

MobiShuffNetV2: An Efficient Hybrid Deep Learning Framework for Accessible Eye Disease Screening Using Simple Ocular Photography

Mahe Zabin

Human and Digital Interface Department, JW Kim College of Future Studies,
Woosong University, Daejeon, South Korea
Email: mahezabin@wsu.ac.kr

Abstract—Early diagnosis of ocular pathologies is critical to avoiding the loss of vision; nevertheless, existing diagnostic imaging modalities are often based on expensive imaging devices, including fundus cameras or optical coherence tomography scanners, which are often unavailable in rural and low-resource environments. In this study, we introduce MobiShuffNetV2, a powerful hybrid deep-learning model that incorporates depth wise separable convolutions of MobileNetV2 with channel-shuffle components typical of ShuffleNetV2 and is capable of classifying five ocular conditions, including normal, cataract, conjunctivitis, uveitis, and eyelid disorders, without any extra information other than just simple ocular photography. The dataset used in this study was collected from Mendeley and was preprocessed with a domain-specific pipeline that included resizing, grayscale conversion, region-of-interest detection, contrast enhancement, and denoising, followed by targeted augmentation to balance underrepresented classes, resulting in 9861 images. The model achieved a test accuracy of 99.70% (95% confidence interval: [0.9911, 0.9990]) with perfect precision and recall on the categories of Normal and Uveitis, and a Cohen's kappa of 0.9962, which reflects near-perfect interrater agreement. These findings were supported by cross-validation that gave a result of $99.54\% \pm 0.84\%$ accuracy. The proposed architecture exhibited an inference latency of 0.0016 seconds per sample, which equates to an overall latency of 1.61 seconds, thus allowing for real-time usage with end-to-end predictions achieved within a three-second window. Also, Grad-CAM (Class Activation Mapping) visualization supported the accurate localization of pathological areas and thus improved the interpretability of the model. This study proposes a highly effective and scalable pre-screening tool in early diagnosis of ocular disorders by making use of simple ocular photographs and not advanced imaging tools, especially in underserved populations.

Keywords—eye disease classification, medical image analysis, eye disease detection, computer-aided diagnosis

I. INTRODUCTION

Vision impairment and preventable blindness pose significant global health challenges, especially in

low-resource and rural areas where access to advanced ophthalmic care is limited [1]. The most affected regions often lack sufficient equipment and face logistical barriers that make advanced care difficult to reach. According to the World Health Organization, at least 2.2 billion people worldwide experience various forms of vision impairment or blindness, and nearly half of these cases could have been prevented or can still be addressed through timely interventions [2, 3]. Early detection of eye conditions like cataracts, conjunctivitis, uveitis, and eyelid disorders plays a vital role in preventing irreversible vision loss. However, standard imaging techniques such as retinal fundus photography and Optical Coherence Tomography (OCT) rely on expensive equipment, highly trained professionals, and clinical setups that are rarely available in underserved areas [4]. As a result, millions of people miss out on even the most basic screening services, widening the gap in eye care access. Researchers have utilized deep learning models to automate disease screening as a recent measure to address existing limitations [5–7]. They achieved promising results using large-scale retinal and OCT datasets, demonstrating near-expert-level performance in detecting diabetic retinopathy, glaucoma, and age-related macular degeneration [8–10]. However, these models rely on complicated imaging devices, making them unsuitable for low-cost and accessible screening programs despite their success [11, 12]. Additionally, many existing works face method design challenges, such as dependence on single-source datasets, insufficient external validation, lack of explainability, and high computational complexity, which hinder real-time performance on mobile or embedded devices [7, 13]. These challenges highlight the urgent need for methods that balance accuracy, efficiency, and accessibility. One lesser-known area involves using simple ocular photography—pictures taken with low-cost or consumer-grade cameras—to detect and classify eye diseases. Ocular photographs are affordable, require minimal skill, and are accessible in rural outreach programs, telemedicine applications, and

mobile health platforms. Despite their simplicity, few deep learning frameworks have been specifically developed for this modality, creating a significant barrier to developing pre-screening tools for early intervention [14, 15].

We have created MobiShuffNetV2, which is a unique hybrid deep learning design that combines the depthwise-separable convolutions from MobileNetV2 with channel-shuffling and point-wise operations from ShuffleNetV2 to close this gap. Simply by integrating these complementary feature extractors, MobiShuffNetV2 achieves the goal of a small model that is still highly discriminative for use with ocular photographs and requires very little computational overhead. The model is designed to be as efficient as possible for real-time mobile and edge device deployment, making it more suitable for use in low-resource healthcare settings, which have practical constraints. Our method is fundamentally different from the past methods that heavily depend on the use of high-tech retinal imaging. In fact, the framework we have developed is a demonstration of how more accessible ocular photography can be a reliable input modality for multi-class eye disease detection.

The contributions of this study are fourfold:

- **Accessible Modality:** We unveil an innovative deep learning system that uses only basic eye photos to detect eye diseases, thus minimizing the need for expensive fundus or OCT imaging.
- **Hybrid Architecture:** We propose MobiShuffNetV2, a novel architecture that integrates MobileNetV2 and ShuffleNetV2, thereby achieving a balance between accuracy, productivity, and lightweight deployment.
- **Comprehensive Evaluation:** We prove our model in different scenarios- training, validation, cross-validation showing a high degree of generalization.
- **Real-Time and Explainable AI Deployment:** We demonstrate on-the-fly prediction (1.6 ms per sample) with Grad-CAM attributions that point out the affected areas by the disease, thus, facilitating interpretability and clinical trust.
- **Benchmark Contribution:** The main contribution of this work is to fill a significant knowledge gap by establishing a benchmark for eye disease detection using basic eye photographs, an approach that has been relatively underexplored compared with conventional fundus- and OCT-based imaging methods.

Our work, which is made possible by these contributions, tackles both the accuracy and the accessibility challenges in the screening of eye diseases. Our work emphasizes a method that is widely applicable and does not require special equipment. This makes MobiShuffNetV2 a potential pre-screening tool for community health programs, teleophthalmology platforms, and clinical settings with limited resources.

The rest of the paper is structured as follows: in Section II, we explore previous works that utilized fundus and OCT-based methods, highlighting their limitations in

terms of generalizability and accessibility. Section III details the techniques we employed, including our preprocessing approach, augmentation methods, and architecture design. In Section IV, we present the experimental results and compare them with state-of-the-art baselines. Additionally, Section V discusses interpretability, resourcefulness, and external validation. Finally, Section VI summarizes the paper and suggests potential directions for accessible ophthalmic AI-based future research.

