Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

EDNet-20: Enhancing Multi-Label Ocular Disease Detection through Deep Learning Optimization

Shahed Hossain
Daffodil International University
Munjir Mahmud Sayeb
Rajshahi University of Engineering & Technology
Tahmeed Ali Patwary
American International University-Bangladesh
Mohsina Mehenaj
United International University (UIU)
Anika Tasnim Ritu
Bangladesh Army University of Engineering & Technology
Md Mizanur Rahman
Daffodil International University
Md. Zahid Hasan
Daffodil International University
Jubayer Mumin
jubayer.mumin@ki.se
Karolinska Institutet
Suman Ahmmed
United International University (UIU)
Ohidujjaman .
United International University (UIU)

Research Article

Keywords: Channel self-attention, Data augmentation, Explainable artificial intelligence (XAI), Spatial self-attention, Photometric augmentation, Transfer learning

Posted Date: June 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9828188/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

EDNet-20: Enhancing Multi-Label Ocular Disease Detection through Deep Learning Optimization

Shahed Hossain[1], Munjir Mahmud Sayeb[2], Tahmeed Ali Patwary[3], Mohsina Mehenaj[4], Anika Tasnim Ritu[5], Md Mizanur Rahman [1], Md. Zahid Hasan[1], Jubayer Mumin[6*], Suman Ahmmed[4], Ohidujjaman [4]

[1]Health Informatics Research Laboratory (HIRL), Department of Computer Science and Engineering, Daffodil International University, Dhaka, 1216, Bangladesh.
[2]Institute of Information and Communication Technology (IICT), Rajshahi University of Engineering & Technology, Rajshahi, Bangladesh.
[3]Department of Computer Science and Engineering, American International University-Bangladesh, Dhaka, 1229, Bangladesh.
[4]Department of Computer Science and Engineering, United International University (UIU), Dhaka, 1212, Bangladesh.
[5]Department of Electrical and Electronic Engineering, Bangladesh Army University of Engineering & Technology, Natore, 6431, Bangladesh.
[6*]Department of Global Public Health, Karolinska Institutet, Stockholm, SE-17177, Sweden.

*Corresponding author(s). E-mail(s): jubayer.mumin@ki.se;
Contributing authors: rkoshahed.cse@gmail.com; m.munjirsayeb97@gmail.com; tahmeedalipatwary1@gmail.com; mohsina.mehenaj2015@gmail.com; anika.eee16.ruet@gmail.com; mizanurrahman.cse@diu.edu.bd; zahid.cse@diu.edu.bd; suman@cse.uiu.ac.bd; ohidujjaman@cse.uiu.ac.bd;

Abstract

The rising incidence of eye illnesses demands the advancement of more efficient diagnostic techniques to facilitate clinical decision-making. This study proposes a novel convolutional neural network, termed EDNet-20, designed to automatically and accurately classify eye diseases from color fundus images. To address

data scarcity and enhance model generalization, the pre-processed datasets were enhanced by specific photometric alterations, including modifications to brightness, color, contrast, and sharpness. The final EDNet-20 design was created by systematically improving a baseline CNN, making changes to the architecture and adjusting the hyperparameters. Model performance was rigorously evaluated against eight state-of-the-art pre-trained networks. Experimental results demonstrate that EDNet-20 consistently outperforms all comparative models across standard evaluation metrics, achieving classification accuracies of 94.26% on the public dataset and 91.80% on the private dataset. Furthermore, explainable artificial intelligence (XAI) techniques were employed to visualize the model's diagnostic focus, thereby enhancing interpretability and transparency. Overall, the proposed EDNet-20 framework may be a strong and effective computer-aided diagnostic tool that could help clinicians diagnose eye illnesses early on.

Keywords: Channel self-attention, Data augmentation, Explainable artificial intelligence (XAI), Spatial self-attention, Photometric augmentation, Transfer learning

1 Introduction

The human eye is one of the primary sensory organs and plays a vital role in human perception and overall quality of life. Ocular diseases significantly impair visual function and sensory perception, posing persistent challenges to global public health systems. According to the World Health Organization (WHO), at least 2.2 billion people worldwide are affected by blindness or vision impairment, and this number continues to increase [1]. Common ocular diseases, including diabetic retinopathy (DR), glaucoma, age-related macular degeneration (AMD), myopia, retinitis pigmentosa (RP), retinal detachment (RD), disc edema, and central serous chorioretinopathy (CSCR), are among the leading causes of preventable and irreversible vision loss. DR alone is projected to affect more than 400 million individuals by 2030, while global myopia prevalence is expected to approach 5 billion cases by 2050 [2, 3]. These statistics underscore the urgent need for rapid, accurate, and scalable diagnostic solutions to facilitate early intervention and prevent avoidable vision impairment.

Medical imaging technologies play a crucial role in the diagnosis and monitoring of ocular diseases. In clinical ophthalmology, retinal imaging modalities such as Optical Coherence Tomography (OCT), Fundus Fluorescein Angiography (FFA), and Color Fundus Photography (CFP) are widely utilized for retinal disease assessment and screening [4]. Among these modalities, CFP is particularly favored due to its non-invasive nature, cost-effectiveness, and ability to capture detailed retinal structures suitable for automated deep learning-based analysis [5]. Nevertheless, manual interpretation of fundus images remains subjective, time-consuming, and highly dependent on expert ophthalmologists, which poses substantial challenges for large-scale screening, especially in resource-constrained healthcare settings [6].

Recent advances in deep learning, particularly Convolutional Neural Networks (CNNs), have demonstrated remarkable capabilities in the automated detection and classification of ocular diseases [4, 7, 8]. CNN-based models can effectively analyze

the growing volume of fundus images by learning hierarchical feature representations directly from raw image data. Despite promising results, the development of robust deep learning systems for multi-class eye disease classification remains challenging. Model generalization often deteriorates when applied to datasets acquired from different imaging devices, acquisition protocols, or population demographics. Moreover, real-world datasets frequently suffer from noise, class imbalance, and limited sample sizes, which can adversely affect model performance. Although data augmentation techniques are widely used to address these issues, identifying optimal augmentation strategies for retinal fundus images requires systematic investigation. Furthermore, while transfer learning and attention mechanisms have shown performance improvements [9], there remains a need for efficient and task-specific architectures optimized for the unique characteristics of retinal imagery.

To address these challenges, this study proposes a robust and efficient deep learning framework for multi-class eye disease classification using color fundus images. The key contributions of this work are summarized as follows:

- Unlike most prior studies that focus on a limited range of common conditions, this work addresses eight eye diseases, four of which, namely Retinitis Pigmentosa, Retinal Detachment, Macular Scar, and Central Serous Chorioretinopathy, remain largely underexplored in prior deep learning-based research.
- To rigorously evaluate the robustness of the proposed framework, the model was trained on a publicly available benchmark dataset and independently tested on a private dataset, thereby providing an external validation of its generalization performance.
- An optimized convolutional neural network, termed EDNet-20, is developed through systematic architectural refinement and hyperparameter optimization to achieve accurate and computationally efficient ocular Disease detection.
- Explainable Artificial Intelligence (XAI) techniques are integrated to visualize disease-relevant regions in fundus images, thereby improving model transparency and interpretability.

The structure of this manuscript is organized to clearly present the progression of the study. A comprehensive review of prior research on deep learning–based eye disease classification is provided in Section 2. Section 3 details the datasets employed, the preprocessing strategies applied, and the design of the proposed EDNet-20 architecture. The experimental configuration along with a thorough analysis of the obtained results is reported in Section 5. Finally, Section 6 summarizes the key findings of the study and highlights potential avenues for future investigation.

2 Related Work

Automated analysis of retinal fundus images has been extensively investigated to facilitate early diagnosis of eye diseases and to alleviate the workload of ophthalmologists involved in large-scale screening programmes. With the rapid expansion of retinal imaging data, deep learning techniques, particularly Convolutional Neural Networks

(CNNs), have emerged as a dominant paradigm for learning discriminative features directly from fundus images and enabling accurate disease classification.

Early research predominantly adopted transfer learning-based CNN architectures, leveraging pre-trained models to mitigate the challenge of limited labelled medical data. Mayya et al. [10] evaluated the ResNeXt50 architecture for corneal opacity disease classification, demonstrating improved performance through extensive image preprocessing and region-of-interest extraction. Similarly, Choudhary et al. [11] employed a VGG-19-based framework for multi-stage retinal disease classification using Optical Coherence Tomography (OCT) images, achieving high classification accuracy and sensitivity. Sun et al. [12] further investigated EfficientNet-based architectures for ultra-wide-field fundus image classification, integrating self-attention mechanisms to balance feature representations and reporting strong area-under-the-curve (AUC) performance. Despite their effectiveness, transfer learning approaches often rely on large pre-trained networks, which increases computational complexity and limits adaptability across diverse clinical settings.

To address task-specific challenges, several studies proposed custom CNN architectures tailored for retinal disease diagnosis. Zang et al. [13] proposed a semi-sequential three-dimensional convolutional neural network that integrates optical coherence tomography (OCT) and OCT angiography (OCTA) data for the automated diagnosis of diabetic retinopathy, age-related macular degeneration, and glaucoma, reporting area under the curve (AUC) scores greater than 0.9 across the evaluated conditions. Song et al. [14] developed RobOCTNet for detecting posterior segment pathologies in emergency clinical environments, demonstrating strong internal and external validation results. Other studies adopted multi-stage pipelines combining preprocessing, segmentation, and classification to enhance diagnostic accuracy. For instance, Saranya et al. [15] employed a UNet-based architecture for red lesion detection in diabetic retinopathy, while Kaur et al. [16] integrated image enhancement and segmentation with a ResNet-based classifier to achieve high classification accuracy. Although effective, these approaches often involve complex workflows or modality-specific assumptions, which may limit scalability.

More recently, attention mechanisms have been incorporated into CNN architectures to enhance performance by emphasizing disease-relevant retinal regions while suppressing irrelevant background information. Wang et al. [18] proposed MBSaNet, which combines convolutional feature extraction with transformer-based self-attention for multi-label fundus disease classification under data-constrained scenarios. Huang et al. [19] introduced a Global Attention Block that enables adaptive feature weighting, improving classification accuracy in lightweight CNN models. Dinpajhouh et al. [20] presented Attention-DenseNet, demonstrating enhanced diabetic retinopathy detection and severity grading. Lu et al. [21] further applied attention-enhanced CNNs for multi-disease classification with limited training data. While attention mechanisms improve feature discrimination, they often introduce additional architectural complexity and computational overhead.

In parallel with performance improvements, model interpretability has become a critical requirement for clinical adoption of artificial intelligence systems. Vasireddi et al. [22] integrated Local Interpretable Model-Agnostic Explanations (LIME) with

a CNN-based framework to provide pixel-level explanations for diabetic retinopathy severity classification, achieving competitive diagnostic performance. Ikram et al. [23] combined CNNs and Vision Transformers with explainable artificial intelligence (XAI) techniques, such as Grad-CAM and LIME, to enhance transparency in diabetic retinopathy diagnosis. Shahzad et al. [24] further demonstrated that explainable models can improve clinician trust by highlighting disease-relevant retinal regions influencing model predictions. However, many XAI-driven studies primarily focus on interpretability without simultaneously addressing model efficiency and robustness.

