The classes include healthy images (NORM*),* choroidal neovascularization (CNV), drusen, diabetic macular oedema (DME), macular hole (MH), central serous chorioretinopathy (CSR), diabetic retinopathy (DR), and age-related macular degeneration (AMD). The dataset is designed to support research and model training for automated classification of macular diseases using machine learning techniques. The images are evenly split into training, testing, and validation subsets, with 2,300 training, 350 testing, and 350 validation images per class [7].

Based on a detailed analysis of the images and consultation with ophthalmologist, the dataset was reduced from eight to six classes. The main reason was the presence of possible clinical overlap between certain diagnoses, which could potentially decrease classification reliability. One possible clinical overlap can be observed between DR and DME class, with preserving DME class, next one is between drusen, CNV and AMD classes, from which we selected drusen and CNV classes to cover dry and exudative AMD classes, respectively. A summary of the final six class dataset structure used in this study is presented in Tab. 1.

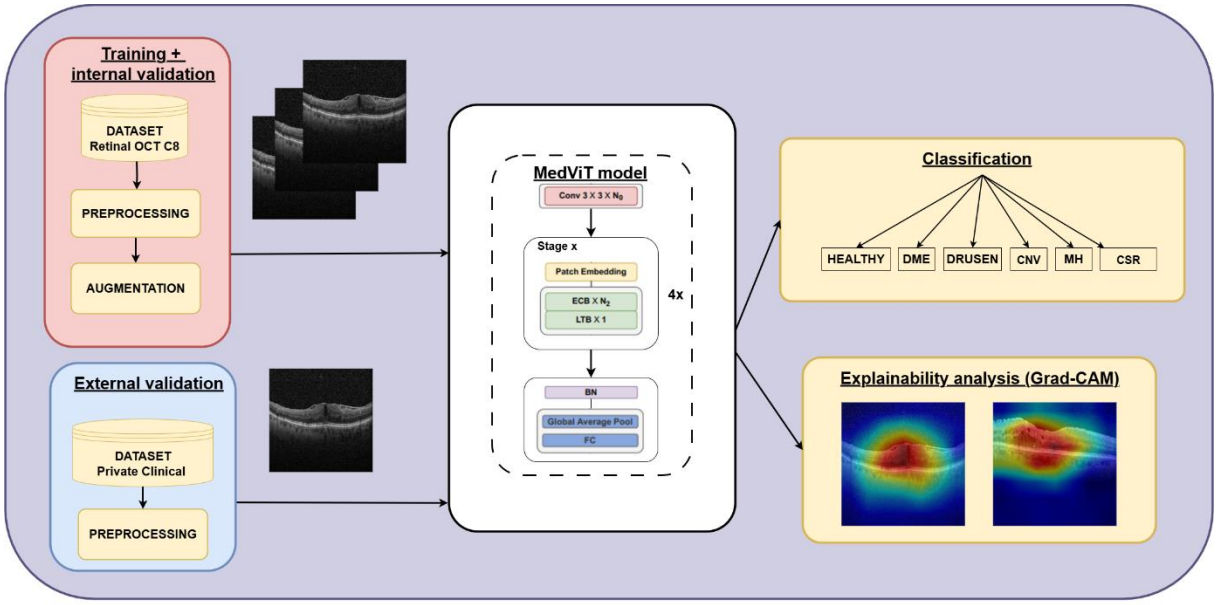

Fig. 1. Diagram of the proposed method

Table 1. Modified retinal OCT C8 dataset [7] structure

Data split	Class*	Total
Train	2 300	13 800
Validation	350	2 100
Test	350	2 100
Total	3 000	18 000

* The same number of images for all classes (CNV, DME, DRUSEN, NORMAL, CSR, MH)

For external independent validation, we used a subset of a private clinical dataset collected within the scientific project DIARET – "Research center for serious diseases and their complications" (project ITMS: 26240120038). This dataset was acquired at the Malacky Hospital and provided by Nemocničná a.s., Holubyho 35, 902 01 Pezinok, Slovakia, for research purposes. We used 10 OCT images of the macular region from this dataset. The dataset contained 5 images diagnosed with diabetic macular edema (DME) and 5 images from healthy eyes. These images were collected from routine clinical practice and are not used during the training or internal validation phases.