II. LITERATURE REVIEW

In recent years, Machine Learning (ML) and Deep Learning (DL) have rapidly transformed medical image analysis by enabling automated, accurate, and scalable disease diagnosis across diverse clinical domains, ranging from cancer detection and histopathological image interpretation to dermatological, neurological, and ophthalmic disease screening [16–20]. The increasing availability of medical imaging datasets, together with advances in Convolutional Neural Networks (CNNs), vision transformers, lightweight architectures, and explainable artificial intelligence, has encouraged researchers to develop intelligent diagnostic systems that can support clinicians in early disease detection and decision-making. In ophthalmology, these developments are particularly important because many eye diseases, such as cataract, glaucoma, diabetic retinopathy, and other retinal or ocular abnormalities, may progress silently and lead to irreversible vision loss if not detected at an early stage. Consequently, DL-based eye disease classification has gained significant research attention as a promising approach for improving screening accessibility, reducing diagnostic burden, and enabling cost-effective deployment in resource-limited healthcare environments.

A. Fundus-Based Eye Disease Diagnosis

Recent advances in deep learning have significantly improved the detection and classification of eye diseases using retinal fundus images, achieving notable accuracy and efficiency. Arnob *et al.* [21] proposed a lightweight CNN for multiclass retinal disease screening, addressing computational efficiency, model interpretability, and class imbalance. Utilizing a dataset of 5335 images with Synthetic Minority Over-sampling Technique (SMOTE) for class balancing, their model achieved 87.9% accuracy. However, the reliance on a single-source dataset, the lack of cross-validation, the absence of severity grading, and the limited evaluation of explainability hinder clinical applicability. Alsohemi and Dardouri [22] implemented EfficientNetB3 for retinal disease classification. Their approach, based on standard pre-processing, data augmentation, and transfer learning, achieved 95.12% accuracy. While demonstrating robust performance, the method lacked methodological innovation, external validation on independent datasets, and comprehensive explainability analysis, which limited its real-world deployment. Ensemble-based methods have been explored to improve accuracy and robustness. Yu and Dong [23] developed RetinaDNet, integrating dual-branch

inputs (fundus and vessel segmentation), six backbone networks, and a soft voting ensemble. The system achieved 99.2% accuracy on diabetic retinopathy and 98.8% on general retinal disease classification. Despite excellent performance, it suffers from high computational complexity, dependency on accurate vessel segmentation, and limited external validation. Hybrid architectures combining CNNs with handcrafted features were proposed by Shamsan *et al.* [24], fusing MobileNet and DenseNet121 with Principal Component Analysis (PCA), achieving 98.5% accuracy and 99.23% Area Under the Curve (AUC). Although effective, dataset specificity and computational overhead constrain generalizability. Similarly, Imaduddin and Sakina [25] evaluated multiple pre-trained networks, reporting an accuracy of up to 96.2% with DenseNet-121; however, external validation remained lacking. Multi-label approaches for the simultaneous detection of numerous ocular diseases have gained attention.

Al-Fahdawi *et al.* [26] introduced Fundus-DeepNet, which integrates HRNet, attention mechanisms, SENet blocks, and bilateral fundus fusion, achieving F1-Scores above 88% and an AUC of over 99.7%. Sait [27] combined denoising autoencoders, whale optimization, and optimized ShuffleNetV2, reaching 99.1–99.4% accuracy. Despite their technical innovations, both methods face challenges in dataset imbalance and real-world deployment. Ensemble feature fusion strategies were applied by Jeny *et al.* [28], combining a 20-layer CNN with VGG16, ResNet50, and DenseNet201, achieving 96.96% accuracy. Nguyen *et al.* [29] and Shyamalee *et al.* [30] employed ultra-wide-field fundus imaging and attention-enhanced frameworks, reporting high classification accuracy while highlighting challenges in multi-modal integration, real-time validation, and disease progression assessment. Vadduri and Kuppusamy developed comprehensive pipelines for multi-class diabetic eye disease classification [31], incorporating advanced preprocessing and segmentation techniques, achieving up to 98.33% accuracy for diabetic retinopathy. Ejaz *et al.* [32] utilized multiple CNN architectures on the RFMiD dataset, achieving 90.34% validation accuracy, but deeper models exhibited overfitting and single-label classification limited clinical applicability. Focused studies on specific conditions, such as Abd El-Khalek *et al.* [33], reported high accuracy for age-related macular degeneration (96.85%) using machine learning-based classifiers, yet clinical validation was limited. Wang *et al.* [34] proposed MBSaNet, a hybrid CNN-self-attention network for multi-label fundus disease classification, achieving 88.1% accuracy with 9.4 M parameters. While demonstrating efficient local and global feature learning, challenges remain for underrepresented disease categories and optimizing the architecture.

B. OCT-Based Eye Disease Diagnosis

Deep learning applied to OCT imaging has advanced auto-mated retinal disease detection, progressing from traditional CNNs to attention and ensemble-based architectures. Early work by Han *et al.* [35] demonstrated the potential of VGG and ResNet models with transfer

learning and mix-up augmentation on 6063 images, achieving 99.7% accuracy for three-class and 91.1% for six-class classification. This study highlighted the advantage of deep learning over human experts but was limited by single-device data and a lack of cross-validation.

Building on multi-stage approaches, Chen *et al.* [36] combined ensemble binary classifiers with YOLOv3-based pathology detection using 37,138 OCT images. By integrating multiple backbone networks and utilizing Grad-CAM for interpretability, the system achieved an accuracy of over 99%. While demonstrating the benefit of ensemble and multi-task strategies, the study remained constrained by single-center data and rectangular labeling.

To explore foundation models, Du *et al.* [37] pretrained RETFound on 1.6 million unlabeled OCT scans and fine-tuned it on 1770 labeled B-scans to detect subretinal and intraretinal fluid, drusen, and pigment epithelial detachment. Compared to ResNet-50, RETFound achieved an average AUC-ROC of 0.80, highlighting the promise of large-scale pretraining for feature-rich OCT representations. However, small labeled datasets and limited pathology diversity restricted the broader applicability.

Specialized detection of retinal signs was addressed by Inferred *et al.* [38], who trained nine VGG-16 models on 10,770 OCT scans to identify abnormalities such as drusen, macular holes, and vitreomacular adhesion, achieving 93–99% accuracy. This work emphasized the importance of multi-label and sign-specific detection, which contrasts with general disease classification approaches.

Further refinement of standard CNNs was demonstrated by Jaimes *et al.* [39], who used a modified VGG16 on 33,792 Kermany images, optimizing hyperparameters and performing 10-fold Monte Carlo cross-validation, to achieve 95.19% accuracy. The study demonstrated that careful optimization and cross-validation can enhance generalizability, although dataset limitations remained a concern.

Hybrid architectures were explored by Gencer and Gencer [40], combining EfficientNetB0 and Xception with Squeeze-and-Excitation blocks to enhance feature representation, reaching over 99% accuracy on UCSD and Duke datasets. Similarly, Puneet *et al.* [41] integrated attention mechanisms into VGG16 for 84,484 OCT images, achieving 95.6% testing accuracy. Both studies highlighted the advantage of architectural enhancements in improving OCT classification performance, while also exposing the need for broader clinical validation.

Attention-based networks were further advanced by Talaat *et al.* [42] with OCTNet, an attention-augmented DenseNet trained on 84,495 high-resolution images, reaching 97.92% validation accuracy. The model emphasized relevant OCT features and utilized extensive data augmentation, demonstrating the potential of attention mechanisms to improve performance, though XAI and multi-center validation were still lacking.

