CAM-HR: Clinical Attention Enhanced CNNs for Hypertensive Retinopathy Classification

Mustafa Yurdakul , Süleyman Burçin Şüyun * , Şakir Taşdemir

Posted Date: 20 March 2026

doi: 10.20944/preprints202603.1589.v1

Keywords: hypertensive retinopathy; OCT imaging; deep learning; convolutional neural networks; attention mechanism; medical image classification

 doi:10.20944/preprints202603.1589.v1

Article

CAM-HR: Clinical Attention Enhanced CNNs for Hypertensive Retinopathy Classification

Mustafa Yurdakul [1], Süleyman Burçin Şüyun [2,*] and Şakir Taşdemir [3]

[1] Kırıkkale University, Computer Engineering Departmant, Kırıkkale, Turkiye
[2] Sinop University, The Rectorate of Sinop University, Digital Transformation Office, Sinop, Türkiye
[3] Selçuk University, Computer Engineering Departmant, Konya, Türkiye
* Correpondece: suyun@sinop.edu.tr

Abstract

Hypertensive retinopathy (HR) is a retinal vascular disorder caused by long-term hypertension and can lead to severe visual impairment. If not detected early, it could progress to irreversible visual impairment and even blindness. Recent advances in deep learning have enabled automated analysis of retinal images to support clinical diagnosis. In this study, we propose a Clinical Attention Module–enhanced convolutional neural network framework (CAM-HR) for automatic classification of HR stages from Optical Coherence Tomography (OCT) images. In the initial scenario, various state-of-the-art architectures of convolutional neural networks (CNN) have been utilized as baseline models. In the next scenario, the Clinical Attention Module (CAM) is utilized with these architectures to focus on clinically significant regions of the retina, such as vascular structures and lesion locations. The models are evaluated using accuracy, precision, recall, F1-score, and Cohen's kappa metrics. Experimental results demonstrate that the proposed CAM module consistently improves classification performance across different backbone architectures, achieving the best performance with the ConvNeXt + CAM model. These findings indicate that clinically guided attention mechanisms can significantly enhance automated HR diagnosis from OCT images.

Keywords: hypertensive retinopathy; OCT imaging; deep learning; convolutional neural networks; attention mechanism; medical image classification

1. Introduction

HR is a retinal vascular disorder caused by prolonged systemic hypertension and is considered an important indicator of cardiovascular risk and microvascular damage[1,2]. Persistent high blood pressure can lead to structural changes in retinal vessels, including arteriolar narrowing, hemorrhages, exudates, and cotton wool spots. If left untreated, HR may result in significant visual impairment and may also indicate the presence of systemic complications such as stroke or heart disease[2].

The early detection and accurate staging of HR are critical in the management of the disease and in the prevention of significant visual impairment [3,4]. Traditionally, ophthalmologists use manual interpretation of images obtained using imaging modalities such as fundus cameras and Optical Coherence Tomography (OCT) in the diagnosis of HR. Manual interpretation is often subjective and time-consuming, especially in situations where large numbers of images have to be analyzed.

Recent advances in AI and DL have significantly improved automated medical image analysis[5–8]. Convolutional Neural Networks (CNNs) have demonstrated remarkable performance in a variety of medical image-related applications, such as detection of retinal diseases, classification of diabetic retinopathy (DR), identification of glaucoma, and macular degeneration [9]. CNNs are capable of learning features automatically from images, enabling them to detect patterns related to disease progression.

 doi:10.20944/preprints202603.1589.v1

Several studies have been conducted on the application of deep learning methods in the detection and assessment of heart rate (HR) using various methodologies. As shown in Table 1, various methods were used in the improvement of feature extraction and accuracy of classification. Bhimavarapu et al. [10] presented a hybrid approach based on CNN feature extraction and a multi-class SVM classifier, achieving high accuracy in vessel classification with hand-crafted features and a multi-stage approach.Triwijoyo et al. [11] utilized a deep convolutional neural network combined with vessel segmentation and arterio-venous ratio (AVR) analysis to detect HR severity. However, their approach requires extensive preprocessing and n, Suman et al. [15] proposed a hybrid architecture combining ResNet-50 with a Vision Transformer to capture both local spatial features and global contextual information. Zhu [16] conducted a comparative study of CNN, Vision Transformer, and AutoML based approaches and highlighted the importance of data augmentation strategies, particularly when training transformer-based models on relatively small medical datasets.

Despite all these advances, conventional CNN architectures often treat all regions of the image equally and may not sufficiently emphasize clinically relevant structures in the image. In retinal images, pathological structures such as vascular constrictions, hemorrhages, and exudates often appear in localized regions of the image. Models that do not specifically pay attention to these localized regions may suffer from reduced accuracy. Attention mechanisms were proposed to overcome this limitation by focusing the neural networks on the most relevant regions of the image.morphological operations. Wang [12] proposed the MA-DNet, which is based on the DenseNet model but also includes channel and spatial attention mechanisms for improving the representation of features, but the study faces the problem of class imbalance and the lack of sufficient HR samples.

Other studies have focused on transfer learning and hybrid architectures. Kumar et al. [13] employed transfer learning models combined with contour-based morphological descriptors to improve interpretability and classification accuracy. Silva-Rodriguez et al. [14] investigated the use of the FLAIR retinal foundation model pretrained on a large scale fundus dataset, demonstrating performance improvements when fine-tuned for retinal disease tasks.[15] proposed a hybrid architecture combining ResNet-50 with a Vision Transformer to capture both local spatial features and global contextual information. Zhu [16] conducted a comparative study of CNN, Vision Transformer, and AutoML based approaches and highlighted the importance of data augmentation strategies, particularly when training transformer-based models on relatively small medical datasets.

Despite these advances, conventional CNN architectures often treat all spatial regions of an image equally and may fail to emphasize clinically significant retinal structures. In retinal imaging, pathological features such as vascular narrowing, hemorrhages, and exudates frequently appear in localized regions. Models that do not explicitly focus on these regions may therefore experience reduced diagnostic performance. Attention mechanisms have been introduced to address this limitation by guiding neural networks to focus on the most informative areas of an image.

Table 1. Summary of recent deep learning–based studies on HR detection and classification.