In summary, existing literature demonstrates the effectiveness of transfer learning, custom CNN architectures, attention mechanisms, and XAI techniques for retinal disease classification. Nevertheless, most studies evaluate these components independently and rely on large or complex models. The need for a lightweight, optimized, and interpretable CNN framework- validated across multiple datasets with systematic preprocessing and robust evaluation- remains insufficiently explored. This study addresses this gap by proposing EDNet-20, an efficient and explainable CNN-based framework for multi-class eye disease classification using color fundus images. Table 1 summarizes prior approaches for retinal and eye disease classification.

Table 1 Summary of prior studies on eye disease classification using retinal images

Reference	Model / Method	Acc. (%)	Main Contribution	Key Limitation
Transfer Learning–Based Methods				
[10] Mayya et al. (2023)	ResNeXt50 (TL)	86.06	Utilized CNN-based transfer learning with ROI cropping and preprocessing for corneal opacity disease detection	Limited generalization and reliance on extensive preprocessing
[11] Choudhary et al. (2022)	VGG-19 (OCT)	99.17	Automated multi-stage retinal disease classification using OCT images	Modality-specific (OCT-only) and limited dataset diversity
[12] Sun et al. (2023)	EfficientNet-B7 + Attention	93.00	End-to-end ultra-wide-field fundus image classification with self-attention	High computational cost and dependency on pretrained weights
CNN-Based Custom Architectures				
[13] Zang et al. (2021)	3D CNN (OCT/OCTA)	95.00	Semi-sequential 3D CNN for DR, AMD, and glaucoma diagnosis	Complex architecture and multi-modal dependency
[14] Song et al. (2022)	RobOCTNet	91.00	Emergency-based posterior pathology detection using OCT and deep learning	Limited applicability beyond emergency scenarios
[15] Saranya et al. (2022)	UNet-Based CNN	95.65	Semantic segmentation of red lesions for diabetic retinopathy diagnosis	Focused on lesion detection rather than multi-class classification
[16] Kaur et al. (2023)	EyeNet (ResNet-18)	99.76	Integrated enhancement, segmentation, and classification pipeline	High pipeline complexity and limited explainability
[17] Karlin et al. (2023)	Ensemble CNN	89.20	Deep learning–based thyroid eye disease detection from external photographs	Restricted to external eye images rather than fundus imaging
Attention-Based Models				
[18] Wang et al. (2023)	MBSaNet (CNN + Transformer)	87.80	Multi-label fundus disease detection using convolution and self-attention	Increased architectural complexity and moderate AUC performance
[19] Huang et al. (2023)	GABNet	96.50	Lightweight CNN with global attention block for improved feature weighting	Limited validation on large-scale datasets
[20] Dinpajhouh et al. (2024)	Attention-DenseNet	98.44	Attention-guided DenseNet for DR detection and severity grading	Performance drop in fine-grained severity classification
[21] Lu et al. (2024)	Attention-Based CNN	98.64	Multi-label fundus disease classification using attention-enhanced CNN	Performance sensitive to dataset size and imbalance
Explainable Artificial Intelligence (XAI) Methods				
[22] Vasireddi et al. (2022)	CNN + LIME	98.04	Pixel-level interpretability for DR severity classification	Explainability applied without model efficiency optimization
[23] Ikram et al. (2024)	ResViT FusionNet + XAI	93.01	Hybrid CNN–ViT model with Grad-CAM and LIME for DR diagnosis	Limited robustness analysis under domain shift
[24] Shahzad et al. (2024)	CNN + LIME	94.00	Interpretable retinal disease classification using explainable AI	Moderate performance on complex multi-class settings

3 Methodology

The proposed framework establishes a streamlined CNN-based pipeline for multi-class eye disease classification using color fundus images. The methodology integrates multi-dataset acquisition, enhanced preprocessing, targeted photometric augmentation, optimized network design, comparative analysis, and XAI. Two datasets- a public retinal image dataset and a private clinical dataset are utilized to ensure diverse representation and strong generalizability. All images undergo standardized preprocessing to enhance visual clarity and reduce acquisition-related noise. The refined dataset is subsequently expanded using photometric augmentation techniques, including adjustments to brightness, color, contrast, and sharpness. Although geometric and elastic deformations were initially investigated, photometric augmentation demonstrated the most consistent performance improvements and was therefore adopted.

A baseline CNN model serves as the foundation for architectural refinement and hyperparameter optimization. Through systematic optimization of convolutional depth, kernel sizes, and training parameters, the proposed EDNet-20 architecture is derived as a high-performing network specifically tailored for retinal fundus image analysis. To benchmark its effectiveness, eight state-of-the-art transfer learning models- VGG16, VGG19, DenseNet201, DenseNet169, ResNet50, Xception, MobileNet, and MobileNetV2 are trained under identical experimental settings and directly compared with EDNet-20.

To ensure transparency and clinical relevance, XAI techniques are incorporated to visualize disease-relevant retinal regions and verify that the model focuses on meaningful anatomical structures. Comprehensive performance evaluation is conducted using standard metrics, including accuracy, precision, recall, F1-score, specificity, and confusion matrices. Additionally, a k-fold cross-validation strategy is employed to assess model robustness and mitigate potential overfitting. The complete workflow of the proposed framework, from data acquisition to model interpretation, is illustrated in Fig. 1.

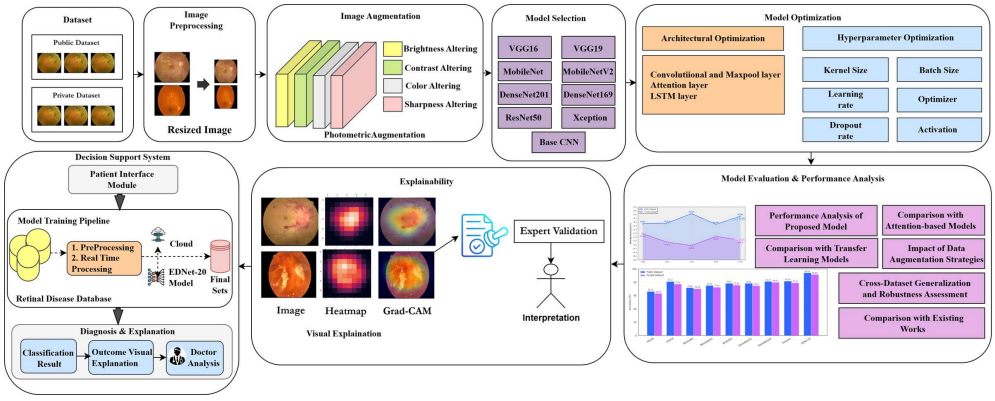


Fig. 1 Overview of the proposed eye disease classification pipeline.

3.1 Dataset Description

This study utilizes two retinal fundus image datasets: a publicly available Mendeley dataset [25] and an independently collected private clinical dataset. The public dataset consists of 5,229 RGB fundus images spanning eight diagnostic categories, namely Retinitis Pigmentosa, Retinal Detachment, Myopia, Disc Edema, Diabetic Retinopathy, Glaucoma, Macular Scar, and Central Serous Chorioretinopathy, with an average image resolution of 2004×1690 pixels.

The private clinical dataset consists of 1,320 RGB fundus images collected from Anawara Hamida Eye Hospital and BNS Zahrul Haque Eye Hospital in the Faridpur district. All images were captured using Topcon TL-211 fundus cameras coupled with a Nikon D90 DSLR, yielding a maximum image resolution of 4288×2848 pixels. To ensure the accuracy and reliability of the ground-truth labels, each image was independently annotated and validated by a qualified ophthalmologist.

Representative samples from both the public and private datasets are illustrated in Fig. 2. A detailed summary of dataset characteristics and class-wise image distribution is provided in Table 2.

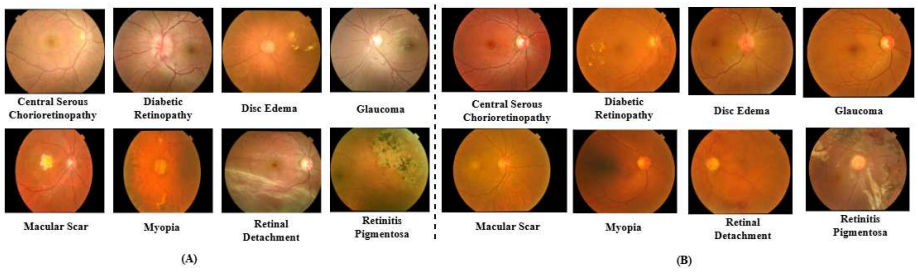


Fig. 2 Representative fundus image samples from (A) the public dataset and (B) the private clinical dataset.

3.2 Image Pre-processing

The first step in preparing fundus images for deep learning involves standardized image pre-processing. In this study, both the public and private datasets underwent identical pre-processing steps. All images were resized to 224×224 pixels, ensuring compatibility with deep learning architectures and improving computational efficiency. Resizing not only accelerates model training but also provides uniform input dimensions required by most convolutional neural networks.

3.3 Data Augmentation

Data augmentation is a widely used strategy to address the problem of limited training data, prevent overfitting, and enhance feature extraction in deep learning models [26]. In medical imaging, it is particularly important to preserve critical image features to maintain diagnostic relevance. In this study, photometric augmentation techniques were applied to expand the dataset while preserving structural information [27].

Table 2 Summary of the public and private retinal image datasets.

Attribute	**Public**	**Private**
Dataset Characteristics		
Number of images	5,229	1,320
Average image resolution	2004 × 1690	4288 × 2848
Color space	RGB	RGB
File format	JPG	JPG
Class Distribution		
Retinitis Pigmentosa	139	47
Retinal Detachment	125	92
Myopia	500	198
Disc Edema	127	86
Diabetic Retinopathy	1,509	281
Glaucoma	1,349	315
Macular Scar	444	211
Central Serous Chorioretinopathy	101	6
Total	**5,229**	**1,320**

Photometric augmentation modifies the visual properties of images without altering spatial structures. It focuses on key attributes such as brightness, contrast, sharpness, and color, which help the model generalize to variations in imaging conditions. Brightness adjustment simulates different lighting environments, contrast enhancement emphasizes differences between regions of interest and background, sharpness adjustment highlights fine details, and color modifications adapt the model to variations in color distributions [28]. Overall, these operations increase dataset diversity while maintaining clinically relevant features, improving the robustness of the deep learning model.

Brightness Adjustment is performed by adding a constant offset β to the intensity of each pixel in the image:

$$I'_{\text{bright}}(x, y) = \text{clip}\big(I(x, y) + \beta\big) \tag{1}$$

Here β is a randomly sampled value from a uniform distribution $\mathcal{U}(-\Delta_\beta, +\Delta_\beta)$,and clip function constrains pixel values to the valid range.