Elkholy and Marzouk [43] similarly applied VGG16 with preprocessing enhancements on 35,468 OCT images, improving accuracy from 94% to 97%. Their work reinforced the benefit of image quality improvement and fine-tuning, while still lacking explainable AI and cross-validation. Ai *et al.* [44] extended this idea to Fusion Networks (FN-OCT) combining multiple Inception variants with attention modules, achieving 98.7% accuracy and 99.1% AUC on 108,312 images, illustrating the power of ensemble and attention-based fusion strategies. Building on ensemble and hybrid methods for fundus-based classification, Khalid *et al.* [45] developed a novel system named Computer-Aided Diagnosis (CAD)-EYE, a feature fusion system combining MobileNet and EfficientNet with a novel Fluorescence Imaging Simulation preprocessing algorithm. Evaluated on 65,871 fundus images, the system achieved 98% accuracy in classifying normal, diabetic retinopathy, hypertensive retinopathy, glaucoma, and contrast-related eye diseases, outperforming individual CNN models. Despite its strong performance, the study lacks explainable AI techniques, cross-validation, and comprehensive multi-center clinical validation, highlighting persistent challenges in generalizability and interpretability for real-world deployment. Khalid *et al.* [45] extended fundus-based classification with CAD-EYE, a feature fusion system that combines MobileNet and EfficientNet, utilizing Fluorescence Imaging Simulation preprocessing. Tested on 65,871 fundus images, it achieved 98% accuracy across multiple eye diseases. However, the study lacks explainable AI, cross-validation, and multi-center clinical validation, limiting generalizability. Finally, hyperparameter tuning was shown to optimize performance further. Shamsan *et al.* [24] compared six CNN architectures on the OCT-C8 dataset, with DenseNet121 achieving 97.79% accuracy for eight-class classification. While outperforming previous models, the study highlighted persistent gaps in cross-validation, external validation, and clinical interpretability.

C. Research Gaps

Although significant advancements have been made in the application of deep learning to detect eye diseases, several limitations are important to note. Most studies base their results on single-source or single-device data, thereby limiting their ability to generalize to a wide range of populations and imaging environments. Validation procedures are frequently suboptimal, and cross-validation, external validation, and multi-center validation are often limited or absent, which reduces confidence in clinical applicability. Explainable AI, such as Grad-Cam, is used inconsistently, and clinical interpretability is often not readily available. The computational cost and model complexity further restrict real-time deployment, particularly in resource-limited settings. In addition, there are minimal research works that involve expert validation, thus limiting the evaluations of practical utility. These gaps should be addressed through

effective validation procedures, explainable AI integration, model efficiency optimization, and expert review to transform those methods into practical, reliable clinical tools.

In addition, most studies are based on OCT- and fundus-based methods of detection, which involve specialised and costly equipment that is not available in rural or low-resource settings. This generates a practical gap: very few pre-screening tools can use simple eye images to detect and thus prevent early, making it challenging to detect eye diseases in populations who do not have access to high-end imaging equipment.

Based on this, our research aims to address this gap by developing a pre-screening instrument that utilizes simple eye images to detect eye diseases at an early stage, thereby improving access to eye disease screening in rural and low-resource areas.

III. METHODOLOGY

In this study, we propose a systematic approach for eye disease detection, integrating advanced preprocessing, data augmentation, and hybrid deep learning architectures. The workflow is designed to maximize feature extraction efficiency while maintaining high computational performance, making it suitable for both research and real-time clinical applications. Fig. 1 illustrates the overall methodology, highlighting each critical stage from the preparation of the input image to the final classification.

A. Data Acquisition

The dataset used in this research is based on the Mendeley repository [46], which includes images labeled in five ophthalmic conditions: Eyelid disease, Normal, Cataract, Uveitis, and Conjunctivitis. The conditions are commonly found in clinical contexts, and the dataset is maintained to support the training and development of machine learning models to classify ocular diseases using visual features. The dataset includes photographs illustrating different phases and forms of these ocular illnesses with each picture designated to one of the five diagnostic groups. They are annotated images used in the training and testing of computational models that attempt to identify and classify ocular pathologies based on visual analysis.

Fig. 2 illustrates the distribution of images per five classes of diseases.

This dataset contains a substantial amount of visually labeled data for each type of disease, enabling the analysis of eye disease classification based on image characteristics. Its various collection of images, covering different stages of severity in each disease, assists in training robust machine learning algorithms to allow them to diagnose eye diseases based on medical images.

To further illustrate this, Fig. 3 displays sample images of each class, visually representing the various conditions included in the data and highlighting the versatility within each disease category.

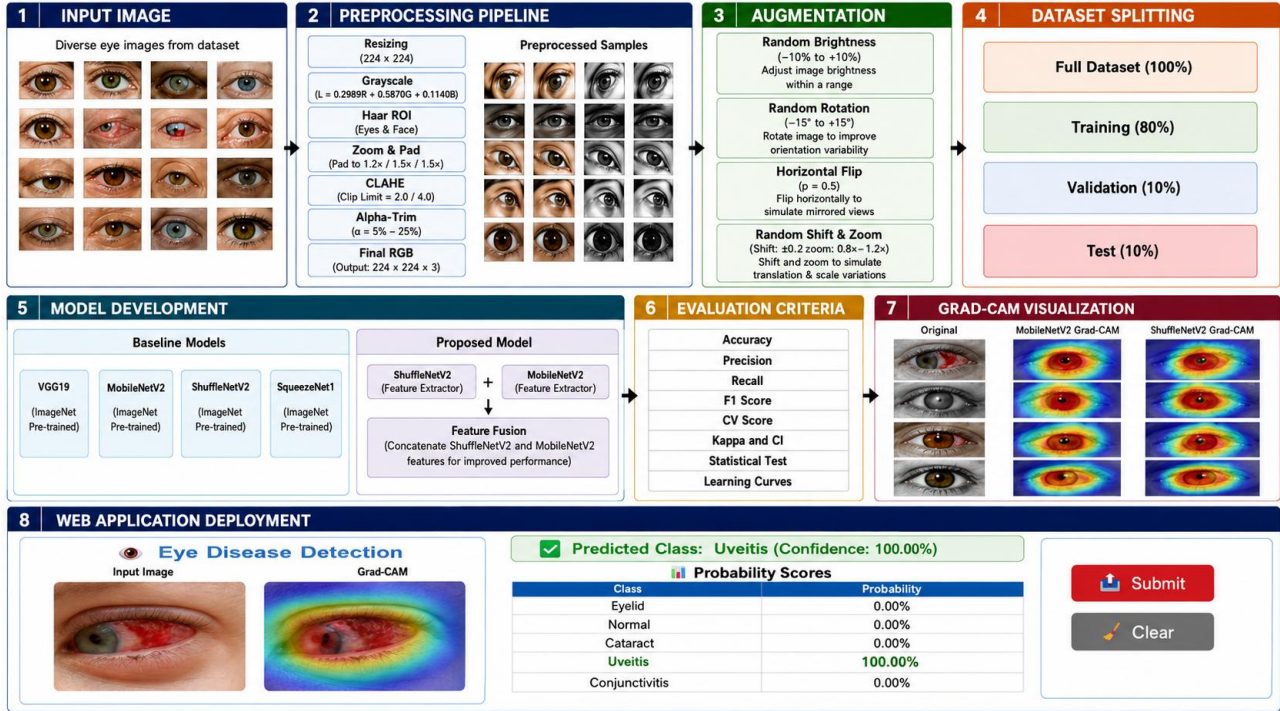


Fig. 1. Overview of the proposed workflow for eye disease detection.