3.2 Image preprocessing and augmentation

The first step prior to model implementation was a thorough dataset analysis, which revealed that the images were of varying sizes and resolutions and exhibited a high level of noise. Therefore, several preprocessing techniques were applied to standardize image dimensions and reduce noise while preserving clinically relevant details. Image resizing was performed using bilinear interpolation [15].

To reduce noise, multiple denoising approaches were evaluated, including bilateral filtering, non-local means denoising, Gaussian blurring, and median blurring. The best classification performance was achieved using the Non-local Means Denoising algorithm, which, unlike traditional local filters, leverages the redundancy of similar patterns across the entire image. This approach effectively reduces noise while preserving textures and fine structural details [16].

An example of the preprocessing process is shown in Fig. 2. The original image (a) contains significant noise, particularly in darker regions, while the processed image (b) shows these regions significantly smoothed, with the key retinal layers preserved and clearly distinguishable.

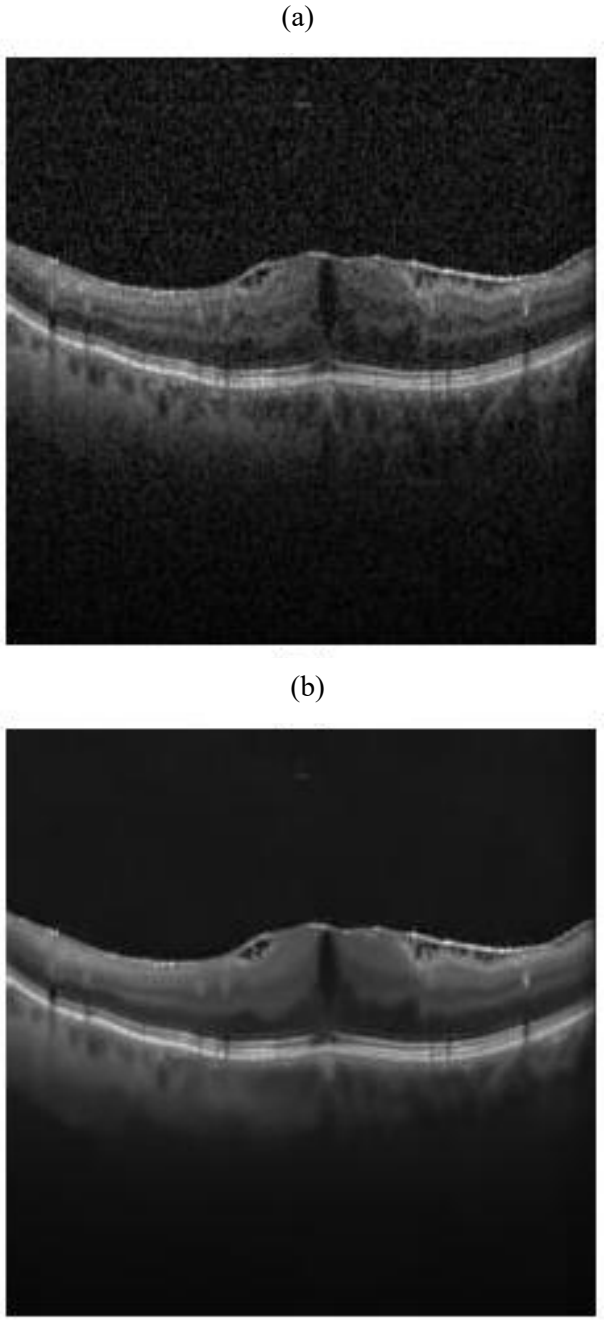

Fig. 2. Preprocessing of the image: (a) original image, (b) image after Non-local Means denoising

To further enhance the robustness of the model and reduce the risk of overfitting, data augmentation was applied to the training set. Advanced augmentation method based on the random combination of multiple simple transformation operations (rotation, contrast adjustment, translation, scaling) was used. The resulting transformations are combined through a convex combination, generating new image variants with high diversity while maintaining their essential features [17]. Additionally, novel augmentation technique known as the Patch Momentum Changer (PMC) [3] specifically integrated in our implemented deep learning model was used, leading to improved robustness and generalization capabilities by combining feature normalization with token-level data augmentation.