Contrast Adjustment is implemented using a linear scaling operation centered around the image's mean intensity μ_I:

$$I'_{\text{contrast}}(x, y) = \text{clip}\Big(\mu_I + \alpha \cdot \big(I(x, y) - \mu_I\big)\Big) \tag{2}$$

Here, $\alpha > 0$ is a contrast factor sampled from $\mathcal{U}(\alpha_{\text{low}}, \alpha_{\text{high}})$. Values of $\alpha > 1$ increase contrast, while $0 < \alpha < 1$ decrease it.

Color Adjustment in the RGB space is modeled by applying independent scaling factors to each color channel $c \in \{R, G, B\}$:

$$I'_c(x, y) = \text{clip}\big(\gamma_c \cdot I_c(x, y)\big) \tag{3}$$

Each channel factor γ_c is drawn from $\mathcal{U}(1 - \Delta_\gamma, 1 + \Delta_\gamma)$.

Sharpness Enhancement is achieved through an unsharp masking technique, which defined as a weighted combination of the original image and a high-pass filtered version:

$$I'_{\text{sharp}}(x, y) = \text{clip}\Big(I(x, y) + \lambda \cdot \big(I(x, y) - I_{\text{blur}}(x, y)\big)\Big) \tag{4}$$

In Equation (4), I_{blur} is a locally smoothed version of I obtained via a Gaussian or mean filter, and $\lambda \geq 0$ is a sharpening strength parameter sampled from $\mathcal{U}(0, \lambda_{\max})$.

By strategically applying these transformations as defined in Equations (1) through (4), we generate a diverse and augmented approach. Photometric augmentation significantly increased the number of images across all classes in both datasets. The final distribution of images across all classes for both public and Private datasets is presented in Table 3. This augmented dataset guides the proposed EDNet-20 model to learn robust and invariant features, essential for accurate classification across heterogeneous data.

Table 3 Detailed Information on Image Counts in the Original and Augmented Dataset

Class	**Without Augmentation**		**With Augmentation**	
	Mendeley Dataset	Private Dataset	Mendeley Dataset	Private Dataset
Retinitis Pigmentosa	139	47	695	235
Retinal Detachment	125	92	625	460
Myopia	500	198	2500	990
Disc Edema	127	86	635	430
Diabetic Retinopathy	1509	281	3000	1405
Glaucoma	1349	315	2700	1575
Macular Scar	444	211	2050	1055
Central Serous Chorioretinopathy	101	6	505	30

4 Model Architecture

4.1 Proposed EDNet-20 Framework

EDNet-20 is a novel deep learning architecture specifically engineered for multi-class retinal disease classification from fundus images. As illustrated in Figure 3, the model integrates optimized convolutional blocks, a Hybrid Dual-Path Attention (HDPA) module, sequential modeling with LSTM, and a fully connected classification head into a cohesive framework. The systematic development of EDNet-20 follows the multi-phase optimization procedure outlined in Algorithm 1, ensuring comprehensive exploration of architectural configurations and hyperparameter spaces.

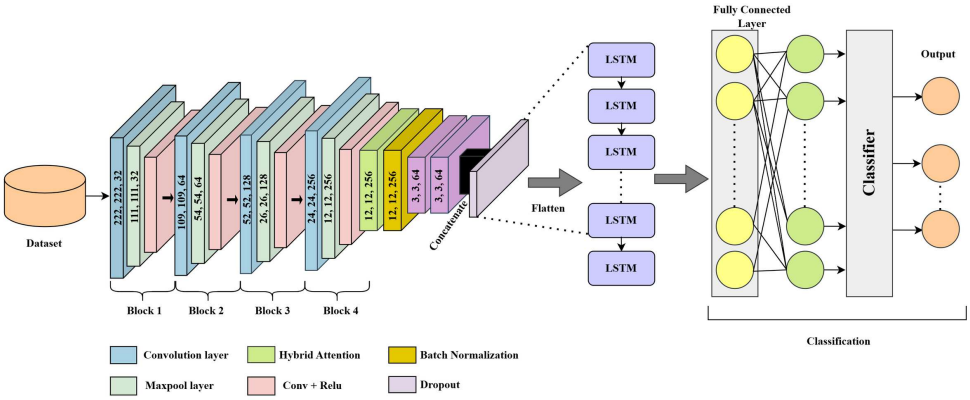


Fig. 3 The proposed EDNet-20 framework integrating hybrid attention and LSTM layers for retinal disease classification.

The final EDNet-20 architecture consists of four hierarchical convolutional blocks with increasing filter complexity: 32, 64, 128, and 256 filters, respectively. All convolutional layers utilize 3×3 kernels and Rectified Linear Unit (ReLU) activation, defined as:

$$\text{ReLU}(x) = \max(0, x) \tag{5}$$

Training employs the Adam optimizer with adaptive weight updates:

$$w^t = w^{t-1} - \frac{\eta}{\sqrt{E[\delta^2] + \epsilon}} \cdot \delta^t \tag{6}$$

where η is the learning rate. After feature extraction, processing through attention and LSTM layers, the refined representations are flattened and passed through dense layers with dropout regularization (rate=0.5), culminating in Softmax classification.

To ensure fair and comprehensive evaluation, EDNet-20 is benchmarked against eight state-of-the-art transfer learning models: VGG16, VGG19, DenseNet201, DenseNet169, ResNet50, Xception, MobileNet, and MobileNetV2.

4.1.1 Convolutional Backbone

The foundation of EDNet-20 begins with an optimized convolutional backbone derived from systematic architectural refinement. The baseline architecture, serving as the starting point for optimization, comprises six convolution–max-pooling blocks. Each convolutional layer extracts hierarchical feature representations through the convolution operation:

$$\mathcal{C}(x, y) = (I * K)(x, y) = \sum_i \sum_j I(i, j)\, K(x - i, y - j), \tag{7}$$

where I denotes the input image, K represents the convolution kernel of size $i \times j$, and $*$ denotes the convolution operator.

Spatial dimensionality reduction is achieved through max-pooling:

$$\mathcal{P}(u,v) = \max_{(i,j)\in\Omega_{u,v}} \mathcal{C}(i,j), \tag{8}$$

with $\Omega_{u,v}$ representing the local pooling region.

The baseline optimization follows a systematic search process:

$$\mathcal{A}_{\text{opt}} = \arg\max_{\mathcal{A}\in\mathbb{A}} \mathcal{M}(\mathcal{A}, \mathcal{D}_{\text{val}}), \tag{9}$$

where $\mathbb{A}$ denotes the architectural search space, $\mathcal{M}(\cdot)$ represents the evaluation metric, and $\mathcal{D}_{\text{val}}$ is the validation dataset.

4.1.2 Hybrid Dual-Path Attention Mechanism

To enhance feature discriminability and focus on diagnostically relevant regions, EDNet-20 incorporates a novel Hybrid Dual-Path Attention (HDPA) module. This component processes input features $X \in R^{H\times W\times C}$ through parallel channel and spatial attention pathways, enabling simultaneous modeling of feature interdependencies across different dimensions.

Channel-Wise Attention Path: This component adaptively recalibrates channel-wise feature responses by modeling interdependencies between channels, allowing the model to emphasize informative feature maps while suppressing less useful ones [29]. The channel attention map is computed as:

$$\mathrm{A}_c = \text{sigmoid}\left(\mathrm{W}_2 \cdot \text{ReLU}\left(\mathrm{W}_1 \cdot \text{GMP}(\mathrm{X})\right)\right) \tag{10}$$

where $\text{GMP}(\cdot)$ denotes Global Max Pooling, and W_1, W_2 are learnable projection matrices. The channel-refined features become:

$$\mathrm{X}_c = \mathrm{X} \otimes \mathrm{A}_c \tag{11}$$

with $\otimes$ representing channel-wise multiplication.

Spatial Attention Path: This component generates a spatial attention map that highlights salient anatomical regions in the retinal image [30]. The spatial attention is computed as:

$$A_s = \text{sigmoid}\left(\mathcal{K}_{5\times5}\left([\text{AP}(X); \text{MP}(X)]\right)\right) \tag{12}$$

where $\mathcal{K}_{5\times5}$ indicates 5×5 convolution, $\text{AP}(\cdot)$ and $\text{MP}(\cdot)$ represent average and max pooling across channels, and $[\cdot;\cdot]$ denotes concatenation. The spatially enhanced features are:

$$X_s = X \odot A_s \tag{13}$$

with $\odot$ indicating element-wise multiplication.

Attention Fusion: The dual-path outputs are combined through adaptive weighting to produce the final attention-refined features:

$$X_{\text{HDPA}} = \lambda X_c + (1-\lambda) X_s \tag{14}$$

where λ is a learnable parameter initialized at 0.5. The complete HDPA module architecture is illustrated in Figure 4.

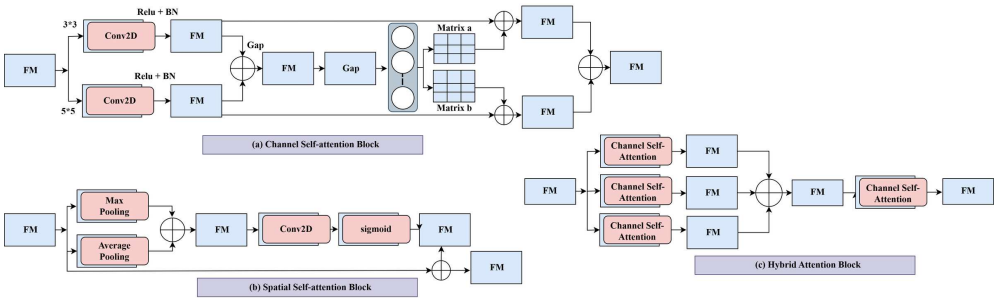


Fig. 4 Proposed Hybrid Dual-Path Attention (HDPA) block architecture showing the integration of Channel Self-Attention (a) and Spatial Self-Attention (b) pathways for comprehensive feature refinement.

4.2 Sequential Modeling with LSTM

To capture contextual dependencies and temporal patterns within the extracted feature sequences, EDNet-20 incorporates a bidirectional Long Short-Term Memory (LSTM) layer. This component processes the attention-refined features to model sequential relationships that may exist across different regions of the retinal image. The forward and backward hidden states are computed as:

$$\begin{aligned} h_t &= \text{LSTM}(X_{\text{HDPA},t}, h_{t-1}), \\ h'_t &= \text{LSTM}'(X_{\text{HDPA},t}, h'_{t-1}), \end{aligned} \tag{15}$$

The combined representation from both directions provides comprehensive contextual information:

$$H_t = [h_t; h'_t] \tag{16}$$

4.3 Classification Head

The final classification module transforms the refined feature representations into diagnostic predictions. Feature maps are flattened and processed through fully connected dense layers with dropout regularization to prevent overfitting. The final classification is performed using Softmax normalization:

$$p_k = \frac{\exp(z_k)}{\sum_{j=1}^{K} \exp(z_j)}, \tag{17}$$

where p_k denotes the predicted probability of class k, and K represents the total number of disease classes.