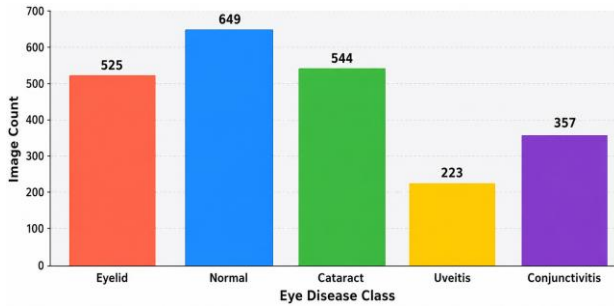


Fig. 2. Distribution of image counts per eye disease class in the Mendley dataset.



Fig. 3. Sample images from each eye disease class, illustrating the variations in visual characteristics across different conditions in the dataset.

B. Image Processing and Feature Enhancement

A structured preprocessing pipeline was used to prepare ocular images for accurate disease classification. To ensure a fixed input size for the deep learning models and reduce computational variation during training, all images were first resized to 224×224 pixels. The images were then converted to grayscale to minimize unnecessary complexity in the color channels and enable the model to focus on intensity-based visual patterns, such as edges, inflammatory regions, opacity, texture changes, and contrast differences. This step also re-duces preprocessing cost before ROI extraction.

After grayscale conversion, the Region of Interest (ROI) was detected using Haar Cascade classifiers. The detector first searched for the eye region, as the eyes contain the most disease-relevant visual features. When the eye region was detected, the corresponding bounding box was extracted as the ROI. If no eye was detected, the system switched to face detection as a fallback strategy. If neither the eye nor the face was found, the entire image was retained as the ROI to prevent the loss of valuable samples. The detected ROI was then zoomed using a factor of 1.3 and padded to preserve the main ocular region while maintaining contextual information.

Next, Contrast-Limited Adaptive Histogram Equalization (CLAHE) was applied to improve local contrast and enhance subtle disease-related patterns that may not be clearly visible in the original image. Since important structural information must be preserved while reducing extreme pixel variations, noise reduction was also performed using an alpha-trimmed mean filter. Finally, the image was converted back to RGB format to ensure compatibility with pretrained deep learning models designed for three-channel images. In summary,

the preprocessing approach ensures consistency in image content, highlights key features for disease diagnosis, and supports effective model training. Table I summarizes the

preprocessing steps applied to the image dataset, detailing the specific operations performed and their respective parameters.

TABLE I. PREPROCESSING STEPS, DESCRIPTIONS, AND PARAMETERS

Step	Description	Parameter
Resizing	Resizes the image to 224×224 pixels for uniformity and efficient processing	224×224 pixels
Grayscale Conversion	Converts the image to grayscale to focus on intensity variations and reduce data size	-
Haar Cascade ROI Detection	Detects relevant features (eyes or face) using Haar Cascade classifiers	Eyes first, fallback to face detection
Zoom and Padding of ROI	Zooms into the detected ROI with a factor of 1.3, ensuring contextual information is retained	$zoom_factor = 1.3$
CLAHE (Contrast Limited Adaptive Histogram Equalization)	Enhances the contrast of the image for better feature visibility	$clipLimit = 2.0$, $tileGridSize = (8, 8)$
Alpha-Trimmed Mean Filter	Reduces noise by removing extreme pixel values within a kernel	$kernel_size = 3 \times 3$, $alpha = 2$
Final RGB Conversion	Converts the image back to RGB format for consistency with model inputs	-

Additionally, Fig. 4 includes examples of processed im-ages from all classes, demonstrating how the preprocessing pipeline impacts the dataset. These pictures illustrate how the preprocessing procedures have optimized key characteristics and ensured the consistency of the data set.



Fig. 4. Sample processed images from each class, demonstrating the effects of the preprocessing pipeline on the dataset.

C. Augmenting Images to Diversify Classes

Data augmentation is an important step that introduces greater diversity into a small image database by applying a set of transformations to the original images. This helps the model become more generalizable, less sensitive to different visual conditions, and less prone to overfitting. For the current dataset, a series of augmentation operations were performed on the images, including random brightness changes, rotation, horizontal flipping, translation, and zooming. Such changes introduce variability in the visual information, allowing the model to learn a wider range of patterns, which is especially relevant in medical image-classification applications. The number of samples in some categories, particularly Uveitis, was lower, as illustrated in Fig. 2. Targeted

augmentation was used to reduce this imbalance and increase the number of samples in underrepresented classes. Specifically, Uveitis images were multiplied by 6, and images from other disease classes were adjusted accordingly. This sampling approach enabled every class to be well represented and minimized the initial bias toward dominant classes. Table II shows the total number of images in the dataset after augmentation.

TABLE II. CLASS-WISE IMAGE DISTRIBUTION AFTER TARGETED AUGMENTATION

Class Name	Number of Images
Eyelid	2100
Normal	2596
Cataract	2176
Uveitis	1561
Conjunctivitis	1428
Total	9861

These enhancements brought the number of images in the dataset to 9861. This augmentation is essential for enhancing the model's capability to generalize, as it subjects the network to diverse lighting conditions, orientations, and scales. To illustrate these augmentation steps, along with the corresponding descriptors and parameter values, Table III summarizes the augmentation steps, descriptions, and parameter settings.

D. Dataset Splitting

Once the images were augmented, they were divided into training, validation, and testing datasets in the ratios of 80%, 10%, and 10%, respectively. The 80-10-10 split is widely used in machine learning, as it provides a reasonable size of training data and a large enough amount of data to validate and test the model. To provide a representative distribution of the sets, the images were randomly shuffled before the split. The distribution of images by class across each split is presented in the Fig. 5. The training set is slightly imbalanced, but the imbalance does not significantly impact the model's ability to generalize. For example, in an augmented image scenario, the "Uveitis" class had fewer original images, which were

represented in a greater proportion of augmented images in the training data to ensure sufficient representation. This slight imbalance is unlikely to significantly hamper the

model's performance, as the overall distribution is relatively balanced.