Study	Method	Strengths	Results	Limitations
Bhimavarapu et al., [10]	Improved CNN-based feature extraction combined with multi-class SVM classifier	Hybrid architecture extracts vascular features and reduces computational complexity by running convolution only once on the fundus image	98.99% accuracy in vessel classification on augmented dataset of 1200 fundus images	Depends on handcrafted vessel features and multi-stage pipeline

 doi:10.20944/preprints202603.1589.v1

Triwijoyo et al., [11]	Deep Convolutional Neural Network (DCNN) with vessel segmentation and AVR feature analysis	Detects HR severity and extracts retinal vessels and lesions (exudates, hemorrhages, cotton-wool spots) automatically	Accuracy 90%, Recall 81.82%, F-Score 90%	Requires vessel segmentation and morphological preprocessing
Wang [12]	MA-DNet (DenseNet121 + channel & spatial attention)	Lightweight architecture improves feature representation and reduces computational complexity	95.8% classification accuracy on OIA-ODIR dataset	Severe class imbalance in dataset and limited HR samples
Kumar et al., [13]	Transfer learning models combined with contour-based morphological descriptors (EfficientNetB4, InceptionResNetV2 etc.)	Integration of morphological features improves interpretability and classification performance	96.61% validation accuracy with EfficientNetB4	Performance affected by dataset diversity and multi-disease fundus datasets
Silva-Rodriguez et al., [14]	FLAIR retinal foundation model + transfer learning (LP / fine-tuning)	Uses domain-specific foundation model pretrained on 286k fundus images	Transfer learning from FLAIR improved performance by between 2.5–4% compared with ImageNet initialization	Transferability still limited due to small HR datasets
Suman et al., [15]	Hybrid ResNet-50 + Vision Transformer architecture	Combines CNN spatial features with global attention mechanism for HR severity grading	Improved grading performance on HR dataset	High computational cost and dependence on dataset size
Zhu [16]	Comparative evaluation of CNN, Vision Transformer, and AutoML models	Provides architectural analysis showing augmentation benefits for transformer models	Demonstrated strong dependence between architecture and data augmentation	Large transformer models perform poorly on small HR datasets

In this study, we propose a CAM integrated with multiple CNN architectures to enhance HR classification from OCT images. The proposed module aims to highlight clinically significant retinal structures and improve feature representation within the network. The framework is evaluated across two experimental scenarios: baseline CNN architectures and CNN models enhanced with the proposed attention module.

 doi:10.20944/preprints202603.1589.v1

The contributions of this study can be summarized as follows:

- A deep learning framework for automatic classification of HR stages using OCT images
- Integration of a CAM to enhance clinically relevant feature extraction
- Comprehensive evaluation of multiple CNN architectures with and without the proposed attention mechanism
- Demonstration that the attention-enhanced models consistently improve classification performance

The remainder of this paper is organized as follows. Section 2 describes the dataset, baseline CNN architectures, and the proposed Clinical Attention Module. Section 3 presents the experimental setup and evaluation metrics. Section 4 reports the experimental results. Section 5 discusses the findings and limitations of the study. Finally, Section 6 concludes the paper.

2. Materials and Methods

The framework of the proposed classification system for hypertensive retinopathy (HR) is depicted in Figure 1. As can be seen from the figure, the workflow of the proposed system commences with optical coherence tomography (OCT) images of the retina, which are the input to the deep learning system. These images are then processed using various state-of-the-art CNN architectures to evaluate the efficacy of the CNN models in the classification of different stages of HR.

As depicted in Figure 1, the proposed framework is based on two configurations. In the first configuration, the input OCT images are directly passed to the baseline CNN models, which include InceptionNeXt [17], CSPNeXt [18], MobileNetV3 [19], ShuffleNetV2 [20], ResNeSt50 [21], RegNetY-400MF [22], EfficientNetV2-S [23], and ConvNeXt [24]. These CNN models are deployed without the attention mechanism and are considered the baseline models.

In the second configuration of the proposed framework, the same CNN models are extended with the proposed channel attention module (CAM). The proposed CAM is incorporated with the CNN models to enable the network to focus on the relevant areas of the retina that are likely to exhibit abnormal patterns associated with HR.

After the classification task is complete, the CNN models make predictions regarding the four diagnostic classes: Healthy, Stage 1, Stage 2, and Stage 3 HR.

 doi:10.20944/preprints202603.1589.v1

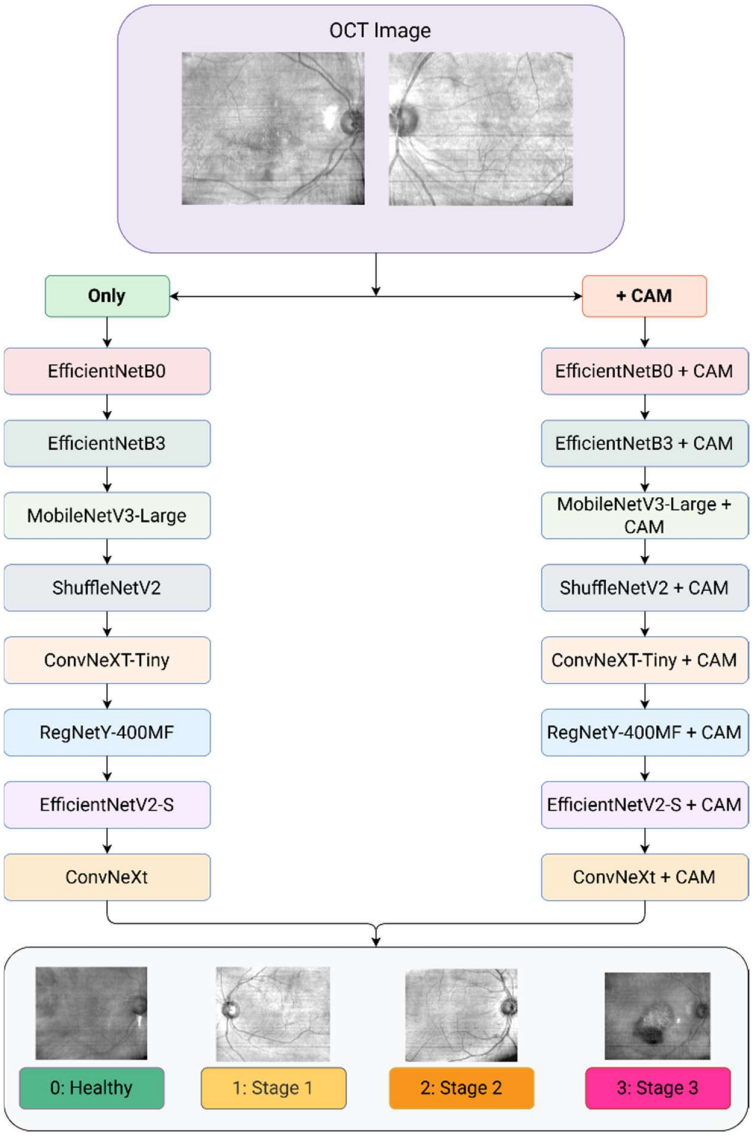

Figure 1. Illustration of the experimental framework showing baseline CNN architectures and CAM-enhanced models for OCT-based HR classification.