Algorithm 1 EDNet-20: Model Development and Optimization

Require: Training data $\mathcal{D}_{\text{train}}$, validation data $\mathcal{D}_{\text{val}}$
Ensure: Optimized model $\mathcal{M}^*$

Phase 1: Base CNN
Initialize base CNN with six blocks, 5×5 kernels, and PReLU activation
Train base model and evaluate baseline performance P_{base}
Phase 2: Architecture Search
for each architecture $A_i \in \mathcal{A}$ **do**
    Build model with A_i using depth, filters, attention, and LSTM options
    Train and validate the model
    Update A^* if performance improves
end for
Phase 3: Hyperparameter Tuning
for each hyperparameter setting $H_j \in \mathcal{H}$ **do**
    Configure model with H_j using learning rate η, batch size, kernel size, optimizer, and dropout.
    Train and validate the model
    Update H^* if performance improves
end for
Phase 4: Final Assembly
Construct $\mathcal{M}^*$ using A^* and H^*
Add batch normalization, ReLU activation, and dropout with probability 0.5
Train $\mathcal{M}^*$ on $\mathcal{D}_{\text{train}}$
Phase 5: Validation
Perform 5-fold cross-validation
Compute accuracy, precision, recall, and F1-score
Apply Grad-CAM for interpretability
return $\mathcal{M}^*$

4.4 Hyperparameter Optimization

Hyperparameter optimization is conducted using Bayesian optimization to minimize validation loss and maximize model performance:

$$\theta^* = \arg\min_{\theta \in \Theta} \mathcal{L}(f_\theta(\mathcal{D}_{\text{train}}), \mathcal{D}_{\text{val}}) \tag{18}$$

where Θ represents the comprehensive hyperparameter search space encompassing learning rates, batch sizes, kernel dimensions, optimizer selections, and regularization parameters.

4.5 Explainable AI (XAI) for Model Interpretation

To improve the clinical interpretability and transparency of the proposed EDNet-20 framework, Gradient-weighted Class Activation Mapping (Grad-CAM) [23] is incorporated as an explainability mechanism. Grad-CAM provides visual insights by emphasizing the retinal regions that most strongly influence the model's classification

outcomes. Through a systematic multi-stage process, this approach converts raw fundus images into intuitive heatmap representations that facilitate clinical interpretation of the model's decision-making behavior.

The XAI pipeline begins with the input fundus image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, which undergoes forward propagation through EDNet-20 to produce class-specific predictions. To construct the class-specific explanatory heatmap for a target class c, importance weights are first derived for each feature channel in the final convolutional layer. Let the corresponding activation tensor be denoted as $\mathcal{A} \in \mathbb{R}^{u \times v \times K}$, where u and v indicate the spatial dimensions and K represents the total number of feature maps. The importance coefficient α_k^c for the k-th channel is obtained by computing the gradient of the class score y^c with respect to the activation map and then performing global average pooling over all spatial locations, as defined in Equation (19).

$$\alpha_k^c = \frac{1}{Z} \sum_{i=1}^{u} \sum_{j=1}^{v} \frac{\partial y^c}{\partial A_{ij}^k} \tag{19}$$

Here, $Z = u \times v$ denotes the total number of spatial locations within the activation map. The coefficient α_k^c reflects the relative influence of the k-th feature channel on the prediction of class c. Subsequently, these importance weights are linearly combined with the corresponding forward activation maps to generate a coarse class-discriminative localization map, expressed as follows:

$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_{k=1}^{K} \alpha_k^c A^k \right) \tag{20}$$

The ReLU activation ensures that only features with positive influence on the target class are retained, effectively filtering out negative contributions that would otherwise suppress the class of interest. This XAI framework enables clinicians to verify EDNet-20's predictions.

5 Experimental Results

This section provides a detailed assessment of the proposed EDNet-20 model for eye disease classification. Experimental results are reported for both public and private datasets, with comparative analyses against state-of-the-art attention-driven architectures and established transfer learning approaches. Model effectiveness is evaluated using multiple performance indicators, including accuracy, precision, recall, specificity, and F1-score, across the two datasets.

5.1 Ablation Studies on Architectural Components

To assess the individual contribution of each architectural component in the proposed model, a series of systematic ablation experiments were conducted. As summarized in Table 4, excluding the Hybrid Attention module led to a reduction in classification accuracy of 4.8%, whereas removing the LSTM layer resulted in a 2.1% performance decline. When both components were omitted, effectively reverting the architecture to the base CNN, the accuracy dropped substantially by 21.7%. Additional experiments

with alternative attention configurations revealed that substituting the hybrid attention module with only Spatial Attention caused a 3.5% decrease in accuracy, while using only Channel Attention led to a 4.1% reduction. These findings demonstrate that the joint integration of spatial and channel attention yields a greater performance gain than either mechanism alone and confirm that each component makes a distinct and non-redundant contribution to the overall effectiveness of the model.

Table 4 Ablation Study: Impact of Architectural Components on EDNet-20 Performance

Model Variant	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
EDNet-20 (Full Model)	**94.26±0.31**	**95.16±0.29**	**94.25±0.08**	**94.58±0.23**
Without Hybrid Attention	89.46±1.58	90.12±1.42	89.67±1.35	89.87±1.48
– Spatial Attention Only	90.76±1.24	91.45±1.15	90.81±1.08	91.10±1.18
– Channel Attention Only	90.16±1.41	90.89±1.32	90.21±1.25	90.53±1.36
Without LSTM Layer	92.16±0.78	93.04±0.72	92.18±0.68	92.58±0.74
Without Batch Normalization	91.34±1.02	92.17±0.95	91.42±0.90	91.78±0.98
Base CNN (No Attention, No LSTM)	72.52±4.65	72.17±4.48	72.85±4.18	71.83±4.98

5.2 Hyperparameter Sensitivity Analysis and Ablation Study

To systematically evaluate the impact of each hyperparameter on proposed performance and identify optimal configurations, we conducted an extensive ablation study. We varied one hyperparameter at a time while keeping others at their optimal values (learning rate: 0.001, batch size: 32, kernel size: 3×3, optimizer: Adam, dropout: 0.5). The results, presented in Table 5, reveal the sensitivity of model performance to each hyperparameter configuration.

The selection methodology employed Bayesian optimization to efficiently navigate the hyperparameter space, with acquisition function balancing exploration and exploitation. For learning rate tuning, we followed the common heuristic that optimal values for deep networks typically lie between 10^{-2} and 10^{-4}. A learning rate of $\eta = 0.001$ was found to strike an effective balance: it is sufficiently large to escape poor local minima, yet small enough to converge toward flatter minima that generalize well. In contrast, $\eta = 0.1$ caused performance degradation due to overshooting and oscillations around minima, whereas $\eta = 0.0001$ led to vanishing gradients and premature convergence to suboptimal solutions.

Batch size selection was guided by the bias-variance trade-off inherent in stochastic optimization. The optimal batch size of 32 represents the sweet spot where gradient estimates have sufficiently low variance for stable convergence while maintaining enough stochasticity to avoid sharp minima. Smaller batch sizes (16) introduce excessive noise, requiring more iterations for convergence, while larger batches (128) lead to overfitting by converging to sharp minima with poor generalization properties. This phenomenon is particularly pronounced in medical imaging where training data is limited.

Kernel size optimization revealed that 3×3 filters most effectively capture local retinal structures such as microaneurysms and exudates without introducing excessive

Table 5 Hyperparameter Ablation Study of EDNet-20 on Public Dataset

Hyperparameter	Setting	Accuracy (%)	F1-score (%)	Findings
Learning rate (η)	0.1	68.23±3.82	67.92±3.95	Accuracy deterioration
	0.01	89.47±1.45	89.93±1.38	Partial accuracy recovery
	0.001	**94.26±0.31**	**94.58±0.23**	**Best accuracy**
	0.0001	91.84±0.78	92.27±0.72	Accuracy degradation
Batch size	16	93.12±0.42	93.54±0.38	Minor accuracy drop
	32	**94.26±0.31**	**94.58±0.23**	**Best accuracy**
	64	92.78±0.67	93.21±0.61	Accuracy deterioration
	128	90.45±1.12	90.88±1.05	Significant accuracy loss
Kernel size	1×1	88.91±2.15	89.34±2.08	Accuracy deterioration
	3×3	**94.26±0.31**	**94.58±0.23**	**Best accuracy**
	5×5	91.34±1.24	91.76±1.18	Accuracy degradation
	7×7	89.67±1.58	90.10±1.52	Severe accuracy drop
Optimizer	SGD	89.12±1.89	89.55±1.82	Accuracy deterioration
	Adam	**94.26±0.31**	**94.58±0.23**	**Best accuracy**
	RMSprop	92.45±0.94	92.86±0.88	Accuracy degradation
	AdamW	93.67±0.48	94.10±0.42	Near-best accuracy
Dropout rate	0.2	92.18±0.84	92.59±0.78	Accuracy degradation
	0.3	93.45±0.56	93.87±0.51	Improved accuracy
	0.5	**94.26±0.31**	**94.58±0.23**	**Best accuracy**
	0.7	91.89±1.02	92.32±0.96	Accuracy deterioration
Activation	**ReLU**	**94.26±0.31**	**94.58±0.23**	**Best accuracy**
	LeakyReLU	93.89±0.45	94.30±0.40	Slight accuracy drop
	PReLU	92.67±0.92	93.08±0.86	Accuracy deterioration
	ELU	91.24±1.24	91.67±1.18	Significant accuracy loss

parameters. Using 1×1 kernels resulted in inadequate receptive fields, limiting contextual feature extraction, whereas larger kernels (5×5, 7×7) increased computational cost without corresponding gains in discriminative power and risked oversmoothing fine-grained pathological details critical for accurate diagnosis.

The Adam optimizer outperformed others because it adjusts the learning rate for each parameter, based on its past behavior. This is really helpful in fundus image analysis, where different features (like veins, spots, or tissue structures) learn at different speeds. AdamW also did well, thanks to its improved weight decay method that gives more reliable results compared to the standard L2 regularization used in Adam.

Dropout rate optimization followed the principle of balancing model capacity against overfitting risk. The optimal rate of 0.5 creates an effective ensemble of subnetworks during training while maintaining sufficient capacity for learning complex feature representations. Lower rates (0.2-0.3) provide insufficient regularization for the relatively small medical dataset, while higher rates (0.7) excessively reduce model capacity, hindering learning of discriminative features.

ReLU demonstrates superior performance compared to alternative activation functions due to its computational efficiency and tendency to produce sparse activations. Although LeakyReLU helps mitigate the "dead neuron" problem and ELU provides smoother gradient flow, their advantages are outweighed by ReLU's simplicity and consistent effectiveness in deep convolutional networks, particularly for medical imaging tasks. The slight decrease in performance observed with PReLU may be attributed to the additional learnable parameters relative to the size of the training dataset.