3.3 MedViT model architecture

MedViT is a robust hybrid CNN-Transformer model designed for the classification of medical images across various domains. It combines the strengths of convolutional neural networks (CNNs) and vision transformers (ViTs) to address key challenges in medical image analysis, such as modeling long-range dependencies and improving robustness to image variability [3].

The core components of the architecture are the Efficient Convolution Block (ECB), responsible for capturing local features, and the Local Transformer Block (LTB), designed to extract multi-frequency signals. The integration of these two modules enables enhanced extraction of both global and local features. This architecture has demonstrated high accuracy and efficiency on large-scale medical imaging datasets such as MedMNIST-2D [3].

MedMNIST-2D is a comprehensive collection of normalized and preprocessed 2D biomedical images from diverse sources, including X-ray, ultrasound, and CT scans. It is specifically designed for classification tasks and serves as a benchmark dataset for comparing machine learning algorithms in medical image analysis [18].

The MedViT model is available in three versions based on the number of trainable parameters: MedViT-T (Tiny, 10.8M), MedViT-S (Small, 23.6M), and MedViT-L (Large, 45.8M) [3]. In this work, we employed the medium-sized variant, MedViT-S, referred to as MedViT_base in available implementations.

3.4 Model training and evaluation

The MedViT model was initialized with weights pretrained on the ImageNet-1K dataset [19]. Training was conducted for up to 100 epochs using the AdamW optimizer, the cross-entropy loss function, and a batch size of 32. The initial learning rate was set to 1×10^{-4} and subsequently adjusted using the CosineAnnealingLR scheduler. To prevent overfitting, the early stopping technique with a patience of 7 epochs was employed, and the best model checkpoint was selected based on the validation performance. A complete overview of the training parameters is provided in Tab. 2.

Table 2. Training parameters

Parameter	Value
Model	MedViT-S (MedViT_base)
Model weights	ImageNet-1K
Optimizer	AdamW
Learning rate (init.)	0.0001
Learning rate scheduler	CosineAnnealingLR (100 epochs)
Loss function	Cross-entropy
Early stopping	Enabled (patience = 7)
Batch size	32
Max. epochs	100

During training, the evolution of accuracy and loss was continuously monitored on both the training and validation sets. After training, the model was evaluated on the test set, where overall accuracy and loss were recorded. To obtain a more detailed assessment of model performance, we employed standard classification metrics, precision, recall, F1-score, and ROC-AUC, complemented by visualizations. These included a confusion matrix, which provides a detailed overview of correctly and incorrectly classified samples across all classes, and ROC curves illustrating the model's discriminative ability. The combination of these metrics and visualizations enabled a comprehensive evaluation of the model's performance in the multi-class classification of OCT images.

3.5 Explainability analysis

To analyze the explainability of the trained models, we utilized a complementary library for the PyTorch framework that provides state-of-the-art interpretability methods in the field of computer vision. This library supports a wide range of architectures, from traditional convolutional neural networks (CNNs) to vision transformer models, and is applicable to various tasks such as classification, semantic segmentation, and object detection [20].

In our work, we employed the Grad-CAM (Gradient-weighted Class Activation Mapping) technique, which visualizes weighted 2D activation maps of convolutional layers based on the average values of their gradients. In the context of vision transformer models, these activations can be interpreted as weighted representations of individual patches – the image segments into which the input image is divided [20].

The Grad-CAM outputs were visualized as heatmaps, where the color scale ranged from dark blue (indicating the lowest attention) to dark red (indicating the highest attention). These visualizations were subsequently reviewed and discussed with an ophthalmologist to assess the clinical relevance of the highlighted regions.