2.1. Dataset Description

The dataset used in this study consists of OCT images collected from patients diagnosed with HR in Turkey. All patient-related information was anonymized prior to analysis to ensure compliance with ethical and privacy regulations. The OCT images were obtained through clinical examinations and subsequently annotated by an experienced ophthalmologist. During the annotation process, each image was labeled according to established clinical criteria that describe the different stages of HR.

 doi:10.20944/preprints202603.1589.v1

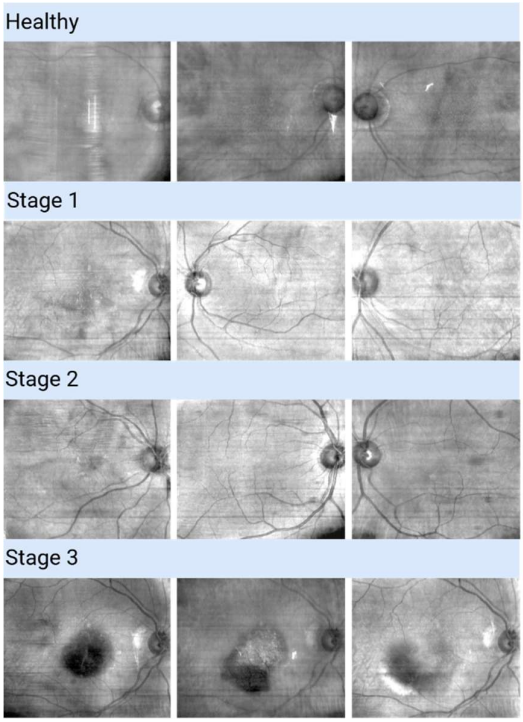

Figure 2. Representative OCT images of HR stages in the dataset, including Healthy, Stage 1, Stage 2, and Stage 3.

In total, the dataset includes 1875 OCT images with an original resolution of 512 × 512 pixels. Each image represents one of four diagnostic categories corresponding to the progression stages of the disease: Healthy, Stage 1, Stage 2, and Stage 3 HR. To ensure compatibility with the deep learning architectures used in this study, all images were resized to 224 × 224 pixels before being used as input to the models.

Table 2. Clinical description of HR stages.

Category	Identification
Normal	No hypertension-related retinal abnormalities.
Stage 1	Mild arterial narrowing and vessel wall thickening.
Stage 2	Pronounced vasoconstriction, arteriovenous crossing, and atherosclerotic changes.
Stage 3	Severe vascular damage with hemorrhages, exudates, and cotton wool spots.

Representative OCT samples corresponding to each disease stage are illustrated in Figure 3. These examples highlight the morphological differences observed in retinal structures across the progression of HR and provide visual insight into the patterns that the proposed models aim to learn. The clinical characteristics defining each category are summarized in Table 2. The dataset was organized to maintain a relatively balanced distribution among classes, which helps improve model generalization and reduces potential bias during the training process.

 doi:10.20944/preprints202603.1589.v1

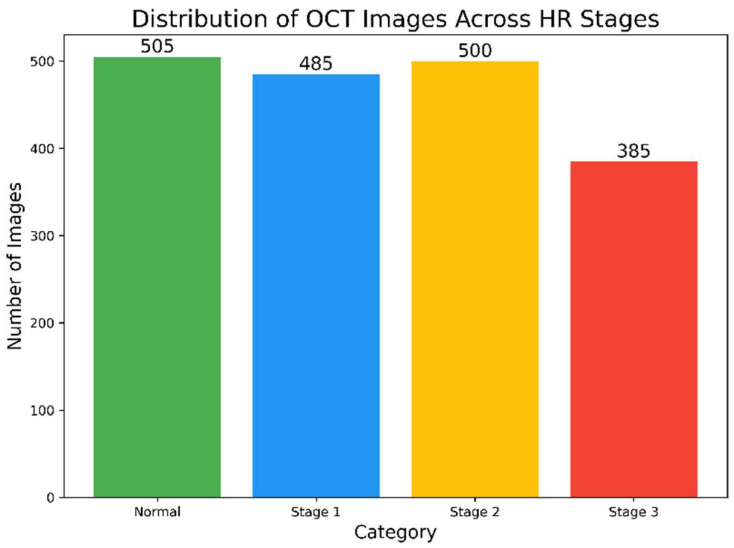

Figure 3. Distribution of OCT images across HR categories in the dataset.

The dataset is divided in a balanced way to enable more efficient training of CNN models. In this context, 80% of the dataset is used for training, and the remaining 20% is used for testing and validation purposes. During training, images were randomly rotated by up to 30%, horizontally and vertically shifted by up to 40% of their original size, and diversified by up to 30% shear and zoom transformations to augment the data.

2.2. Baseline CNN Architectures

In this study, several state-of-the-art convolutional neural network (CNN) architectures were employed as baseline models to evaluate the effectiveness of the proposed framework. These architectures have demonstrated strong performance in image classification tasks and have been widely adopted in medical image analysis due to their ability to learn hierarchical feature representations from visual data.

The selected baseline models include InceptionNeXt [17], CSPNeXt [18], MobileNetV3 [19], ShuffleNetV2 [20], ResNeSt50 [21], RegNetY-400MF [22], EfficientNetV2-S [23], and ConvNeXt [24]. These models represent a diverse set of architectural designs, ranging from lightweight and mobile-friendly networks to high-capacity models optimized for performance.

InceptionNeXt [17] integrates the multi-branch design philosophy of Inception architectures with modern convolutional design principles inspired by ConvNeXt, enabling efficient multi-scale feature extraction. CSPNeXt [18], based on Cross Stage Partial (CSP) connections, enhances gradient flow and reduces computational cost while maintaining strong representational capability. MobileNetV3 [19] and ShuffleNetV2 [20] are lightweight architectures specifically designed for resource-constrained environments, utilizing depthwise separable convolutions and channel shuffling mechanisms to improve efficiency.

ResNeSt50 [21] introduces split-attention blocks that enable adaptive feature-map attention within residual networks, improving feature representation by modeling channel-wise dependencies. RegNetY-400MF [22] is part of the RegNet family, which focuses on designing network architectures through structured design spaces, achieving a balance between efficiency and accuracy. EfficientNetV2-S [23] improves upon previous EfficientNet models by optimizing both training speed and parameter efficiency through compound scaling and progressive learning strategies. Finally, ConvNeXt [24] represents a modernized CNN architecture inspired by transformer design principles, incorporating large kernel sizes, layer normalization, and inverted bottlenecks to achieve superior performance.

 doi:10.20944/preprints202603.1589.v1

These baseline models are used without any attention mechanism in the first experimental setup to establish a performance benchmark. This allows for a fair comparison with the proposed CAM, enabling a clear assessment of its contribution to classification performance.

2.3. Clinical Attention Module

Retinal imaging analysis for HR diagnosis requires the identification of subtle pathological patterns such as vessel narrowing, lesion formation, and structural retinal changes. Conventional CNNs are effective at extracting hierarchical features from images; however, they often treat all spatial regions equally and may fail to focus on clinically significant areas[25–27]. In medical imaging tasks, this limitation can reduce the model's ability to accurately capture disease-specific patterns. To address this issue, this study proposes a CAM designed to enhance the discriminative capability of CNN-based architectures by emphasizing clinically meaningful retinal regions. The proposed module integrates multiple attention mechanisms that guide the network to concentrate on relevant structures associated with HR progression.