TABLE III. DATA AUGMENTATION STEPS, DESCRIPTIONS, AND PARAMETERS

Augmentation Step	Description	Parameter
Random Brightness	Adjusts the brightness of the image by a random factor to simulate lighting variations.	Random factor between 0.7 and 1.3
Random Rotation	Rotates the image within a specified range to introduce orientation variability.	Angle range: -15 to 15 degrees
Horizontal Flip	Flips the image horizontally to simulate mirrored views.	Flip mode: Horizontal (1)
Random Shift and Zoom	Shifts and zooms the image to simulate translations and scale variations	Shift ratio: 0.1, Zoom range: 0.1

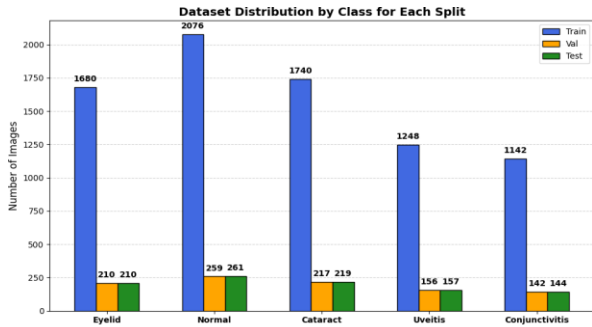


Fig. 5. Distribution of images by class across the training, validation, and test splits. The figure demonstrates the class distribution after applying the data split.

E. Dataset Leakage Analysis

A thorough similarity analysis was carried out to ensure proper partitioning of the dataset and to assess possible information leakage between the training, validation, and test sets. A dual-hashing approach, using MD5 hashing to identify exact duplicates and perceptual hashing (pHash) to identify visually similar images, was used in the analysis. This combination allows strong detection of identical and similar images that may result from augmentation or repeated acquisition of similar frames.

As shown in Table IV, no images matched exactly between any dataset splits, indicating that there was no direct data leakage due to the partitioning strategy. However, the training set still showed moderate similarity to the validation and test sets, as indicated by perceptual hashing. This is common in augmentation-based pipelines, where transformed samples share structural and semantic properties with the original images. However, the similarity between the validation and test sets was relatively low, suggesting good separation between the evaluation sets.

TABLE IV. CROSS-SPLIT DATASET LEAKAGE ANALYSIS USING MD5 AND PERCEPTUAL HASHING

Comparison	Exact Dup	Exact (%)	Near Dup	Near (%)
Train vs Val	0	0.00	187	19.00
Train vs Test	0	0.00	191	19.27
Val vs Test	0	0.00	26	2.64

In addition, the within-split duplication analysis provided in Table V indicates good data consistency, with almost no internal duplication found in any subset of the

data. A single duplicate was found in the training dataset, corresponding to 0.013% duplication, but none were detected in the validation and test datasets. Overall, the dataset is free from exact duplication, with only perceptual similarities introduced by augmentation operations.

TABLE V. WITHIN-SPLIT DUPLICATE IMAGE ANALYSIS

Split	Self Duplicates	Self Dup (%)
Train	1	0.013
Val	0	0.000
Test	0	0.000

F. Baseline Classifiers: VGG19, MobileNetV2, ShuffleNetV2, SqueezeNet1

The need to develop efficient and lightweight models, particularly where real-time processing is required, such as on resource-limited devices like smartphones, IoT devices, and embedded systems, has increased significantly in recent years. In response to these issues, the current work employs state-of-the-art classifiers, which offer a trade-off between performance and precision. The models considered are the well-known VGG19, MobileNetV2, ShuffleNet V2, and SqueezeNet1, which are optimized to achieve high performance with minimal memory usage and computational complexity. Although such architectures are ideal in low resource consumption, they represent a variety of deep neural network optimisation strategies suitable in low-resource environments. Each model, along with its salient architectural characteristics, memory-consumption properties, and relative performance in terms of speed and accuracy, is described in detail, allowing for an assessment of the suitability of each model for on-device workloads in the real world.

1) VGG19

VGG19 network is one of the most potent deep-learning models used in various image-classification tasks, such as the detection of ocular diseases. It consists of nineteen layers with sixteen convolutional and three fully connected. The convolutional layers are organized into five separate blocks, separated by max-pooling layers that discard the spatial resolution of the feature maps while retaining salient information. Such architectural design enables VGG19 to respond to both low- and high-level visual features, which becomes critical when the task of identifying eye diseases is subtle. Based on the current

empirical research, it can be stated that VGG19 is an effective classification system for eye disease images since it finds fine-grained visual representations. The model has an input size of up to 0.60 MB, and its memory consumption for forward and backward propagation is 118.88 MB. The total number of parameters is 558.36 MB, resulting in a total model size of 677.85 MB. This large memory footprint can be attributed to the network's ability to learn non-linear visual patterns, thereby supporting its high disease-detecting performance.

Table VI tabulates the specific layer specifications of the VGG19 architecture used in detecting ocular diseases, listing the types of layers, the dimensions of the outputs, and the number of parameters per layer. The table does not present the intermediate computations (e.g., the activation functions) or the memory requirements per layer; it is preferred to illustrate the overall architecture and computational requirements of the model. It therefore provides a concise yet comprehensive overview of how the input data is modified using multiple convolutional, pooling, and fully connected layers.

TABLE VI. VGG19 KEY LAYERS FOR EYE DISEASE DETECTION

Layer (type:depth-idx)	Output Shape	Number of Parameter
Conv2D: 2-1	[1, 64, 224, 224]	1792
Conv2D: 2-3	[1, 64, 224, 224]	36,928
MaxPool2D: 2-5	[1, 64, 112, 112]	-
Conv2D: 2-6	[1, 128, 112, 112]	73,856
Conv2D: 2-8	[1, 128, 112, 112]	147,584
MaxPool2D: 2-10	[1, 128, 56, 56]	-
Conv2D: 2-11	[1, 256, 56, 56]	295,168
Conv2D: 2-13	[1, 256, 56, 56]	590,080
MaxPool2D: 2-19	[1, 256, 28, 28]	-
Conv2D: 2-20	[1, 512, 28, 28]	1,180,160
Conv2D: 2-22	[1, 512, 28, 28]	2,359,808
MaxPool2D: 2-28	[1, 512, 14, 14]	-
Conv2D: 2-29	[1, 512, 14, 14]	2,359,808
MaxPool2D: 2-37	[1, 512, 7, 7]	-
AdaptiveAvgPool2D: 1-2	[1, 512, 7, 7]	-
Linear: 2-38	[1, 4096]	102,764,544
Dropout: 2-40	[1, 4096]	-
Linear: 2-41	[1, 4096]	16,781,312
Dropout: 2-43	[1, 4096]	-
Linear: 2-44	[1, 5]	20,485

2) MobileNetv2

MobileNetV2 architecture is a model of resource efficient and lightweight deep-learning network capable of showing strong performance on devices with resource constraints. It utilizes depth wise separable convolutions and an inverted residual architecture, which introduces a significantly lower number of parameters without compromising classification accuracy. These architecture decisions make MobileNetV2 especially appropriate when the task at hand is computationally limited, e.g. real-time ocular disease detection. MobileNetV2, used in the current work, has proven its effectiveness in detecting eye diseases, demonstrating the computational efficiency and

accuracy of the model. The input size of the model is 19.27 MB; on inference it uses an average of 3419.19 MB of memory, and on training it uses a similar amount. The parameter subset is 8.92 MB, but the total model size is approximated at 3447.38 MB. This efficient memory footprint can be largely attributed to the use of depth wise separable convolutions, which enhance model performance with minimal impact on computational complexity.