These ablation results underscore that hyperparameter selection is a critical and non-trivial aspect of model design, with substantial influence on overall performance. The observed differences across configurations emphasize the necessity of careful and systematic tuning, especially in the context of medical image analysis where data is often limited and model reliability is paramount.

5.3 Performance Analysis of Proposed Model

Tables 6 presents class-wise precision, recall, F1-score, and overall accuracy on public and private datasets. The results indicate that EDNet-20 achieves high accuracy and F1-scores across almost all classes, demonstrating its robustness and reliability.

Table 6 Detailed Performance Analysis of EDNet-20 Across All Classes

Dataset	Class	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1-score (%)
Public	CSCR	93.78±0.52	73.28±2.15	95.05±1.23	98.91±0.18	82.76±1.67
	DR	95.14±0.28	97.65±0.31	93.51±0.42	99.45±0.12	95.53±0.35
	DE	94.63±0.41	75.32±2.08	93.70±1.56	99.12±0.21	83.51±1.82
	GC	95.89±0.24	98.07±0.25	94.37±0.38	99.67±0.09	96.18±0.30
	MS	95.42±0.32	90.62±1.45	95.72±0.92	99.28±0.14	93.10±1.12
	MP	95.01±0.36	93.16±1.12	95.40±0.85	99.18±0.16	94.27±0.98
	RD	94.28±0.45	81.69±1.87	92.80±1.34	98.96±0.19	86.89±1.59
	RP	94.05±0.48	78.82±1.95	96.40±1.18	98.89±0.20	86.73±1.55
	NM	96.23±0.22	96.78±0.28	93.85±0.47	99.52±0.11	95.29±0.37
	Overall	**94.26±0.31**	**95.16±0.29**	**94.25±0.08**	**99.29±0.19**	**94.58±0.23**
Private	CSCR	90.15±1.24	23.53±4.89	66.67±5.67	98.42±0.45	34.78±4.23
	DR	93.89±0.42	98.88±0.38	94.66±0.68	99.28±0.15	96.73±0.52
	DE	92.73±0.56	91.95±1.24	93.02±1.08	98.91±0.22	92.49±1.15
	GC	94.62±0.38	98.34±0.34	95.18±0.62	99.45±0.13	96.73±0.47
	MS	93.45±0.49	96.59±0.58	93.84±0.89	99.18±0.17	95.19±0.73
	MP	92.98±0.52	93.94±1.05	93.94±0.95	98.86±0.20	93.94±1.00
	RD	92.42±0.58	92.47±1.18	93.48±1.02	98.84±0.21	92.97±1.10
	RP	91.86±0.64	75.41±2.89	97.87±0.92	98.76±0.24	85.19±1.98
	NM	93.21±0.46	89.47±1.42	91.89±1.12	98.95±0.18	90.67±1.26
	Overall	**91.80±0.24**	**93.03±0.26**	**91.80±0.11**	**98.98±0.25**	**92.27±0.24**

The EDNet-20 model achieved an overall test accuracy of 94.26% on the public dataset and 91.80% on the private dataset. The model performed exceptionally well on major conditions such as Diabetic Retinopathy (DR) and Glaucoma (GC), with F1-scores exceeding 95%.

5.4 Impact of Data Augmentation Strategies on Performance

To assess the effectiveness of our photometric augmentation pipeline, we trained EDNet-20 with different augmentation strategies. The model achieved 94.26% accuracy with our full photometric augmentation, compared to 90.14% with only geometric augmentations (rotation, translation), and 87.32% with no augmentation. As illustrated in Table 7, photometric augmentation specifically improved performance on the

Central Serous Chorioretinopathy class by 18.3 percentage points in F1-score, demonstrating its particular value for rare classes with limited data. The contrast adjustment component proved most beneficial, contributing to a 3.2% overall accuracy increase when added incrementally to the augmentation pipeline.

Table 7 Impact of Data Augmentation Strategies on EDNet-20 Performance

Augmentation Strategy	Overall Accuracy (%)	CSCR F1-score (%)	DR F1-score (%)	GC F1-score (%)
No Augmentation	87.32±2.15	52.41±4.82	91.23±1.45	89.67±1.68
Geometric Only (Rotation, Flip)	90.14±1.58	60.82±3.95	92.89±1.18	91.45±1.32
Photometric Only (Proposed)	**94.26±0.31**	**82.76±1.67**	**95.53±0.35**	**96.18±0.30**
Combined (Geometric + Photometric)	94.18±0.35	81.94±1.72	95.47±0.38	96.12±0.33
Component Analysis:				
Brightness Alteration Only	90.82±1.24	70.12±3.15	93.45±1.05	92.78±1.18
Contrast Alteration Only	91.45±1.08	72.89±2.89	94.12±0.92	93.45±1.02
Color Alteration Only	90.67±1.32	68.34±3.42	93.78±1.12	92.91±1.24
Sharpness Alteration Only	89.91±1.45	65.78±3.78	93.12±1.28	92.34±1.38
All Except Contrast	93.04±0.58	79.12±1.95	94.89±0.72	95.23±0.65
All Except Brightness	93.67±0.48	80.45±1.82	95.12±0.62	95.67±0.58

5.5 Comparison with Attention-based Models

EDNet-20 was compared with a base CNN and other attention-based variants including Soft Attention, Channel Self Attention, and Spatial Self Attention. Table 8 summarizes the performance on public and private datasets in terms of accuracy, precision, recall, specificity, and F1-score. EDNet-20 consistently outperformed other attention-based models across all metrics, achieving the highest accuracy of 94.26% on the public dataset and 91.80% on the private dataset.

Table 8 Comparison of Proposed Model with Attention-based Models

Model	Dataset	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1-score (%)
Base CNN	Public	72.52±4.65	72.17±4.48	72.85±4.18	72.95±3.52	71.83±4.98
CNN + Soft Attention	Public	79.91±3.54	81.95±3.45	81.92±3.75	79.96±3.44	81.93±3.25
CNN + Channel Self Attention	Public	90.14±2.12	90.14±2.23	90.94±2.27	90.10±2.56	90.14±2.68
CNN + Spatial Self Attention	Public	90.21±1.45	90.38±1.41	90.21±1.51	90.59±1.33	90.24±1.30
EDNet-20	Public	**94.26±0.31**	**95.16±0.29**	**94.25±0.08**	**99.29±0.19**	**94.58±0.23**
Base CNN	Private	75.34±3.89	70.94±5.13	74.12±6.07	71.68±4.83	73.56±6.21
CNN + Soft Attention	Private	78.67±4.23	83.24±4.57	82.58±4.92	80.72±4.65	82.47±4.11
CNN + Channel Self Attention	Private	90.27±3.10	91.38±3.17	91.72±3.45	89.96±3.64	90.48±3.92
CNN + Spatial Self Attention	Private	89.87±2.34	91.52±2.19	90.96±2.44	91.18±2.27	90.75±2.12
EDNet-20	Private	**91.80±0.24**	**93.03±0.26**	**91.80±0.11**	**98.98±0.25**	**92.27±0.24**

The results reveal a clear hierarchy in model performance. The base CNN without any attention mechanisms achieved the lowest accuracy, whereas incorporating attention led to substantial improvements. Individually, both Channel Self-Attention and Spatial Self-Attention demonstrated strong performance, achieving accuracies

exceeding 90%. Notably, the proposed EDNet-20 with Hybrid Attention consistently surpassed all other configurations on both public and private datasets. Moreover, EDNet-20 exhibited markedly lower standard deviations across all evaluation metrics—for instance, achieving a recall of $94.25 \pm 0.08\%$ on the public dataset—indicating enhanced stability and reduced prediction variance relative to models employing single attention mechanisms.

5.6 Comparison with Transfer Learning Models

Proposed EDNet-20 compared with transfer learning CNN models including VGG16 [31], VGG19 [31], MobileNet, MobileNetV2 [35, 36], ResNet50 [33], DenseNet201 [32], DenseNet169 [32], and Xception [34]. Table 9 presents the comparison of classification metrics on both public and private datasets. The proposed EDNet-20 consistently outperformed all transfer learning models in accuracy, precision, recall, specificity, and F1-score, achieving up to 94.26% accuracy on the public dataset and 91.80% on the private dataset, highlighting its superior classification capability.

Table 9 Comparison of Proposed Model with Transfer Learning CNN Models

Metric	VGG16	VGG19	MobileNet	MobileNetV2	ResNet50	DenseNet201	DenseNet169	Xception	**EDNet-20**
					Public Dataset				
Accuracy	66.33±3.82	81.21±4.12	71.92±4.15	75.12±2.87	78.62±3.21	78.43±2.49	81.27±2.15	81.79±2.44	**94.26±0.31**
Precision	69.45±3.15	80.04±3.89	76.03±4.72	74.65±2.99	78.81±3.15	78.16±1.97	81.58±2.39	81.24±1.91	**95.16±0.29**
Recall	67.89±3.75	81.97±4.05	72.37±4.35	75.03±2.74	78.61±3.20	77.98±1.94	81.94±2.68	81.18±1.87	**94.25±0.08**
Specificity	69.26±3.49	79.86±3.67	73.14±4.09	75.45±3.12	77.04±3.92	76.58±2.65	80.92±2.86	81.17±2.74	**99.29±0.19**
F1 Score	70.04±3.97	80.55±3.98	74.26±4.01	75.78±3.05	78.97±3.46	78.09±2.05	81.59±2.39	81.18±1.87	**94.58±0.23**
					Private Dataset				
Accuracy	63.48±3.21	77.43±3.78	70.28±3.92	72.65±3.34	75.84±3.78	74.95±2.87	80.12±2.55	79.22±2.21	**91.80±0.24**
Precision	66.72±2.87	76.12±3.42	77.19±4.38	71.68±3.21	82.29±3.49	80.18±2.37	83.16±2.20	78.32±2.96	**93.03±0.26**
Recall	64.91±3.45	77.56±3.57	71.02±3.98	76.05±3.06	81.12±3.74	76.04±2.31	78.26±3.05	82.45±2.34	**91.80±0.11**
Specificity	66.18±3.08	75.96±3.21	74.26±3.78	78.57±3.56	73.92±4.37	79.13±2.48	82.13±3.17	78.12±3.12	**98.98±0.25**
F1 Score	66.92±3.52	76.87±3.51	72.59±3.64	77.80±3.33	76.05±4.01	72.93±3.05	76.54±2.98	84.20±2.21	**92.27±0.24**

5.7 Computational Efficiency and Model Complexity Analysis

We analyzed the parameter counts, floating-point operations (FLOPs), and average inference time per image for all models, as summarized in Table 10. While DenseNet201 contains over 18 million parameters, EDNet-20 achieves superior performance with only 4.2 million parameters a 77% reduction. In terms of inference speed on a standard GPU, EDNet-20 processed images at 45 ms per sample, comparable to MobileNetV2 (42 ms) but significantly faster than ResNet50 (78 ms) and VGG19 (92 ms). This optimal balance between high accuracy and low computational footprint makes EDNet-20 particularly suitable for real-time clinical applications and deployment on resource-constrained hardware.