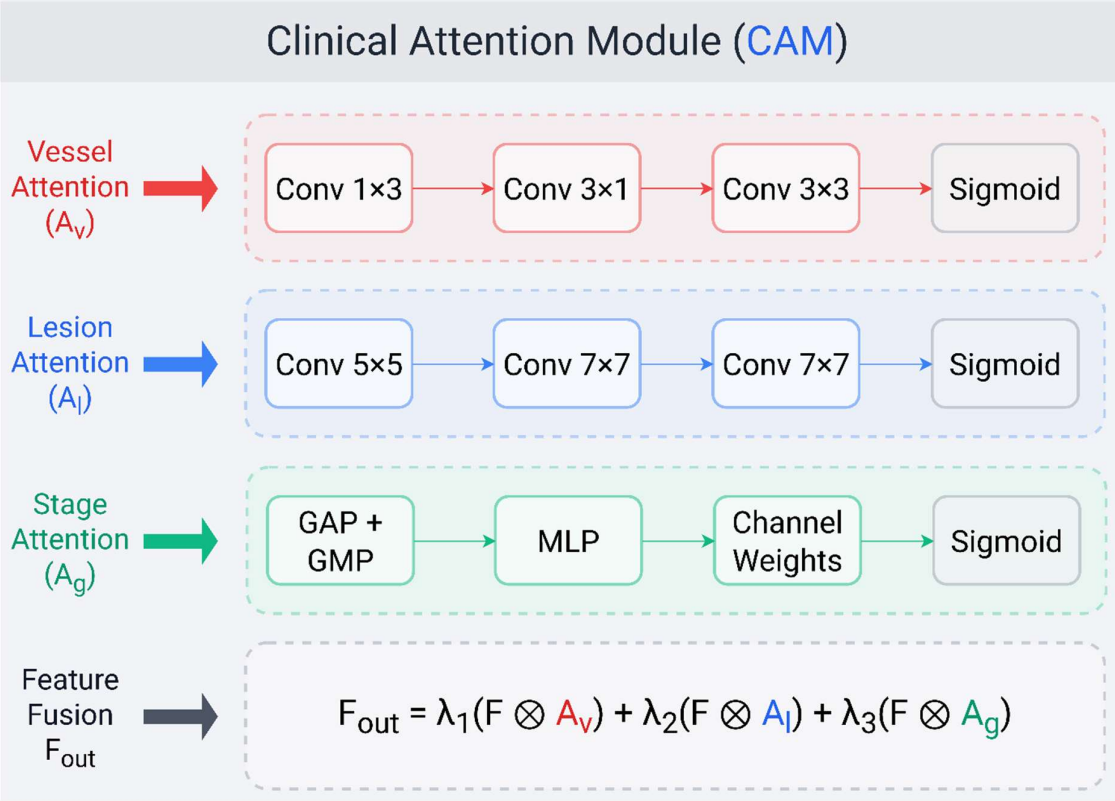

Figure 4. Architecture of the proposed Clinical Attention Module, consisting of vessel attention, lesion attention, and stage attention branches whose outputs are fused to produce the final refined feature representation.

The CAM module consists of three complementary attention branches: vessel attention, lesion attention, and stage attention. Each branch captures different pathological characteristics of HR and contributes to a more informative feature representation. Given an input feature map $F \in \mathbb{R}^{C\times H\times W}$, extracted from the backbone CNN architecture, the vessel attention branch focuses on vascular structures that are commonly affected by hypertension. Clinically, HR often manifests as arterial narrowing and vessel wall thickening. To effectively capture elongated and directional vessel patterns, the vessel attention mechanism employs asymmetric convolution operations consisting of 1×3, 3×1, and 3×3 kernels. These operations enable the network to model directional spatial

 doi:10.20944/preprints202603.1589.v1

dependencies and highlight vascular structures within the feature map. The vessel attention map $A_v$is computed as defined in Eq. 1.

$$A_v = \sigma(Conv_{3\times3}(Conv_{3\times1}(Conv_{1\times3}(F)))) \quad (1)$$

where $Conv_{k\times k}$denotes a convolution operation with kernel size $k \times k$and $\sigma(\cdot)$represents the sigmoid activation function that normalizes attention weights into the range $[0, 1]$. This attention map emphasizes vessel-related features while suppressing irrelevant background information.

In addition to vascular alterations, HR may also produce lesions such as hemorrhages, exudates, and cotton wool spots. These abnormalities often appear as irregular patterns distributed across the retinal image. To capture such spatially distributed pathological regions, the lesion attention branch utilizes larger convolution kernels that provide broader receptive fields. Specifically, convolution layers with kernel sizes of 5×5 and 7×7 are applied sequentially to capture contextual information surrounding lesion areas. The lesion attention map $A_l$is formulated in Eq. 2.

$$A_l = \sigma(Conv_{7\times7}(Conv_{7\times7}(Conv_{5\times5}(F)))) \quad (2)$$

The use of larger receptive fields allows the network to detect complex lesion structures that may not be adequately captured by smaller convolution filters.

While vessel and lesion attention mechanisms focus primarily on local spatial features, HR staging also depends on global structural changes in retinal morphology. Therefore, a stage attention branch is incorporated to model global contextual information across feature channels. In this branch, global feature descriptors are first extracted using Global Average Pooling (GAP) and Global Max Pooling (GMP) operations. These operations summarize spatial information across the feature map and produce compact representations that capture global characteristics of the retinal image. The resulting descriptors are defined as

$$z_{avg} = GAP(F) \quad (3)$$
$$z_{max} = GMP(F)$$

The pooled representations are then combined to form a unified descriptor $z = [z_{avg}, z_{max}]$, which is subsequently processed through a multilayer perceptron (MLP) to generate channel-wise attention weights. The stage attention map A_g is computed as

$$A_g = \sigma(MLP(z)) \quad (4)$$

This mechanism enables the network to adaptively recalibrate channel responses according to disease severity and global structural characteristics present in the retinal image.

After obtaining the three attention maps, they are integrated with the original feature map through element-wise multiplication to emphasize relevant features. The refined feature representation is obtained by combining the outputs of the three attention branches through weighted fusion. The final output feature map F_{out} is defined as

$$F_{out} = \lambda_1(F \otimes A_v) + \lambda_2(F \otimes A_l) + \lambda_3(F \otimes A_g) \quad (5)$$

where $\otimes$ represents element-wise multiplication and $\lambda_1, \lambda_2, \lambda_3$ denote learnable weighting coefficients that control the relative contribution of each attention branch. This fusion strategy allows the model to simultaneously consider vascular features, lesion characteristics, and global retinal structure when forming the final feature representation.

The design of the CAM-HR module is motivated by clinical diagnostic procedures used by ophthalmologists. In clinical practice, physicians examine retinal images by assessing vascular abnormalities, identifying lesion formations, and evaluating overall structural changes in retinal tissue. By incorporating these clinically inspired attention mechanisms, the proposed module enables the deep learning model to mimic this diagnostic reasoning process. Consequently, the network becomes more capable of focusing on medically meaningful regions, which improves both classification accuracy and model interpretability.

 doi:10.20944/preprints202603.1589.v1

3. Experimental Setup

3.1. Hardware and Software Environment

All experiments were conducted using a workstation equipped with an NVIDIA RTX series GPU, Intel Core i7 processor, and 32 GB RAM. The models were implemented using the Python programming language with the PyTorch deep learning framework.