The pre-trained MobileNetV2 model has 53 layers, and it provides a compact, but powerful model, which can be deployed in real-life scenarios where resources are scarce. The major layers are further broken down into finer details, as shown below in the context of eye disease detection. The following Table VII shows the key layers of the MobileNetV2 architecture for eye disease detection.

TABLE VII. MOBILENETV2 KEY LAYERS FOR EYE DISEASE DETECTION

Layer (type:depth-idx)	Output Shape	Number of Parameter
MobileNetV2	[32, 5]	-
Conv2dNormActivation: 2-1	[32, 32, 112, 112]	864
InvertedResidual: 2-2	[32, 16, 112, 112]	896
InvertedResidual: 2-3	[32, 24, 56, 56]	5136
InvertedResidual: 2-4	[32, 24, 56, 56]	8832
InvertedResidual: 2-5	[32, 32, 28, 28]	10,000
InvertedResidual: 2-6	[32, 32, 28, 28]	14,848
InvertedResidual: 2-7	[32, 32, 28, 28]	14,848
InvertedResidual: 2-8	[32, 64, 14, 14]	21,056
InvertedResidual: 2-9	[32, 64, 14, 14]	54,272
InvertedResidual: 2-10	[32, 64, 14, 14]	54,272
InvertedResidual: 2-11	[32, 64, 14, 14]	54,272
InvertedResidual: 2-12	[32, 96, 14, 14]	66,624
InvertedResidual: 2-13	[32, 96, 14, 14]	118,272
InvertedResidual: 2-14	[32, 96, 14, 14]	118,272
InvertedResidual: 2-15	[32, 160, 7, 7]	155,264
InvertedResidual: 2-16	[32, 160, 7, 7]	320,000
InvertedResidual: 2-17	[32, 160, 7, 7]	320,000
InvertedResidual: 2-18	[32, 320, 7, 7]	473,920
Conv2dNormActivation: 2-19	[32, 1280, 7, 7]	409,600
Dropout: 2-20	[32, 1280]	-
Linear: 2-21	[32, 5]	6405

3) ShuffleNet_V2

The ShuffleNetV2 is a lightweight, mobile, and embedded cognitive design for deep neural architectures. It utilizes channel-shuffle operations and point-wise group convolution to reduce the number of calculations, while maintaining significant performance. Such optimizations make ShuffleNetV2 applicable to real-time use in resource-constrained applications, including ophthalmic disease detection. The current paper shows that ShuffleNetV2 has a desirable trade-off between performance and precision in the analysis of eye-diseases. The model uses an input tensor of 0.60 MB; during forward and backward propagation, the memory peak of the model increases to 31.20 MB. The trainable parameters have a size of 5.03 MB, which results in a projected on-disk footprint of 36.84 MB. This improved

computational efficiency is primarily due to the combination of point-wise convolution and channel-shuffle mechanisms, which significantly reduce the computational burden without compromising accuracy.

The ShuffleNetV2 pretrained model has fifty layers, and this depth has been determined to provide an optimum balance between model capacity and edge device deployability. Its architecture can be outlined by a brief overview of the key layers that are most important for detecting eye diseases. The main convolutional and pooling stages that the backbone of ShuffleNetV2 consists of are listed in Table VIII.

TABLE VIII. SHUFFLENETV2 KEY LAYERS FOR EYE DISEASE DETECTION

Layer (type:depth-idx)	Output Shape	Number of Parameter
ShuffleNetV2	[1, 5]	-
Conv2d: 2-1	[1, 24, 112, 112]	648
BatchNorm2d: 2-2	[1, 24, 112, 112]	48
ReLU: 2-3	[1, 24, 112, 112]	-
MaxPool2d: 1-2	[1, 24, 56, 56]	-
InvertedResidual: 2-4	[1, 116, 28, 28]	1772
InvertedResidual: 2-5	[1, 116, 28, 28]	5626
InvertedResidual: 2-6	[1, 116, 28, 28]	7598
InvertedResidual: 2-7	[1, 116, 28, 28]	7598
InvertedResidual: 2-8	[1, 232, 14, 14]	14,964
InvertedResidual: 2-9	[1, 232, 14, 14]	28,652
InvertedResidual: 2-10	[1, 232, 14, 14]	28,652
InvertedResidual: 2-11	[1, 232, 14, 14]	28,652
InvertedResidual: 2-12	[1, 232, 14, 14]	28,652
InvertedResidual: 2-13	[1, 232, 14, 14]	28,652
InvertedResidual: 2-14	[1, 232, 14, 14]	28,652
InvertedResidual: 2-15	[1, 232, 14, 14]	28,652
InvertedResidual: 2-16	[1, 464, 7, 7]	56,840
InvertedResidual: 2-17	[1, 464, 7, 7]	111,128
InvertedResidual: 2-18	[1, 464, 7, 7]	111,128
InvertedResidual: 2-19	[1, 464, 7, 7]	111,128
Conv2d: 2-20	[1, 1024, 7, 7]	475,136
BatchNorm2d: 2-21	[1, 1024, 7, 7]	2048
ReLU: 2-22	[1, 1024, 7, 7]	-
Linear: 1-7	[1, 5]	5125

4) SqueezeNet1

SqueezeNet1 is a compact and straight-forward deep learning model optimized for use in resource-constrained settings, making it particularly suitable for mobile devices and embedded systems. The design includes a Fire module that consists of a squeeze layer (1×1 convolutions), and an expand layer (1×1 convolution and 3×3 convolution) that significantly decreases the number of parameters but does not lead to a decrease in performance. This is achieved by implementing a few highly expressive parameters that enable the network to operate with a minimal computational load. In the current research, SqueezeNet1 will be utilized to detect eye diseases, where its lightweight nature makes it effective in carrying

out detection on low-resource devices. The model input size is 0.60 MB, and the forward/backward pass size is 19.37 MB. The parameter size measures 2.90 MB, with an estimated firm size of 22.87 MB. The model's efficiency is attributed to its smaller number of parameters and the reduced computational requirements of the Fire module, without compromising accuracy.

The trained SqueezeNet1 instance consists of 58 layers and is appropriate for implementation in resource-intensive environments in real-time. The key layers used in the SqueezeNet1 architecture to detect eye diseases are briefly described in Table IX.