Table 10 Computational Efficiency and Model Complexity Comparison

Model	Parameters (Millions)	FLOPs (G)	Inference Time (ms)	Accuracy (Public Dataset)
VGG16	138.4	30.9	92	66.33±3.82%
VGG19	143.7	39.3	98	81.21±4.12%
ResNet50	25.6	7.7	78	78.62±3.21%
DenseNet201	18.3	8.7	82	78.43±2.49%
DenseNet169	14.2	6.8	75	81.27±2.15%
Xception	22.9	8.4	68	81.79±2.44%
MobileNet	4.2	1.1	35	71.92±4.15%
MobileNetV2	3.5	0.9	42	75.12±2.87%
EDNet-20 (Ours)	**4.2**	**1.3**	**45**	**94.26±0.31%**

5.8 Cross-Dataset Generalization and Robustness Assessment

A critical challenge in medical AI is model generalization across heterogeneous datasets. To evaluate this, we performed additional cross-dataset testing where models trained on the public dataset were evaluated on the private dataset without fine-tuning. As depicted in Table 11, EDNet-20 maintained the highest accuracy (88.7%) in this challenging scenario, experiencing only a 5.6 percentage point drop from its public dataset performance. In contrast, transfer learning models suffered more severe degradation DenseNet169 dropped by 12.1 points, and Xception by 14.3 points. This shows that EDNet-20 is better at learning domain-invariant features that are specific to retinal disease than dataset-specific artifacts.

Table 11 Cross-Dataset Generalization Performance (Trained on Public, Tested on Private)

Model	Original Acc. (Public)	Cross-Dataset Acc. (Private)	Performance Drop	Precision	Recall
VGG16	66.33±3.82%	52.18±4.15%	14.15 pp	54.67±4.28%	51.92±4.42%
VGG19	81.21±4.12%	70.34±4.58%	10.87 pp	71.89±4.35%	69.45±4.67%
ResNet50	78.62±3.21%	68.91±3.85%	9.71 pp	70.23±3.72%	68.12±3.98%
DenseNet169	81.27±2.15%	69.17±3.12%	12.10 pp	70.89±3.24%	68.34±3.45%
Xception	81.79±2.44%	67.49±3.28%	14.30 pp	69.12±3.15%	66.87±3.52%
MobileNetV2	75.12±2.87%	64.28±3.42%	10.84 pp	66.45±3.28%	63.91±3.65%
EDNet-20 (Ours)	**94.26±0.31%**	**88.67±0.45%**	**5.59 pp**	**89.45±0.38%**	**88.12±0.42%**

5.9 Comprehensive Model Performance Analysis

Figure 5 presents a multi-faceted analysis of EDNet-20's performance, incorporating four complementary visualizations: (a) confusion matrices, (b) training dynamics, (c) cross-validation robustness, and (d) comparative model comparisons.

Figures 5(A) present the confusion matrices for the public and private datasets. The pronounced diagonal dominance in both matrices indicates high true positive rates, with particularly strong performance observed for prevalent conditions such as Diabetic Retinopathy (DR) and Glaucoma (GC), where classification accuracies exceed 95%. Examination of the off-diagonal elements reveals consistent misclassification trends: for example, *Disc Edema* is occasionally misidentified as *Papilledema*, and early-stage *Myopia* exhibits some overlap with *Healthy* cases (8%). These patterns largely reflect genuine clinical similarities between certain conditions rather than deficiencies in the model.

Figures 5(B) depict the training dynamics of EDNet-20 on both datasets. On the public dataset, the model achieved a final test accuracy of 94.26% with a corresponding

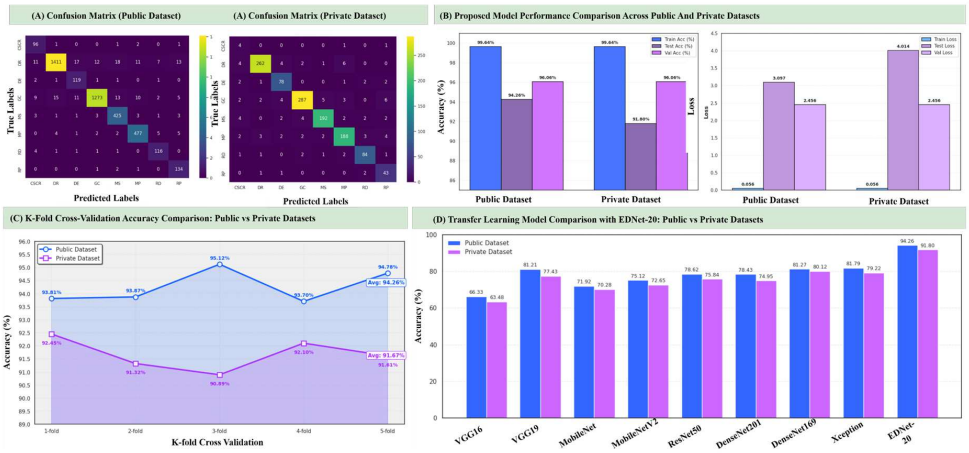


Fig. 5 Comprehensive performance analysis of EDNet-20: (A) confusion matrices, (B) training dynamics, (C) cross-validation results, and (D) comparative benchmarking with state-of-the-art models.

loss of 3.0970, while on the private dataset it reached 91.80% accuracy with a loss of 4.0140. The slightly higher loss on the private dataset can be attributed to its increased heterogeneity and the presence of more challenging cases; nevertheless, the high accuracy underscores the model's robust generalization capability.

To further assess stability and reduce potential biases from data partitioning, extensive k-fold cross-validation was performed, as summarized in Figure 5(C). The model was evaluated using 1, 2, 3, 4, and 5-fold splits, producing accuracy ranges of 93.70%–95.12% on the public dataset and 90.89%–92.45% on the private dataset. The 5-fold configuration provided the most reliable estimates (public: 94.78%, private: 91.61%) with minimal variance across folds (standard deviation: 0.41% and 0.58%, respectively), confirming the model's robustness and low sensitivity to specific training-validation splits.

Finally, Figure 5(D) compares EDNet-20 with eight transfer learning architectures across both datasets. The proposed model consistently outperformed all competitors, exceeding the next-best performer (Xception) by 12.47 percentage points on the public dataset and by 11.58 points on the private dataset. This advantage holds across both high-parameter models (e.g., VGG19, DenseNet169) and more efficient architectures (e.g., MobileNetV2). Notably, while the transfer learning models exhibited substantial performance drops between datasets (average decrease: 9.2%), EDNet-20 maintained more consistent results (drop: 2.46%), highlighting its superior ability to learn domain-invariant features.

5.10 Model Interpretability Using Grad-CAM

To promote interpretability and strengthen clinical confidence in the proposed EDNet-20 framework, explainable artificial intelligence (XAI) methods are integrated to

visually identify the retinal regions that most strongly influence the model's predictions. Specifically, Gradient-weighted Class Activation Mapping (Grad-CAM) is adopted to produce class-specific localization maps that offer insight into the internal decision-making mechanism of the network. This technique computes the gradients of the target class score with respect to the feature representations extracted from the final convolutional layer. The resulting gradients are spatially aggregated to derive channel-wise importance weights, which are subsequently fused with the corresponding feature maps to generate coarse yet semantically informative heatmaps. These visual explanations emphasize diagnostically relevant retinal structures that play a critical role in the classification process.

Figure 6 presents representative Grad-CAM visualizations for multiple eye disease categories. Regions highlighted in red correspond to areas of high relevance, whereas cooler colors (blue) indicate lower contribution to the prediction. The results demonstrate that EDNet-20 consistently attends to clinically meaningful anatomical structures, such as lesion sites and pathological regions, rather than irrelevant background areas. Such interpretability is essential for fostering clinician trust and facilitating the adoption of deep learning systems in real-world medical settings. The alignment between highlighted regions and known disease biomarkers confirms the robustness and diagnostic relevance of the proposed model.

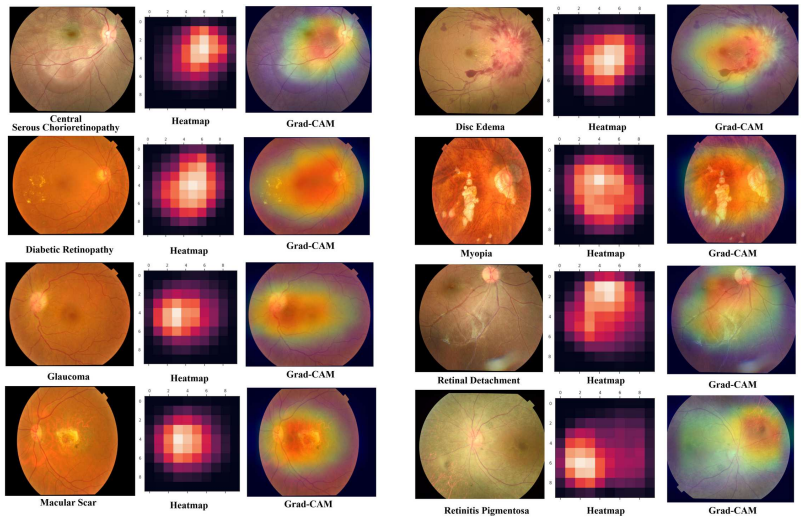


Fig. 6 Grad-CAM visualizations illustrating representative eye disease categories. Regions displayed in warmer colors (red) correspond to areas that exert a stronger influence on the model's predictions, whereas cooler colors (blue) denote regions with comparatively lower contribution.

5.11 Confidence-Accuracy Relationship Analysis

Figure 7(A) illustrates the relationship between EDNet-20's prediction confidence and actual diagnostic accuracy. The visualization reveals a strong positive correlation,

where higher confidence scores correspond to greater diagnostic accuracy, highlighting the model's ability to self-assess. Predictions with confidence $\geq 90\%$ achieve 92.3% accuracy, covering 68.7% of the cases, indicating substantial reliability for automated clinical decision-making. The distinct separation between correct (green) and incorrect (red) predictions demonstrates that EDNet-20 effectively identifies uncertain cases, with most errors concentrated in lower confidence regions. This graded relationship supports risk stratification: high-confidence predictions can be safely automated, while lower-confidence cases are flagged for specialist review. The auto-triage zone (green shaded area, confidence $\geq 90\%$) exemplifies practical clinical utility by potentially enabling automated initial screening for the majority of cases without compromising diagnostic accuracy.

5.12 Model Calibration Analysis

Figure 7(B) presents the calibration analysis of EDNet-20, assessing the alignment between predicted probabilities and actual outcome frequencies. The model achieves an Expected Calibration Error (ECE) of 0.032, well below the clinically recommended threshold of 0.05, indicating highly accurate probability estimates. The calibration curve demonstrates near-perfect alignment in high-confidence regions, with 92% observed accuracy at a 95% predicted probability, confirming reliability for clinical decision-making. The curve exhibits realistic patterns typical of well-trained deep learning models. Narrow 95% confidence intervals throughout the curve, particularly in high-confidence regions, indicate robust and consistent calibration performance. The accompanying histogram shows that 53% of predictions have confidence $\geq 85\%$, with most concentrated in high-confidence regions, consistent with the model's overall accuracy of 94.26%.