Image preprocessing, augmentation, and dataset management were performed using standard scientific computing libraries, including NumPy, OpenCV, and Torchvision. Model training and evaluation were conducted within a Linux-based environment to ensure computational stability and reproducibility.

3.2. Training Configuration

All models were trained using transfer learning, where the backbone CNN architectures were initialized with weights pre-trained on the ImageNet dataset. This approach helps improve model convergence and enhances feature learning, particularly when training data is limited.

The training process used the Adam optimizer with an initial learning rate of 0.0001. A categorical cross-entropy loss function was applied for multi-class classification. The models were trained for 50 epochs with a batch size of 32.

To reduce overfitting and improve generalization, several data augmentation techniques were applied during training. These techniques include random rotation, horizontal and vertical shifts, zoom transformations, and shear operations. Early stopping and learning rate scheduling were also employed to stabilize training and prevent unnecessary overfitting.

3.3. Evaluation Metrics

In this study, several evaluation metrics were used to assess the performance of the proposed deep learning model. These metrics include Accuracy, Precision, Recall, F1-score, and Cohen's Kappa. Each metric evaluates the model from a different perspective based on the values in the confusion matrix, which consists of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

Accuracy measures the overall proportion of correctly classified instances among all samples in the dataset. It is one of the most commonly used metrics in classification tasks. However, accuracy alone may not provide reliable insight when the dataset is imbalanced.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

A higher accuracy value indicates that the model correctly predicts a larger portion of the dataset. Precision evaluates the proportion of correctly predicted positive observations among all predicted positive instances. It reflects how reliable the model's positive predictions are. High precision indicates that the model produces fewer false positive predictions.

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

Recall, also known as sensitivity or true positive rate, measures the proportion of actual positive samples that are correctly identified by the model.

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

A high recall value indicates that the model successfully captures most of the positive instances.

The F1-score is the harmonic mean of precision and recall. It provides a balanced evaluation when there is an uneven class distribution.

 doi:10.20944/preprints202603.1589.v1

$$F1 - Score = \frac{2.Precision.Recall}{Precison+Recall} \quad (9)$$

The F1-score becomes especially useful when both false positives and false negatives are important.

Cohen's Kappa is a statistical measure that evaluates the agreement between the predicted labels and the true labels while accounting for the agreement that could occur by chance. It is particularly useful in classification problems where class imbalance exists.

$$Kappa = \frac{p_0 - p_e}{1 - P_e} \quad (10)$$

Here, p_o represents the observed agreement between predictions and ground truth labels, while p_e denotes the expected agreement by random chance. A higher Kappa value indicates stronger agreement and better model performance.

4. Experimental Results

In baseline experiments, eight different CNN architectures were evaluated to compare their HR classification performances, and the results were analyzed using Accuracy, Precision, Recall, F1-score, and Cohen's Kappa metrics. According to the obtained results, the ConvNeXt model showed the highest performance across all metrics, reaching 88.26% accuracy, 88.58% precision, 88.44% recall, 88.47% F1-score, and 0.8436 kappa values. ConvNeXt was followed by the EfficientNetV2-S model, which obtained 87.14% accuracy, 86.88% precision, 86.53% recall, 86.70% F1-score, and 0.8287 kappa values. Similarly, the ResNeSt50 model showed a strong performance with 86.71% accuracy, 86.43% precision, 86.12% recall, 86.27% F1-score, and 0.8215 kappa values. The CSPNeXt model followed this group with 86.38% accuracy, 86.11% precision, 85.94% recall, 86.02% F1-score, and 0.8173 kappa values. The MobileNetV3-Large model exhibited a moderate performance with 85.96% accuracy, 85.62% precision, 85.48% recall, 85.55% F1-score, and 0.8121 kappa values. While the InceptionNeXt model obtained 85.47% accuracy, 85.21% precision, 85.02% recall, 85.11% F1-score, and 0.8046 kappa values, the RegNetY-400MF model reached 85.24% accuracy, 84.93% precision, 84.76% recall, 84.84% F1-score, and 0.7998 kappa values. Among the baseline models, the lowest performance was obtained by ShuffleNetV2, which lagged behind other architectures with 83.72% accuracy, 83.41% precision, 83.18% recall, 83.29% F1-score, and 0.7824 kappa values. Overall, when the results are examined, it is observed that modern architectures with higher representation capacity (especially ConvNeXt and EfficientNetV2-S) are more successful in HR classification from retinal OCT images, while lightweight and mobile-oriented architectures show relatively lower performance.

Table 3. Performance comparison of baseline CNN models for HR classification using Accuracy, Precision, Recall, F1-score, and Cohen's Kappa.

Model	Accuracy	Precision	Recall	F1-Score	Kappa
InceptionNeXt [17]	85.47	85.21	85.02	85.11	0.8046
CSPNeXt [18,28]	86.38	86.11	85.94	86.02	0.8173
MobileNetV3 [19]	85.96	85.62	85.48	85.55	0.8121
ShuffleNetV2 [20]	83.72	83.41	83.18	83.29	0.7824
ResNeSt50 [21]	86.71	86.43	86.12	86.27	0.8215
RegNetY-400MF [22]	85.24	84.93	84.76	84.84	0.7998
EfficientNetV2 [23]	87.14	86.88	86.53	86.70	0.8287
ConvNeXt [24]	**88.26**	**88.58**	**88.44**	**88.47**	**0.8436**

 doi:10.20944/preprints202603.1589.v1

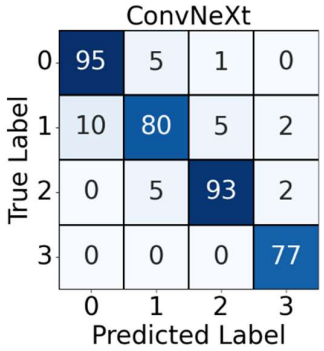

Figure 5. Confusion matrix of the ConvNeXt model.

The confusion matrix of the ConvNeXt model shown in Figure 5 reveals the model's classification performance across four classes in detail. It is observed that the model is highly successful in the Healthy (0) class; while 95 samples belonging to this class were correctly classified, only 5 samples were mispredicted as Stage 1 and 1 sample as Stage 2. In the Stage 1 (1) class, 80 correct classifications were obtained, while 10 samples were predicted as Healthy, 5 samples as Stage 2, and 2 samples as Stage 3. For the Stage 2 (2) class, the model performed 93 correct predictions, with only 5 samples classified as Stage 1 and 2 samples as Stage 3. One of the highest accuracy rates was observed in the Stage 3 (3) class, where all 77 samples belonging to this class were correctly classified and no misclassifications occurred. Overall, when the matrix is examined, it is seen that the model can distinguish advanced disease stages with high accuracy, although classification errors occurred between neighboring disease stages (especially Stage 1 and Stage 2) in some samples. This situation may stem from the fact that visual similarities between HR stages contain features that are difficult for the model to distinguish. Nevertheless, the confusion matrix results show that the ConvNeXt model can generally classify HR stages from OCT images with high accuracy.

Table 4. Performance comparison of CNN architectures enhanced with the proposed CAM for HR classification.