TABLE IX. KEY LAYERS IN SQUEEZENET1 FOR EYE DISEASE DETECTION

Layer (type:depth-idx)	Output Shape	Param #
Conv2d: 2-1	[1, 64, 111, 111]	1792
Fire: 2-4	[1, 128, 55, 55]	-
Conv2d: 3-1	[1, 16, 55, 55]	1040
Conv2d: 3-3	[1, 64, 55, 55]	1088
Fire: 2-7	[1, 256, 27, 27]	-
Conv2d: 3-13	[1, 32, 27, 27]	4128
Conv2d: 3-15	[1, 128, 27, 27]	4224
Fire: 2-11	[1, 384, 13, 13]	-
Conv2d: 3-25	[1, 48, 13, 13]	12,336
Conv2d: 3-27	[1, 192, 13, 13]	9408
Fire: 2-13	[1, 512, 13, 13]	-
Conv2d: 3-37	[1, 64, 13, 13]	24,640
Conv2d: 3-41	[1, 256, 13, 13]	147,712
Dropout: 2-14	[1, 512, 13, 13]	-
Conv2d: 2-15	[1, 5, 13, 13]	2565
AdaptiveAvgPool2d: 2-17	[1, 5, 1, 1]	-

TABLE X. KEY LAYERS IN MOBISHUFFLENETV2 FOR EYE DISEASE DETECTION

Layer (type:depth-idx)	Output Shape	Param #
ShuffleNetFeatureExtractor	[1, 1024]	-
Sequential (Block 1)	[1, 24, 112, 112]	696
MaxPool2d	[1, 24, 56, 56]	3
Sequential (Block 2)	[1, 116, 28, 28]	30,192
Sequential (Block 3)	[1, 232, 14, 14]	244,180
Sequential (Block 4)	[1, 464, 7, 7]	501,352
Sequential (Block 5)	[1, 1024, 7, 7]	477,184
MobileNetFeatureExtractor	[1, 1280]	-
Conv2dNormActivation	[1, 32, 112, 112]	928
InvertedResidual (Block 1)	[1, 16, 112, 112]	896
InvertedResidual (Block 2)	[1, 24, 56, 56]	5136
InvertedResidual (Block 3)	[1, 32, 28, 28]	10,000
InvertedResidual (Block 4)	[1, 64, 14, 14]	21,056
Conv2dNormActivation	[1, 1280, 7, 7]	412,160
FusionModel	[1, 5]	-
BatchNorm1d (ShuffleNet)	[1, 1024]	2048
BatchNorm1d (MobileNet)	[1, 1280]	2560
Linear	[1, 512]	1,180,160
ReLU	[1, 512]	-
Dropout	[1, 512]	-
Linear	[1, 5]	2565

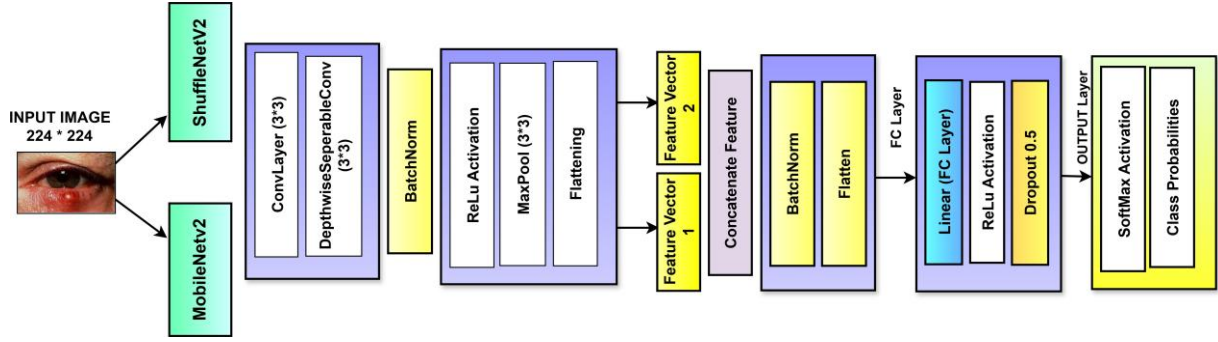


Fig. 6. Architecture of proposed MobiShuffNetV2.

G. Proposed MobiShuffNetV2: A Hybrid Fusion of ShuffleNetV2 and MobileNetV2 for Eye Disease Detection

The proposed MobiShuffNetV2 framework is an ensemble-based deep learning framework comprising the two best baseline classifiers, namely MobileNetV2 and ShuffleNetV2, which are fused using a soft voting strategy. The proposed method fuses the probabilistic outputs from both pretrained models so that classification robustness and diagnostic accuracy are enhanced instead of performing direct feature-level fusion. This ensemble approach enables the model to harness the complementary strengths of both lightweight CNN models while ensuring computational efficiency, which is crucial for resource-constrained applications. MobileNetV2 provides efficient depthwise separable convolutions and inverted residual blocks, which help extract deep semantic representations with low computational complexity. In contrast, ShuffleNetV2 uses channel shuffling and pointwise group convolution to achieve faster inference and lightweight feature extraction. These architectures are combined through a soft voting mechanism, resulting in better representational capability and efficient inference time. During inference, each model independently predicts class probability distributions for the given eye image. After generating probability scores from MobileNetV2 and ShuffleNetV2, the final probability score is computed as the mean of both, and the class with the highest probability score is selected as the final class. This probabilistic decision-level fusion helps enhance model generalization and minimize the chance of individual model bias and misclassification. The proposed MobiShuffNetV2 framework provides an effective compromise between accuracy and computational cost, making it suitable for real-time eye disease diagnosis on low-resource devices. The architecture processes input images with low memory requirements while maintaining high classification performance. The parameter size of the proposed model is 18.66 MB, and the total size of the proposed model is estimated to be 157.34 MB. Additionally, the memory consumption of the forward and backward passes is around 138.08 MB, demonstrating the lightweight nature of the framework. The proposed soft voting ensemble is shown to be appropriate for practical, time-sensitive medical image-classification applications due to the properties.

The key layers of the proposed architecture are listed in Table X. The proposed overall network structure is shown in Fig. 6 and Algorithm 1 outlines the training and evaluation processes of the model.

Algorithm 1: Proposed MobiShuffNetV2 Algorithm

Input: Image $I \in \mathbb{R}^{3 \times H \times W}$

Preprocessing: Normalize the input image:

$$I_{\text{norm}} = \text{Normalize}(I)$$

Feature Extraction: Extract features using ShuffleNetV2 and MobileNetV2:

$$\begin{aligned} x_{\text{shuffle}} &= \text{ShuffleNetV2}(I_{\text{norm}}) \\ x_{\text{mobile}} &= \text{MobileNetV2}(I_{\text{norm}}) \end{aligned}$$

Feature Fusion: Concatenate the extracted features:

$$x_{\text{fused}} = [x_{\text{shuffle}}, x_{\text{mobile}}]$$

Classification: Pass the fused features through fully connected layers:

$$\begin{aligned} z &= \text{Linear}(x_{\text{fused}}) \\ z_{\text{ReLU}} &= \text{ReLU}(z) \\ y_{\text{logits}} &= \text{Linear}(z_{\text{ReLU}}) \end{aligned}$$

Softmax Activation: Compute class probabilities:

$$P = \text{Softmax}(y_{\text{logits}})$$

Prediction:

$$\hat{y} = \arg \max(P)$$

Optimization: Update model parameters using the Adam optimizer.

Evaluation: Accuracy, Precision, Recall, F1-Score, ROC-AUC, and PR-AUC.