3.6 External validation

To evaluate the model's ability to generalize to unseen data, we used part of an external clinical private dataset from DIARET project consisting of 10 OCT images of the macular region. The dataset included 5 images diagnosed with diabetic macular edema (DME) and 5 images of healthy eyes. These images were obtained from clinical practice and were not part of the training or validation process.

Preprocessing of the clinical images followed the same procedure as applied to the Retinal OCT C8 dataset [7] to ensure input consistency. The pre-trained MedViT model was then evaluated on this dataset without any additional fine-tuning.

The evaluation was based on the same set of metrics and visualization techniques used in the internal evaluation. Additionally, Grad-CAM-based visualizations were generated for the external dataset to assess whether the model focused on clinically relevant regions during classification.

4 Results and analysis

This section presents the experimental results corresponding to the methodology described in Section 3. We first report the training and evaluation results on the Retinal OCT C8 dataset [7], including quantitative metrics and performance visualizations. Next, we focus on the explainability analysis using the Grad-CAM method, which enables the assessment of the clinical relevance of the model's decision-making process. Finally, we present the results of external validation on an independent clinical dataset, providing insights into the model's generalization capability in real-world scenarios.

4.1 Training and evaluation results

The MedViT model was trained for up to 100 epochs using the parameters described in Section 3.4. The evolution of accuracy and loss on the training and validation sets is shown in Fig. 3. The curves demonstrate a stable increase in accuracy and a decrease in loss, indicating good model convergence and the absence of significant overfitting.

After training, the model achieved a test accuracy of 95.95% and an F1-score of 0.9595, confirming a well-balanced performance between precision and recall. Detailed results, including additional metrics such as precision and recall, are presented in Tab. 3.

The confusion matrix shown in Fig. 4 (a) provides a detailed view of the model's classification performance across individual classes. High accuracy was observed across all categories, with the lowest classification rate recorded for the DRUSEN class, where misclassifications primarily occurred with CNV and CSR.

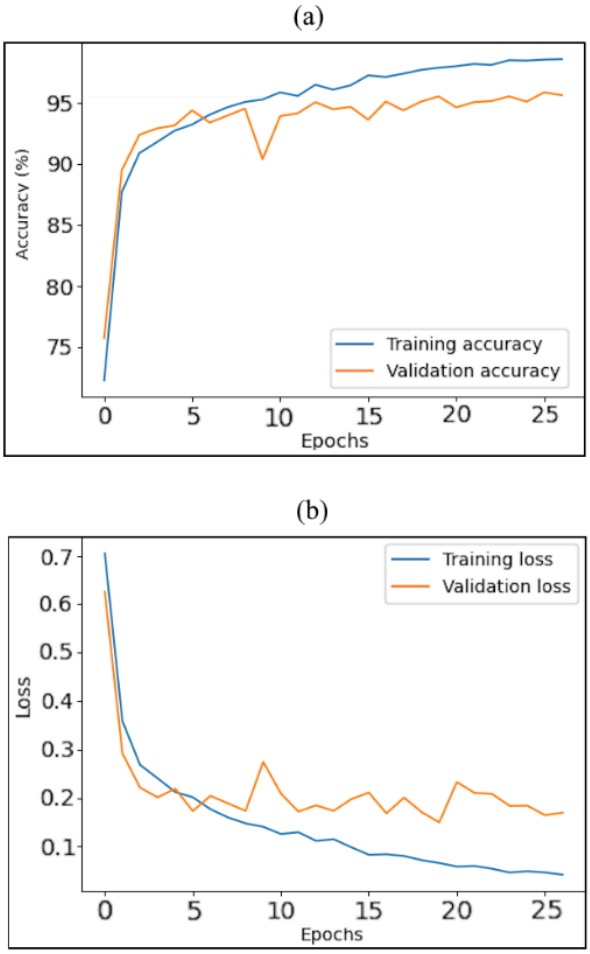

Fig. 3. Training and validation accuracy (a), and loss curves (b)