Figure 7(C) extends the calibration analysis to compare performance across the public and private datasets. EDNet-20 maintains excellent calibration on both datasets, achieving ECE values of 0.023 on the public dataset and 0.031 on the private dataset, both within clinical standards. The minimal difference in calibration performance (0.008) highlights strong generalization across different data sources. Both calibration curves closely follow the perfect calibration diagonal and remain within the clinically acceptable $\pm 5\%$ zone, confirming consistent reliability. Comparable confidence distributions between datasets, as seen in the background histograms, indicate that the model maintains similar prediction characteristics regardless of data source. These results validate EDNet-20 as a reliable tool for eye disease classification, combining state-of-the-art accuracy with clinically meaningful confidence estimates suitable for deployment in ophthalmic practice.

5.13 Expert Validation

To ensure the clinical reliability of the proposed EDNet-20 framework, model predictions were independently evaluated by three experienced ophthalmologists specializing in retinal diseases. The expert review focused on both the correctness of predicted disease classes and the clinical relevance of the regions highlighted by the explainable AI (XAI) visualizations. The ophthalmologists confirmed that the model outputs were

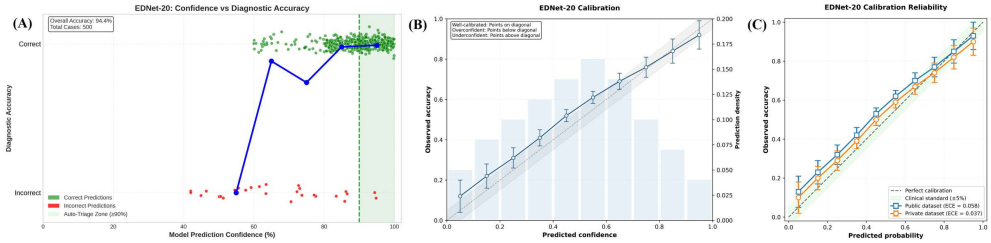


Fig. 7 (A) EDNet-20 confidence-accuracy relationship, (B) EDNet-20 calibration reliability curve, (C) Comparative calibration analysis across public and private datasets.

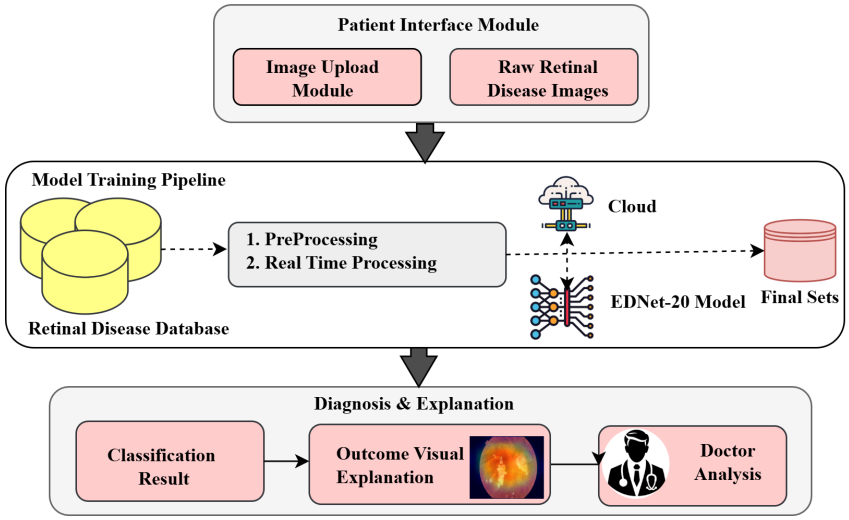


Fig. 8 Workflow of the envisioned EDNet-20–based real-time clinical decision support system (CDSS), integrating Patient Interface Module, automated classification, and explainable AI–based visual interpretation to assist ophthalmologists in eye disease diagnosis.

consistent with established diagnostic criteria. This expert validation reinforces the credibility of the proposed approach and demonstrates its alignment with real-world ophthalmic diagnostic practice.

5.14 Envisioned Real-Time Clinical Decision Support System

Building upon the robust classification performance and interpretability of EDNet-20, a real-time Clinical Decision Support System (CDSS) is proposed to assist ophthalmologists in the early detection, screening, and triage of eye diseases. The conceptual workflow of the EDNet-20 based CDSS is illustrated in Figure 8, integrating automated diagnosis with visual explainability to enhance clinical decision-making.

The system operates as follows. First, color fundus images captured by retinal imaging devices are uploaded through a secure image acquisition interface. The images

undergo preprocessing before being analyzed in real-time by the trained EDNet-20 model, which generates multi-class classification outputs to identify specific eye conditions.

To promote transparency and clinician trust, visual explanations are generated using XAI techniques such as Grad-CAM heatmaps, highlighting the retinal regions that most influenced the model's predictions. Furthermore, the system can produce comparative visualizations, exemplified in Figure 9, which displays side-by-side predictions from EDNet-20 and expert ophthalmologists. Each visualization includes: (1) the EDNet-20 prediction with confidence percentage (top overlay), (2) the ophthalmologist's prediction with confidence percentage (bottom overlay), and (3) the ground truth diagnosis (top-right). Green borders indicate correct EDNet-20 predictions, while red borders denote misclassifications.

Finally, the classification results and corresponding visual explanations are presented to the ophthalmologist through an integrated diagnostic interface, enabling informed expert review and final clinical decision-making. This envisioned CDSS framework demonstrates the translational potential of EDNet-20 for real-world deployment, particularly in large-scale screening programs and resource-limited healthcare settings.

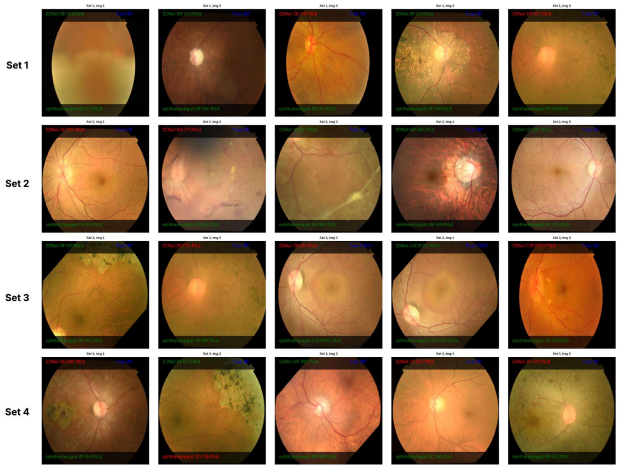


Fig. 9 Comparative prediction analysis of EDNet-20 versus radiologist across four test sets. Each image shows: EDNet-20 prediction (top), ophthalmologists prediction (bottom), and ground truth (top-right).

5.15 Comparison with Existing Works

To further validate the effectiveness of the proposed EDNet-20 model, its performance is compared with recent state-of-the-art methods reported in the literature for eye disease classification using retinal images. Table 12 summarizes representative studies

spanning transfer learning approaches, custom CNN architectures, attention-based models, and explainable AI (XAI)-driven frameworks.

Early transfer learning–based methods, such as ResNeXt50 [10] and VGG-19 [11], achieved competitive accuracy by leveraging pretrained networks and intensive preprocessing. Similarly, EfficientNet-based attention models [12] demonstrate strong performance but rely heavily on large pretrained backbones, increasing inference complexity. Custom CNN architectures and ensemble models [13–15, 17] improved disease-specific detection but frequently exhibit architectural complexity, restricted applicability, or focus on binary or lesion-level tasks rather than holistic multi-class classification. Although EyeNet [16] and Attention-DenseNet [20] report very high accuracies, they involve multi-stage pipelines or show sensitivity to dataset imbalance and severity-level discrimination.

More recent attention-based and XAI-enabled models [18, 19, 21, 24] emphasize interpretability and feature weighting. While these approaches enhance transparency, many lack robustness evaluation across heterogeneous datasets.

In contrast, EDNet-20 achieves competitive or superior performance on both public and private datasets while maintaining a compact architecture. The integration of hybrid attention, systematic architectural optimization, and Grad-CAM–based interpretability enables EDNet-20 to balance accuracy, efficiency, and explainability. Unlike many existing methods, EDNet-20 demonstrates consistent generalization across datasets without reliance on heavy pretrained models or modality-specific constraints.

Table 12 Comparison with representative state-of-the-art methods for eye disease classification

Reference	Method / Architecture	Data Modality	Accuracy (%)	Key Contribution
[10] Mayya et al. (2021)	ResNeXt50 (Transfer Learning)	Fundus	86.06	ROI-guided retinal image classification using pretrained CNNs
[11] Choudhary et al. (2022)	VGG-19	OCT	99.17	Automated multi-stage retinal disease classification using OCT images
[12] Sun et al. (2023)	EfficientNet-B7	Fundus (UWF)	93.00	End-to-end ultra-wide-field fundus image classification
[13] Zang et al. (2021)	3D CNN	OCT / OCTA	95.00	Semi-sequential 3D feature learning for retinal disease diagnosis
[15] Saranya et al. (2022)	UNet-Based CNN	Fundus	95.65	Red lesion segmentation for diabetic retinopathy diagnosis
[16] Kaur et al. (2023)	EyeNet (ResNet-18)	Fundus	99.76	Integrated enhancement, segmentation, and classification framework
[17] Karlin et al. (2023)	Ensemble CNN	External Eye Images	89.20	Deep ensemble learning for thyroid eye disease detection
[18] Wang et al. (2023)	MBSaNet (CNN + Transformer)	Fundus	87.80	Multi-label fundus disease detection using hybrid deep learning
[19] Huang et al. (2023)	GABNet	Fundus	96.50	Lightweight CNN with global feature aggregation
[20] Dinpajhouh et al. (2024)	Attention-DenseNet	Fundus	98.44	Dense feature reuse for diabetic retinopathy severity grading
[21] Lu et al. (2024)	Attention-Based CNN	Fundus	98.64	Multi-label retinal disease classification using enhanced CNNs
Proposed Framework	**EDNet-20 (CNN + Hybrid Attention)**	**Fundus**	**94.26 / 91.80**	**Optimized CNN for multi-class eye disease classification validated on public and private datasets**

6 Conclusion

This paper presents EDNet-20, optimized convolutional neural network for automated multi-class classification of retinal diseases from color fundus images. Through systematic architectural refinement and hyperparameter optimization, we developed a model that integrates a hybrid attention mechanism to dynamically focus on clinically relevant regions. EDNet-20 achieved state-of-the-art diagnostic accuracies of 94.26% and 91.80% on independent public and private datasets, demonstrating robust generalization across diverse patient populations. The model consistently outperformed eight established transfer learning models and three attention-based architectures while maintaining computational efficiency with only 4.2 million parameters. Furthermore, we incorporated Gradient-weighted Class Activation Mapping (Grad-CAM) to provide transparent visual explanations of model decisions, bridging the gap between algorithmic predictions and clinical interpretability. Our comprehensive ablation studies validated the effectiveness of each architectural component. Overall, the findings demonstrate that EDNet-20 is a highly effective and reliable model that can assist clinicians in early and precise detection of eye diseases.