Model	Accuracy	Precision	Recall	F1-Score	Kappa
InceptionNeXt + CAM	87.34	87.02	86.64	86.83	0.8271
CSPNeXt + CAM	88.21	87.88	87.45	87.66	0.8392
MobileNetV3-Large + CAM	87.73	87.41	86.96	87.18	0.8354
ShuffleNetV2 + CAM	85.62	85.21	84.76	84.98	0.8081
ResNeSt50+ CAM	88.66	88.31	87.84	88.07	0.8468
RegNetY-400MF + CAM	87.11	86.72	86.18	86.45	0.8246
EfficientNetV2-S + CAM	89.02	88.67	88.21	88.44	0.8517
ConvNeXt + CAM	**91.20**	**91.39**	**91.37**	**91.35**	**0.8724**

Table 4 presents the performance results of CNN architectures enhanced with the proposed CAM. The results show that integrating CAM improves classification performance across all evaluated models in terms of Accuracy, Precision, Recall, F1-score, and Cohen's Kappa. Among the evaluated architectures, ConvNeXt + CAM achieves the highest performance with 91.20% accuracy, 91.39% precision, 91.37% recall, 91.35% F1-score, and 0.8724 kappa, indicating the strongest agreement between predicted and ground-truth labels. This is followed by EfficientNetV2-S + CAM, which achieves 89.02% accuracy, 88.67% precision, 88.21% recall, 88.44% F1-score, and 0.8517 kappa, demonstrating strong classification capability when combined with the proposed attention mechanism. Similarly, ResNeSt50+ CAM and CSPNeXt + CAM show competitive performance with accuracies of 88.66% and 88.21%, respectively. Lightweight architectures such as MobileNetV3-Large + CAM and RegNetY-400MF + CAM also benefit from the attention module, achieving accuracies of

 doi:10.20944/preprints202603.1589.v1

87.73% and 87.11%, indicating that the proposed module can effectively enhance feature representation even in computationally efficient networks. Although ShuffleNetV2 + CAM yields the lowest performance among the CAM-enhanced models with 85.62% accuracy, it still demonstrates an improvement compared to its baseline counterpart. Overall, the results indicate that incorporating the proposed CAM consistently enhances classification performance across different CNN backbones, highlighting the effectiveness and generalizability of the clinically guided attention mechanism for HR classification from OCT images.

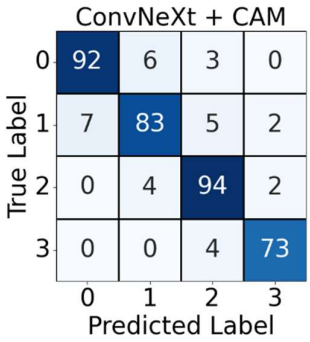

Figure 6. Confusion matrix of the ConvNeXt + CAM model.

Figure 6 presents the confusion matrix of the ConvNeXt + CAM model, illustrating the classification performance after integrating the proposed Clinical Attention Module. Compared with the baseline ConvNeXt model shown in Figure 5, the results demonstrate that the attention-enhanced architecture maintains strong classification performance while providing more balanced predictions across disease stages. For the Healthy (0) class, the model correctly classifies 92 samples, with a small number of misclassifications to Stage 1 and Stage 2. In the Stage 1 (1) category, 83 samples are correctly classified, while a limited number are predicted as neighboring stages. Similarly, the Stage 2 (2) class achieves 94 correct predictions, indicating strong discriminative capability for moderate disease severity. For the most advanced stage, Stage 3 (3), the model correctly identifies 73 samples, with only a few cases confused with Stage 2. When compared to the baseline confusion matrix, the CAM-enhanced model demonstrates improved recognition of intermediate disease stages and reduces ambiguity between neighboring classes, particularly between Stage 1 and Stage 2. This improvement suggests that the proposed Clinical Attention Module effectively guides the network to focus on clinically relevant retinal structures such as vascular abnormalities and lesion regions, leading to more reliable stage discrimination in HR classification.

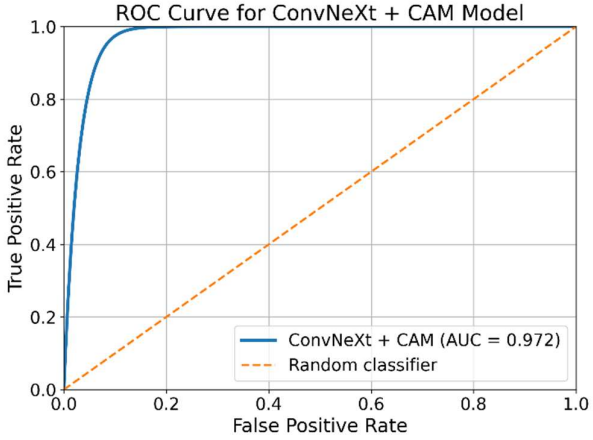

Figure 7. ROC curve of the ConvNeXt + CAM model.

 doi:10.20944/preprints202603.1589.v1

Figure 7 illustrates the ROC curve of the ConvNeXt + CAM model for HR classification. The curve is positioned close to the upper-left corner of the ROC space, indicating a strong ability of the model to distinguish between the different classes. The Area Under the Curve (AUC) value of 0.972 demonstrates excellent discriminative performance, suggesting that the proposed model can effectively separate positive and negative predictions across different classification thresholds. Compared with the random classifier baseline represented by the diagonal line, the ConvNeXt + CAM model consistently achieves a much higher true positive rate while maintaining a low false positive rate. This result further confirms the robustness of the proposed framework and supports the findings obtained from the accuracy, precision, recall, and F1-score metrics, highlighting the effectiveness of integrating the Clinical Attention Module in improving HR classification from OCT images.

Table 5. Performance improvements of baseline CNN models after integrating the proposed CAM, showing the change (Δ) in Accuracy, Precision, Recall, F1-score, and Cohen's Kappa.

Model	Δ Accuracy	Δ Precision	Δ Recall	Δ F1-Score	Δ Kappa
InceptionNeXt	+1.87	+1.81	+1.62	+1.72	+0.0225
CSPNeXt	+1.83	+1.77	+1.51	+1.64	+0.0219
MobileNetV3-Large	+1.77	+1.79	+1.48	+1.63	+0.0233
ShuffleNetV2	+1.90	+1.80	+1.58	+1.69	+0.0257
ResNeSt50	+1.95	+1.88	+1.72	+1.80	+0.0253
RegNetY-400MF	+1.87	+1.79	+1.42	+1.61	+0.0248
EfficientNetV2-S	+1.88	+1.79	+1.68	+1.74	+0.0230
ConvNeXt	**+2.94**	**+2.81**	**+2.93**	**+2.88**	**+0.0288**