Table 3. Training and evaluation results

Metrics	Training accuracy	Validation accuracy	Test accuracy	Precision	Recall	F1 score
Value	98.56	95.62	95.95	0.96	0.96	0.96

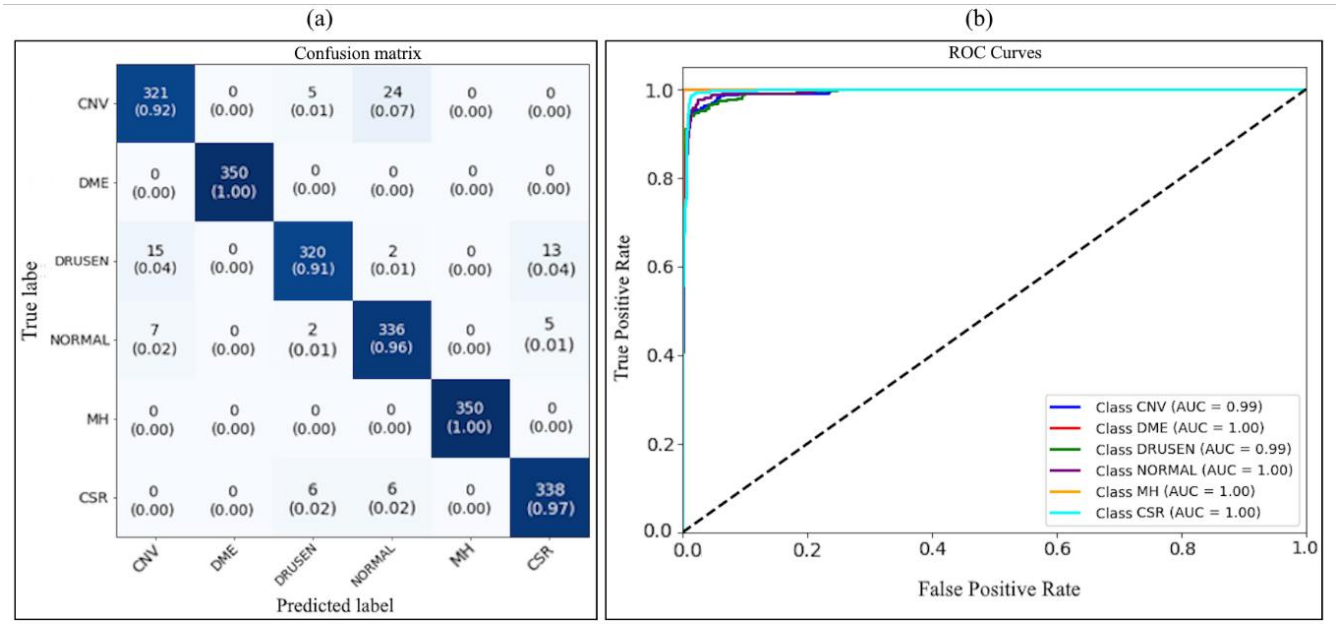

Fig. 4. Evaluation confusion matrix (a), and ROC curves with ROC-AUC metrics (b) of modified Retinal OCT C8 dataset [7]

Additional insights into the model's discriminative capability are provided by the ROC curves shown in Fig. 4 (b). The AUC values ranged from 0.99 to 1.00 across all classes, indicating the model's excellent ability to distinguish between different disease categories.

4.2 Explainability analysis results

For the explainability analysis, we employed the Grad-CAM technique, which allows visualization of the image regions that the model focuses on most during decision-making. The results are presented as heatmaps, where dark blue indicates the lowest level of attention and dark red represents the highest.

As shown in Fig. 5, the model consistently concentrated its attention on the central macular region, which is the most clinically relevant area for diagnosis. This finding confirms that the model not only achieves high classification accuracy but also bases its decisions on image regions that are critical for ophthalmologists during diagnostic assessment. Such alignment between the model's visual focus and clinical practice enhances its credibility and potential as a decision-support tool in ophthalmic diagnostics.