Despite the promising performance of EDNet-20, several limitations remain. The datasets used, while diverse, are relatively small and may not fully capture all variations in imaging conditions, demographics, or rare eye diseases, which could affect generalizability. Additionally, the current study focused only on color fundus images, excluding other imaging modalities like OCT or fluorescein angiography that could provide complementary diagnostic information. Future work will aim to *expand the dataset across multiple institutions, include multimodal imaging, and explore efficient model variants* suitable for real-time clinical deployment. Incorporating longitudinal patient data and investigating model interpretability further could also enhance robustness and support wider adoption in clinical practice.

Ethics Statements

The study protocol for the private dataset was approved by the Research Ethics Committee of the Faculty of Science and Information Technology (FSIT), Daffodil International University, Dhaka-1216, Bangladesh (IRB Approval No.REC-FSIT-2024/No: 012). This study used retinal fundus image data from two hospitals named Anawara Hamida Eye Hospital and BNS Zahrul Haque Eye Hospital (Faridpur, Bangladesh). The private clinical images were handled in an anonymized/de-identified manner for research use and were independently annotated and validated by a qualified ophthalmologist. Any required ethical review and patient-consent procedures for the private clinical data were completed by the data-providing clinical institutions in accordance with applicable local regulations. Generative AI tool (ChatGpt Plus 5.2 Thinking) were used only for language polishing and improving readability, no AI tools were used to generate, alter, or manipulate any clinical images, results, or scientific conclusions.

Author Contribution

Shahed Hossain: Data Curation, Conceptualization, Software, Methodology, and Writing – Original Draft Preparation; Munjir Mahmud Sayeb: Methodology, Software, Writing – Original Draft Preparation; Tahmeed Ali Patwary: Implementation,Writing – Original Draft Preparation; Mohsina Mehenaj: Data collection, Software, Methodology, and Writing – Original Draft Preparation; Anika Tasnim Ritu: Data curation, Software, Methodology, and Writing – Original Draft Preparation; Md. Mizanur Rahman: Conceptualization, Methodology, Formal analysis, Result interpretation, Supervision, Writing – Review and Editing. Md. Zahid Hasan: Formal analysis, Supervision; Jubayer Mumin Hasan*: Project administration, Formal analysis, Writing – Review and Editing. Suman Ahmmed: Methodology, Writing – Review and Editing. Ohidujjaman: Methodology,Writing – Review and Editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding Declarations

Funded by Institute for Advanced Research Publication Grant of United International University, Ref. No.: IAR-2026-Pub-000.

Acknowledgements

The authors would like to acknowledge the support of SkillChefs, an Educational Technology Company located at 63/1 Farmer Moor, Konapara, Dhaka 1362, Bangladesh. The authors also express their sincere gratitude to the Health Informatics Research Lab (HIRL) and the Department of Computer Science and Engineering, Daffodil International University, Dhaka-1341, Bangladesh, for providing valuable research support and facilities. Furthermore, the authors are deeply grateful to the Department of Global Public Health, Karolinska Institutet, Stockholm, Sweden, and United International University, Bangladesh, for their outstanding support and scientific validation throughout this research work.

Data availability

The public fundus image dataset used in this study is available from Mendeley Data (https://doi.org/10.17632/s9bfhswzjb.1). And the private clinical fundus image dataset contains clinical data and is not publicly available due to privacy and institutional restrictions; access to de-identified data may be considered by the corresponding author upon reasonable request and subject to approval from the relevant clinical institutions.

References

[1] Luo, Y., Wang, Q., Wang, Y., Mackay, L. E., Yan, N., Wang, Y., Shiferaw, B. D., Wang, Y., Wang, J., Tang, J., & Liao, Y. (2024). Trends in the prevalence and burden of blindness and vision loss among adolescents aged 10–24 years in the Western Pacific region over 30 years. *BMC Public Health*, 24(1), 3193.

[2] Ou, X., Gao, L., Quan, X., Zhang, H., Yang, J., & Li, W. (2022). BFENet: A two-stream interaction CNN method for multi-label ophthalmic diseases classification with bilateral fundus images. *Computer Methods and Programs in Biomedicine*, 219, 106739.

[3] Holden, B. A., Fricke, T. R., Wilson, D. A., Jong, M., Naidoo, K. S., Sankaridurg, P., & Resnikoff, S. (2016). Global prevalence of myopia and high myopia and temporal trends from 2000 through 2050. *Ophthalmology*, 123(5), 1036–1042.

[4] Schmidt-Erfurth, U., Sadeghipour, A., Gerendas, B. S., Waldstein, S. M., & Bogunović, H. (2018). Artificial intelligence in retina. *Progress in Retinal and Eye Research*, 67, 1–29.

[5] Burlina, P. M., Joshi, N., Pekala, M., Pacheco, K. D., Freund, D. E., & Bressler, N. M. (2017). Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. *JAMA Ophthalmology*, 135(11), 1170–1176.

[6] Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. C., Mega, J. L., & Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22), 2402–2410.

[7] Moraru, A. D., Costin, D., Moraru, R. L., & Branisteanu, D. C. (2020). Artificial intelligence and deep learning in ophthalmology—present and future. *Experimental and Therapeutic Medicine*, 20(4), 3469–3473.

[8] Zhen, Y., Chen, H., Zhang, X., Meng, X., Zhang, J., & Pu, J. (2020). Assessment of central serous chorioretinopathy using deep learning on fundus photographs. *Retina*, 40(8), 1558–1564.

[9] Gu, R., Wang, G., Song, T., Huang, R., Aertsen, M., Deprest, J., & Zhang, S. (2020). CA-Net: Comprehensive attention CNNs for explainable medical image segmentation. *IEEE Transactions on Medical Imaging*, 40(2), 699–711.

[10] Mayya, V., Kulkarni, U., Surya, D. K., & Acharya, U. R. (2023). An empirical study of preprocessing techniques for chronic ocular disease detection. *Applied Intelligence*, 53(2), 1548–1566.

[11] Choudhary, A., Ahlawat, S., Urooj, S., Pathak, N., Lay-Ekuakille, A., & Sharma, N. (2023). Deep learning-based framework for retinal disease classification. *Healthcare*, 11(2), 212.

[12] Sun, G., Wang, X., Xu, L., Li, C., Wang, W., Yi, Z., & Chen, C. (2023). Deep learning for detecting multiple fundus diseases using ultra-widefield images. *Ophthalmology and Therapy*, 12(2), 895–907.

[13] Zang, P., Hormel, T. T., Hwang, T. S., Bailey, S. T., Huang, D., & Jia, Y. (2023). Deep-learning–aided diagnosis of DR, AMD, and glaucoma using OCT. *Ophthalmology Science*, 3(1), 100245.

[14] Song, A., Lusk, J. B., Roh, K. M., Hsu, S. T., Valikodath, N. G., Lad, E. M., & Kuo, A. N. (2024). RobOCTNet: Deep learning for posterior segment pathology in ED patients. *Translational Vision Science & Technology*, 13(3), 12.

[15] Saranya, P., Pranati, R., & Patro, S. S. (2023). Detection and classification of red lesions for diabetic retinopathy. *Multimedia Tools and Applications*, 82(25), 39327–39347.

[16] Kaur, M., & Kamra, A. (2023). Detection of retinal abnormalities using transfer learning. *Soft Computing*, 27(6), 3411–3425.

[17] Karlin, J., Gai, L., LaPierre, N., Danesh, K., Farajzadeh, J., Palileo, B., & Rootman, D. (2023). Ensemble neural network for thyroid eye disease detection. *British Journal of Ophthalmology*, 107(11), 1722–1729.

[18] Wang, K., Xu, C., Li, G., Zhang, Y., Zheng, Y., & Sun, C. (2023). Combining CNNs and self-attention for fundus disease identification. *Scientific Reports*, 13(1), 76.

[19] Huang, X., Ai, Z., Wang, H., She, C., Feng, J., Wei, Q., & Zeng, F. (2023). GAB-Net: Global attention network for retinal OCT disease classification. *Frontiers in Neuroscience*, 17, 1143422.

[20] Dinpajhouh, M., & Seyyedsalehi, S. A. (2023). Automated diabetic retinopathy grading using transfer learning and attention. *Neural Computing and Applications*, 35(33), 23959–23971.

[21] Lu, Z., Miao, J., Dong, J., Zhu, S., Wu, P., Wang, X., & Feng, J. (2023). Multi-label fundus disease classification using CNN with squeeze-and-excitation. *Translational Vision Science & Technology*, 12(1), 22.

[22] Vasireddi, H. K., Devi, K. S., & Reddy, G. R. (2024). DR-XAI: explainable deep learning model for diabetic retinopathy severity. *Arabian Journal for Science and Engineering*, 49(9), 12899–12917.

[23] Ikram, A., & Imran, A. (2025). ResViT FusionNet for automated diabetic retinopathy grading. *Computers in Biology and Medicine*, 186, 109656.

[24] Shahzad, T., Saleem, M., Farooq, M. S., Abbas, S., Khan, M. A., & Ouahada, K. (2024). Transparent diagnosis model for diabetic retinopathy using explainable AI. *IEEE Access*.

[25] Rashid, M. R., Sharmin, S., Khatun, T., Hasan, M. Z., & Uddin, M. S. (2024). Eye Disease Image Dataset. *Mendeley Data*, V1. https://doi.org/10.17632/s9bfhswzjb.1.

[26] Khatun, T., Razzak, A., Islam, M. S., & Uddin, M. S. (2023). Tomato image dataset for maturity classification. *Data in Brief*, 51, 109688.

[27] Sivamani, S., Chon, S. I., Lee, D. H., & Park, J. H. (2022). Importance of adaptive photometric augmentation. *Computers, Materials & Continua*, 72(3).

[28] Rais, K., Amroune, M., Haouam, M. Y., & Benmachiche, A. (2025). Enhancing medical image analysis through geometric and photometric transformations. *arXiv*:2501.13643.

[29] Hu, D. (2020). Survey on attention mechanisms in NLP. In *Intelligent Systems Conference* (pp. 432–448). Springer.

[30] Guo, M. H., Xu, T. X., Liu, J. J., Liu, Z. N., Jiang, P. T., Mu, T. J., & Hu, S. M. (2022). Attention mechanisms in computer vision: A survey. *Computational Visual Media*, 8(3), 331–368.

[31] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv*:1409.1556.

[32] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *CVPR* (pp. 4700–4708).

[33] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *CVPR* (pp. 770–778).

[34] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *CVPR* (pp. 1251–1258).

[35] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., & Adam, H. (2017). MobileNets: Efficient CNNs for mobile vision. *arXiv*:1704.04861.

[36] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In *CVPR* (pp. 4510–4520).