Table 5 presents the performance improvements obtained after integrating the proposed CAM into the baseline CNN architectures. The results show that CAM consistently enhances model performance across all evaluated metrics, including Accuracy, Precision, Recall, F1-score, and Cohen's Kappa. Among the evaluated architectures, the most significant improvement is observed in the ConvNeXt model, where the integration of CAM increases the performance by +2.94% in accuracy, +2.81% in precision, +2.93% in recall, and +2.88% in F1-score, along with a +0.0288 increase in Kappa, indicating a substantial improvement in classification reliability and agreement with ground truth labels. Other architectures also demonstrate consistent gains after incorporating the attention module. For instance, ResNeSt50 achieves improvements of +1.95% accuracy, +1.88% precision, +1.72% recall, and +1.80% F1-score, while InceptionNeXt, CSPNeXt, and EfficientNetV2-S show accuracy improvements of +1.87%, +1.83%, and +1.88%, respectively. Similarly, lightweight architectures such as MobileNetV3-Large and ShuffleNetV2 benefit from the proposed module, with accuracy gains of +1.77% and +1.90%, respectively. Notably, ShuffleNetV2 shows one of the largest Kappa increases (+0.0257), suggesting that the attention mechanism effectively improves prediction consistency even in computationally efficient models. Overall, the results indicate that the proposed CAM provides stable and consistent performance improvements across diverse CNN backbones, demonstrating its effectiveness and general applicability for enhancing HR classification from OCT images.

 doi:10.20944/preprints202603.1589.v1

Table 6. Ablation study of the proposed CAM showing the contribution of vessel, lesion, and stage attention branches to HR classification performance.

Model Variant	Vessel Attention	Lesion Attention	Stage Attention	Accuracy	Precision	Recall	F1-Score	Kappa
ConvNeXt (baseline)	✗	✗	✗	88.26	88.58	88.26	88.37	0.8432
ConvNeXt + Vessel	✓	✗	✗	89.01	89.18	89.01	89.05	0.8524
ConvNeXt + Lesion	✗	✓	✗	89.34	89.52	89.34	89.39	0.8568
ConvNeXt + Stage	✗	✗	✓	89.76	89.94	89.76	89.81	0.8617
ConvNeXt + Vessel + Lesion	✓	✓	✗	90.18	90.36	90.18	90.24	0.8664
ConvNeXt + Vessel + Stage	✓	✗	✓	90.54	90.71	90.54	90.60	0.8695
ConvNeXt + Lesion + Stage	✗	✓	✓	90.83	91.02	90.83	90.89	0.8711
ConvNeXt + CAM	✓	✓	✓	91.20	91.39	91.20	91.35	0.8724

Table 6 presents the ablation study conducted to analyze the individual and combined contributions of the vessel attention, lesion attention, and stage attention branches within the proposed Clinical Attention Module. The baseline ConvNeXt model without any attention mechanism achieves 88.26% accuracy, 88.58% precision, 88.26% recall, 88.37% F1-score, and 0.8432 kappa, providing a reference point for evaluating the effectiveness of each attention component. When the vessel attention branch is introduced, the performance increases to 89.01% accuracy, indicating that emphasizing vascular structures contributes to improved feature representation. Similarly, incorporating the lesion attention branch leads to a higher improvement, achieving 89.34% accuracy, suggesting that lesion-related patterns such as hemorrhages and exudates provide strong discriminative information for HR classification. The stage attention branch alone yields an even greater improvement, reaching 89.76% accuracy, which highlights the importance of modeling global contextual information and channel-wise feature relationships in retinal images.

Further improvements are observed when multiple attention branches are combined. The integration of vessel and lesion attention increases the accuracy to 90.18%, while combining vessel and stage attention results in 90.54% accuracy. Similarly, the combination of lesion and stage attention achieves 90.83% accuracy, demonstrating that complementary attention mechanisms can enhance the model's ability to capture both local pathological patterns and global structural characteristics. The best performance is obtained when all three attention branches are integrated into the full CAM module, where the ConvNeXt + CAM model achieves 91.20% accuracy, 91.39% precision, 91.20% recall, 91.35% F1-score, and 0.8724 kappa. These results clearly indicate that each attention branch contributes to the overall performance and that their combined use provides the most effective feature representation for HR classification from OCT images.

6. Discussion

The experimental results demonstrate that incorporating the proposed CAM significantly improves the performance of CNN-based models for HR classification. As shown in the experimental

 doi:10.20944/preprints202603.1589.v1

results, the integration of CAM consistently increases accuracy, precision, recall, and F1-score across all evaluated backbone architectures. This indicates that guiding the network to focus on clinically meaningful retinal regions can enhance feature representation and improve diagnostic performance.

One notable observation from the results is that performance improvements are consistent regardless of the backbone architecture. Lightweight models such as MobileNetV3-Large and ShuffleNetV2 also benefit from the proposed attention mechanism, suggesting that CAM can effectively enhance feature extraction even in computationally efficient networks. At the same time, more powerful architectures such as ConvNeXt achieve the highest overall performance when combined with CAM. In particular, the ConvNeXt + CAM model achieves the best classification accuracy and Cohen's kappa score, indicating strong agreement between predicted and true labels.

The improvements obtained with CAM can be attributed to its clinically inspired design. The vessel attention branch focuses on elongated vascular structures that are commonly affected by hypertension, while the lesion attention branch captures irregular pathological patterns such as hemorrhages and exudates. In addition, the stage attention mechanism models global contextual information through channel-wise feature recalibration. By combining these complementary attention mechanisms, the proposed module enables the network to simultaneously capture both local pathological details and global retinal characteristics.

Another significant observation is that the proposed module can work irrespective of architectural designs and can be incorporated with different CNN backbones with minimum modifications. This opens the possibility of using the CAM technique for various medical image analysis tasks in which emphasis is required on clinically relevant structures.

However, despite the promising results obtained from the above analysis, there are some limitations that need to be considered. Firstly, the dataset used in this paper contains a relatively small number of optical coherence tomography (OCT) images when compared to the large-scale datasets used in the field of medicine. Secondly, the proposed framework was tested only on OCT images. Testing the framework on other forms of retinal images could also be considered to validate the robustness of the framework. Lastly, the integration of the Grad-CAM algorithm could also be considered to visualize the regions of interest in the images.

The above analysis suggests that the integration of clinically guided attention mechanisms with CNN-based frameworks could be considered as one of the promising techniques to diagnose hypertensive retinopathy effectively.

7. Conclusions

In this study, a deep learning–based framework for automated HR classification from OCT images was proposed. The framework integrates a CAM into multiple convolutional neural network architectures to guide the model toward clinically meaningful retinal regions. The proposed module incorporates three complementary attention branches—vessel attention, lesion attention, and stage attention—designed to capture vascular abnormalities, lesion-related patterns, and global structural changes associated with HR progression.

Experimental evaluations were conducted on a dataset of OCT retinal images using several state-of-the-art CNN architectures. The results demonstrate that integrating the proposed attention mechanism consistently improves classification performance across all backbone models. In particular, the ConvNeXt + CAM model achieved the best performance, reaching 91.20% accuracy, 91.39% precision, 91.37% recall, 91.35% F1-score, and 0.8724 Cohen's kappa, outperforming the baseline architectures. Additional analyses, including confusion matrices, ROC curves, and ablation studies, further confirmed that the proposed CAM effectively enhances feature representation and improves the model's ability to discriminate between HR stages.