4.3 External validation results

To assess the model's generalization capability on clinical data, we used an external independent private dataset consisting of 10 OCT images of the macular region (5 with DME and 5 from healthy eyes). The pre-trained MedViT model was evaluated on this dataset without any additional fine-tuning.

The results showed that the model achieved an overall accuracy of 90%, correctly classifying 9 out of 10 samples. Detailed performance metrics are presented in Tab. 4. The model correctly identified all 5 DME cases, with a single misclassification occurring in the healthy eye category.

In addition to quantitative metrics, we analyzed Grad-CAM visualizations (Fig. 5), which revealed that the model consistently focused on the central macular region during decision-making. This behavior further supports the interpretability of the model and underscores its potential for practical clinical use in ophthalmological diagnostics.

Table 4. External validation results

Metrics	Test accuracy	Test loss	Precision	Recall	F1 score
Value	90.00	0.44	0.92	0.90	0.90

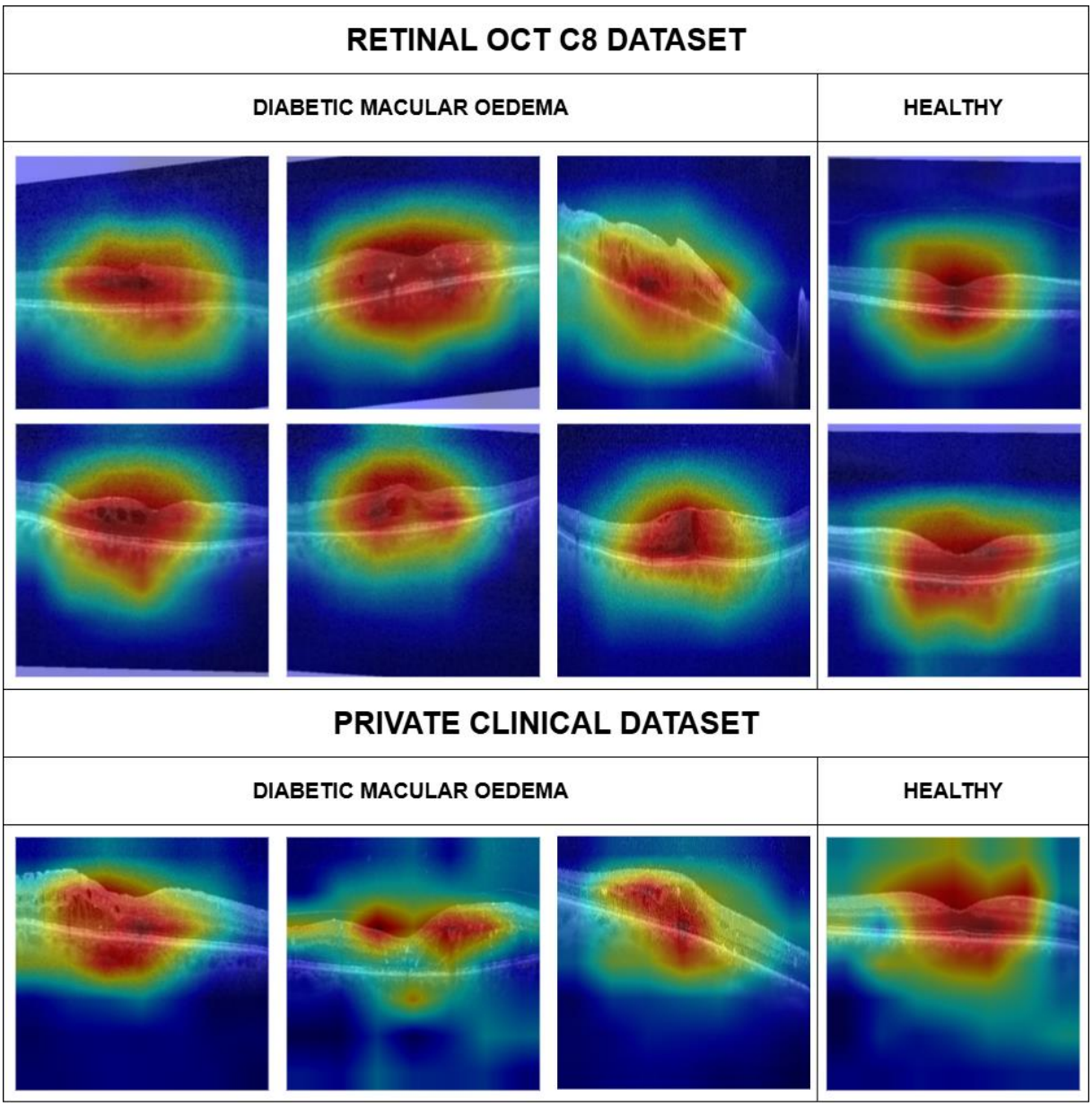

Fig. 5. GRAD-CAM heatmap visualizations of Retinal OCT C8 dataset [7] and external private clinical dataset

5 Conclusion

In this study, we presented a methodology for classifying OCT images of the macular region using the hybrid CNN-Transformer model MedViT, with a particular emphasis on explainability and clinical interpretability. The model, trained on the publicly available Retinal OCT C8 dataset [7], achieved a high classification accuracy of 95.95% and demonstrated balanced performance across all classes, confirming its ability to effectively process medical imaging data.

A key contribution of this work is the integration of explainability analysis through Grad-CAM, which revealed that the model consistently focused on the diagnostically relevant central macular region during decision-making. This insight provides essential context for the clinical interpretation of model outputs. Moreover, external validation on a private independent clinical external dataset confirmed the model's ability to generalize beyond the training data, achieving 90% accuracy even with a limited number of samples.

The obtained results indicate that the proposed approach has strong potential to serve as a reliable decision-support tool for ophthalmologists in the diagnosis of macular diseases, contributing to more efficient and rapid clinical workflows. Future work will focus on evaluating the model on larger clinical datasets, extending it to additional diagnostic categories, and integrating the system into real-world clinical environments.

Acknowledgement

This work was done within the VEGA grant VEGA 1/0202/23 AIDabiomeDIA.

The image dataset created within the scientific project DIARET – "Research center for serious diseases and their complications (project ITMS: 26240120038)" was provided to us by Nemocničná a.s., Holubyho 35, 902 01 Pezinok, Slovakia, for research purposes.

References

[1] A. H. Kashani, C. Chen, J. K. Gahm, F. Zheng, G. M. Richter, P. J. Rosenfeld, Y. Shi, and R. K. Wang, "Optical coherence tomography angiography: A comprehensive review of current methods and clinical applications," *Progress in Retinal and Eye Research*, vol. 60, pp. 66–100, 2017. doi:10.1016/j.preteyeres.2017.07.002.

[2] A. P. Voigt, N. K. Mullin, S. S. Whitmore, A. P. DeLuca, E. R. Burnight, X. Liu, B. A. Tucker, T. E. Scheetz, E. M. Stone, and R. F. Mullins, "Human photoreceptor cells from different macular subregions have distinct transcriptional profiles," *Human Molecular Genetics*, vol. 30, no. 16, pp. 1543–1558, 2021. doi:10.1093/hmg/ddab140.

[3] O. N. Manzari, H. Ahmadabadi, H. Kashiani, S. B. Shokouhi, and A. Ayatollahi, "MedViT: A robust vision transformer for generalized medical image classification," *Computers in Biology and Medicine*, vol. 157, p. 106791, 2023. doi:10.1016/j.compbiomed.2023.106791.

[4] A. Chaddad, J. Peng, J. Xu, and A. Bouridane, "Survey of Explainable AI Techniques in Healthcare," Sensors, vol. 23, no. 2, p. 634, 2023. doi: 10.3390/s23020634.