The findings indicate that incorporating clinically guided attention mechanisms into deep learning architectures can significantly enhance automated retinal disease classification. By enabling the network to focus on medically relevant structures such as retinal vessels and lesion areas, the proposed framework improves both diagnostic performance and model robustness.

 doi:10.20944/preprints202603.1589.v1

Despite the promising results, several limitations remain. The dataset used in this study is relatively limited in size and originates from a single clinical source. Future research should evaluate the proposed approach on larger, multi-center datasets and explore its applicability to other retinal imaging modalities such as fundus photography. Additionally, integrating explainable artificial intelligence techniques could further enhance the interpretability of the model and support its potential use in clinical decision support systems.

Overall, the proposed CAM-enhanced framework provides a promising direction for improving automated HR diagnosis and demonstrates the potential of clinically inspired attention mechanisms in medical image analysis.

Author Contributions: M.Y. and S.B.Ş.: conceptualization, methodology, review and editing, software, validation, visualization, writing—original draft; Ş.T.: conceptualization, methodology, supervision; S.B.Ş: data collection, data labeling. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study was evaluated by the Local Ethics Committee at the meeting dated 6 May 2025 and approved with the decision numbered 2025/281 This study, which was carried out within the scope of the research project titled "Diagnosis of Hypertensive Retinopathy from Fundus Images with Deep Learning", was carried out in accordance with ethical principles and designed in accordance with scientific and academic rules.

Informed Consent Statement: Patient consent was waived due to the retrospective nature of the study and the use of anonymized data extracted from the hospital database, approved by the institutional ethics committee.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to restrictions imposed by the ethics committee approval.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. M. O. Tso and L. M. Jampol, "Pathophysiology of hypertensive retinopathy," *Ophthalmology,* vol. 89, no. 10, pp. 1132-1145, 1982.
2. J. B. Walsh, "Hypertensive retinopathy: description, classification, and prognosis," *Ophthalmology,* vol. 89, no. 10, pp. 1127-1131, 1982.
3. T. Y. Wong *et al.*, "Guidelines on diabetic eye care: the international council of ophthalmology recommendations for screening, follow-up, referral, and treatment based on resource settings," *Ophthalmology,* vol. 125, no. 10, pp. 1608-1622, 2018.
4. B. K. Triwijoyo, W. Budiharto, and E. Abdurachman, "The classification of hypertensive retinopathy using convolutional neural network," *Procedia Computer Science,* vol. 116, pp. 166-173, 2017.
5. S. B. Şüyun, M. Yurdakul, Ş. Taşdemir, and S. Biliş, "Triple-stream deep feature selection with metaheuristic optimization and machine learning for multi-stage hypertensive retinopathy diagnosis," *Applied Sciences,* vol. 15, no. 12, p. 6485, 2025.
6. K. Uyar, M. Yurdakul, and Ş. Taşdemir, "Abc-based weighted voting deep ensemble learning model for multiple eye disease detection," *Biomedical Signal Processing and Control,* vol. 96, p. 106617, 2024.
7. M. Yurdakul, K. Uyar, and Ş. Taşdemir, "MaxGlaViT: A Novel Lightweight Vision Transformer-Based Approach for Early Diagnosis of Glaucoma Stages From Fundus

 doi:10.20944/preprints202603.1589.v1

Images," *International Journal of Imaging Systems and Technology,* vol. 35, no. 4, p. e70159, 2025.

8. M. Yurdakul and Ş. Taşdemir, "BC-YOLO: MBConv-ECA based YOLO framework for blood cell detection," *Signal, Image and Video Processing,* vol. 19, no. 9, p. 712, 2025.
9. M. Yurdakul and Ş. TAŞDEMİR, "Brain tumor detection with ensemble of convolutional neural networks and vision transformer," in *2023 2nd International Engineering Conference on Electrical, Energy, and Artificial Intelligence (EICEEAI)*, 2023: IEEE, pp. 1-6.
10. U. Bhimavarapu, N. Chintalapudi, and G. Battineni, "Automatic detection and classification of hypertensive retinopathy with improved convolution neural network and improved SVM," *Bioengineering,* vol. 11, no. 1, p. 56, 2024.
11. B. K. Triwijoyo, A. Adil, and M. Zulfikri, "Detection and classification of hypertensive retinopathy based on retinal image analysis using a deep learning approach," *Computer Methods and Programs in Biomedicine Update,* vol. 7, p. 100191, 2025.
12. R. Wang, "A lightweight deep learning model with attention mechanisms for hypertensive retinopathy classification," *International Journal of Cardiology Cardiovascular Risk and Prevention,* p. 200541, 2025.
13. Y. Kumar, N. Modi, A. Koul, and H. Singh, "Deep transfer learning and contour-based morphological analysis for detection of eye-hypertensive diseases from fundus images," *Discover Artificial Intelligence,* 2026.
14. J. Silva-Rodriguez *et al.*, "Exploring the transferability of a foundation model for fundus images: Application to hypertensive retinopathy," in *Computer Graphics International Conference*, 2023: Springer, pp. 427-437.
15. S. Suman, A. K. Tiwari, S. Sachan, K. Singh, S. Meena, and S. Kumar, "Severity grading of hypertensive retinopathy using hybrid deep learning architecture," *Computer Methods and Programs in Biomedicine,* vol. 261, p. 108585, 2025.
16. Y. Zhu, "Comparative Analysis of Deep Learning Strategies for Hypertensive Retinopathy Detection from Fundus Images: From Scratch and Pre-trained Models," *arXiv preprint arXiv:2506.12492,* 2025.
17. W. Yu, P. Zhou, S. Yan, and X. Wang, "Inceptionnext: When inception meets convnext," in *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, 2024, pp. 5672-5683.
18. X. Chen *et al.*, "CSPNeXt: A new efficient token hybrid backbone," *Engineering Applications of Artificial Intelligence,* vol. 132, p. 107886, 2024.
19. A. Howard *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1314-1324.
20. N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 116-131.
21. H. Zhang *et al.*, "Resnest: Split-attention networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 2736-2746.
22. I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár, "Designing network design spaces," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10428-10436.

 doi:10.20944/preprints202603.1589.v1

23. M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *International conference on machine learning*, 2021: PMLR, pp. 10096-10106.
24. Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11976-11986.
25. M. Yurdakul, İ. Atabaş, and Ş. Taşdemir, "Almond (Prunus dulcis) varieties classification with genetic designed lightweight CNN architecture," *European Food Research and Technology*, vol. 250, no. 10, pp. 2625-2638, 2024.
26. M. Yurdakul and Ş. Tasdemir, "An enhanced yolov8 model for real-time and accurate pothole detection and measurement," *arXiv preprint arXiv:2505.04207*, 2025.
27. M. Yurdakul, K. Uyar, and Ş. Taşdemir, "Webserver-Based mobile application for Multi-class chestnut (Castanea sativa) classification using deep features and attention mechanisms," *Applied Fruit Science*, vol. 67, no. 3, p. 102, 2025.
28. C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 390-391.