[5] Z. Sadeghi, R. Alizadehsani, M. A. CIFCI, S. Kausar, R. Rehman, P. Mahanta, P. K. Bora, A. Almasri, R. S. Alkhawaldeh, S. Hussain, B. Alatas, A. Shoeibi, H. Moosaei, M. Hladík, S. Nahavandi, and P. M. Pardalos, "A review of Explainable Artificial Intelligence in healthcare," *Comput. Electr. Eng.*, vol. 118, p. 109370, Aug. 2024. doi: 10.1016/j.compeleceng.2024.109370.

[6] T. F. Tan, P. Dai, X. Zhang, L. Jin, S. Poh, D. Hong, J. Lim, G. Lim, Z. L. Teo, N. Liu, and D. S. W. Ting, "Explainable artificial intelligence in ophthalmology," *Current Opinion in Ophthalmology*, vol. 34, no. 5, pp. 422–430, 2023. doi:10.1097/ICU.0000000000000983.

[7] O. S. Naren, "Retinal OCT Image Classification – C8," Kaggle, 2024. doi:10.34740/KAGGLE/DSV/9595300.

[8] M. Subramanian, K. Shanmugavadivel, O. S. Naren, K. Premkumar, and K. Rankish, "Classification of retinal OCT images using deep learning," in *Proc. 2022 Int. Conf. Computer Communication and Informatics (ICCCI)*, 2022, pp. 1–7. doi:10.1109/ICCCI54379.2022.9740985.

[9] K. Karthik and M. Mahadevappa, "Convolution neural networks for optical coherence tomography (OCT) image classification," *Biomedical Signal Processing and Control*, vol. 79, p. 104176, 2023. doi:10.1016/j.bspc.2022.104176.

[10] D. S. Kermany *et al*., "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131.e9, 2018. doi:10.1016/j.cell.2018.02.010.

[11] Ş. Aykat and S. Senan, "Using machine learning to detect different eye diseases from OCT images," *International Journal of Computational and Experimental Science and Engineering*, vol. 9, pp. 62–67, 2023. doi:10.22399/ijcesen.1297655.

[12] J. Jingzhen *et al*., "An interpretable transformer network for the retinal disease classification using optical coherence tomography," *Scientific Reports*, vol. 13, no. 1, p. 3637, 2023. doi:10.1038/s41598-023-30853-z.

[13] I. Khalil, A. Mehmood, H. Kim, and J. Kim, "OCTNet: A modified multi-scale attention feature fusion network with InceptionV3 for retinal OCT image classification," *Mathematics*, vol. 12, no. 19, p. 3003, 2024. doi:10.3390/math12193003.

[14] P. Gholami, P. Roy, M. K. Parthasarathy, and V. Lakshminarayanan, "OCTID: Optical coherence tomography image database," *Computers & Electrical Engineering*, vol. 81, p. 106532, 2020. doi:10.1016/j.compeleceng.2019.106532.

[15] A. C. Bovik, "Basic gray-level image processing," in *Handbook of Image and Video Processing*, 2nd ed., A. C. Bovik, Ed. Burlington, MA, USA: Academic Press, 2005, pp. 21–37. doi:10.1016/B978-012119792-6/50066-8.

[16] A. Buades, B. Coll, and J.-M. Morel, "Non-local means denoising," *Image Processing On Line*, vol. 1, pp. 208–212, 2011. doi:10.5201/ipol.2011.bcm_nlm.

[17] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "AugMix: A simple data processing method to improve robustness and uncertainty," *arXiv* preprint arXiv:1912.02781, 2020. [Online]. Available: https://arxiv.org/abs/1912.02781.

[18] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, and B. Ni, "MedMNIST v2 – A large-scale lightweight benchmark for 2D and 3D biomedical image classification," *Scientific Data*, vol. 10, no. 1, p. 41, Jan. 2023. doi:10.1038/s41597-022-01721-8.

[19] J. Deng *et al*., "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255. doi:10.1109/CVPR.2009.5206848.

[20] J. Gildenblat and contributors, "PyTorch library for CAM methods," GitHub repository, 2021. [Online]. Available: https://github.com/jacobgil/pytorch-grad-cam.

Received 21 July 2025