H. Model Training Settings

The detection models in the research include VGG19, MobileNetV2, ShuffleNetV2, and SqueezeNet1 models, as well as the proposed MobiShuffNetV2, a combination of MobileNetV2 and ShuffleNetV2 models. The models

were trained with A specific configuration to classify eye diseases. In the Table XI below, we listed the essential training environments and hyper parameters involved

among all models including the proposed MobiShuffNetV2 to make the training efficient without sacrificing performance.

TABLE XI. TRAINING CONFIGURATION AND KEY HYPERPARAMETERS FOR ALL MODELS (VGG19, MOBILENETV2, SHUFFLENETV2, SQUEEZE NET1, AND MOBISHUFFNETV2)

Setting	Value
Model Architecture	VGG19, MobileNetV2, ShuffleNetV2, SqueezeNet1
Proposed Model	MobiShuffNetV2 (Combination of MobileNetV2 and ShuffleNetV2)
Input Image Size	224×224
Classes	Eyelid, Normal, Cataract, Uveitis, Conjunctivitis
Number of Classes	5
Optimizer	Adam (learning rate: 1×10^{-4} , weight decay: 1×10^{-4})
Loss Function	CrossEntropyLoss
Learning Rate Scheduler	ReduceLROnPlateau (patience: 3)
Batch Size	32
Epochs	30
Early Stopping	Patience: 5 (stops training if validation loss does not improve for five consecutive epochs)
Image Normalization	IMAGENET1K_V1 mean: [0.485, 0.456, 0.406], std: [0.229, 0.224, 0.225]
Training Time	Variable (dependent on GPU and hardware)
Training GPU	Trained on Kaggle P100 GPU
Evaluation Metrics Framework	Accuracy, Classification Report (Precision, Recall, F1), Confusion Matrix, ROC AUC, Precision-Recall AUC PyTorch

I. Explainable AI

Grad-Cam has been used to both branches of the proposed MobiShuffNetV2 model, since it cannot be used on the fused feature space. We also make sure that the salient regions are sufficiently highlighted on both branches by using Grad-CAM on each branch separately. If both branches can identify the relevant regions, the fused representations will also focus on the appropriate areas to be predicted, thereby making the model more predictable.

J. Real Time Prediction Tool

The proposed model is deployed in Hugging Face Spaces, utilizing Python on the backend. This deployment provides users with a user-friendly web interface that offers real-time predictions. The system takes input data, runs it through the model, and provides predictions instantly, making it easily integrable into real-world use. Scalability, accessibility, and easy extensibility to more uses are also provided by the deployment.

IV. RESULTS AND DISCUSSION

A comparative analysis of the performance of all evaluated models in detecting eye disease, based on training, validation, and test sets, is presented in Table XII. The proposed MobiShuffNetV2 model achieves the highest accuracy in all three partitions, with training, validation, and test accuracies of 99.98%, 99.39%, and 99.70%, respectively. These findings indicate that not only does the proposed MobiShuffNetV2 perform well during training, but it also generalizes effectively to unseen data. Conversely, VGG19, MobileNetV2, Shuff-netV2, and SqueezeNet1 have slightly lower test accuracies, further supporting the performance advantage of the proposed MobiShuffNetV2.

Besides the overall accuracy, Table XIII contains a detailed classification report that provides information on

the precision, recall, and F1-Score for each model across all classes in the test set. MobiShuffNetV2 outperforms all other models in these metrics. It achieves a perfect accuracy, recall, and F1-Score of 1.0000 for the Normal and Uveitis classes, meaning that the model accurately diagnoses all cases of these eye conditions in the dataset. The most promising results with MobiShuffNetV2 are achieved in the Eyelid, Cataract, and Conjunctivitis classes, highlighting the approach's ability to identify a wide range of ocular conditions accurately. In comparison, VGG19, MobileNet V2, ShuffleNet V2, and SqueezeNet1 exhibit lower performance, albeit with reduced precision and recall values, particularly for more challenging diseases such as Uveitis and Conjunctivitis. For example, SqueezeNet1 exhibits a relatively lower recall rate in Uveitis detection, indicating that it may miss some cases compared to MobiShuffNetV2.

TABLE XII. MODEL PERFORMANCE COMPARISON WITH THE BEST MODEL MARKED

Model	Train Accuracy	Validation Accuracy	Test Accuracy
VGG19	0.9985	0.9817	0.9889
MobileNetV2	0.9987	0.9929	0.9919
ShuffleNetV2	0.9992	0.9878	0.9919
SqueezeNet1	0.9996	0.9827	0.9859
MobiShuffNetV2	0.9998	0.9939	0.9970

A. Cross-Validation Performance and Model Robustness

A 5-fold Cross-Validation (CV) procedure was employed to evaluate the generalization and stability of the proposed model across different subsets of the dataset as depicted in Table XIV. The proposed model, MobiShuffNetV2, demonstrated exceptional performance across all five folds, achieving an average accuracy of 0.9954 ± 0.0084 , an F1-Score of 0.9950 ± 0.0080 , an AUC of 0.9999 ± 0.0001 , and a PR AUC of 0.9997 ± 0.0007 .

TABLE XIII. DETAILED CLASSIFICATION REPORT FOR EACH MODEL, SHOWING PRECISION, RECALL, AND F1-SCORE FOR ALL CLASSES. BEST VALUES IN MOBISHUFFNET ARE HIGHLIGHTED

Model	Class	Precision	Recall	F1-Score
VGG19	Eyelid	0.9766	0.9952	0.9858
	Normal	1.0000	0.9885	0.9942
	Cataract	0.9954	0.9909	0.9931
	Uveitis	0.9810	0.9873	0.9841
	Conjunctivitis	0.9860	0.9792	0.9826
MobileNetV2	Eyelid	0.9952	0.9810	0.9880
	Normal	0.9886	1.0000	0.9943
	Cataract	0.9909	0.9909	0.9909
	Uveitis	1.0000	0.9936	0.9968
	Conjunctivitis	0.9862	0.9931	0.9896
ShuffleNetV2	Eyelid	0.9952	0.9857	0.9904
	Normal	0.9924	0.9962	0.9943
	Cataract	0.9954	0.9954	0.9954
	Uveitis	0.9873	0.9873	0.9873
	Conjunctivitis	0.9862	0.9931	0.9896
SqueezeNet1	Eyelid	0.9858	0.9905	0.9881
	Normal	1.0000	0.9847	0.9923
	Cataract	0.9910	1.0000	0.9955
	Uveitis	0.9623	0.9745	0.9684
	Conjunctivitis	0.9790	0.9722	0.9756
MobiShuffNetV2	Eyelid	0.9953	1.0000	0.9976
	Normal	1.0000	1.0000	1.0000
	Cataract	0.9910	1.0000	0.9955
	Uveitis	1.0000	1.0000	1.0000
	Conjunctivitis	1.0000	0.9792	0.9895

TABLE XIV. 5-FOLD CROSS-VALIDATION METRICS FOR DIFFERENT MODELS

Model	Fold	Accuracy	F1-Score (macro)	AUC	PR AUC
VGG19	1	0.9914	0.9905	0.9998	0.9992
	2	0.9863	0.9841	0.9994	0.9977
	3	0.9797	0.9771	0.9989	0.9961
	4	0.9838	0.9818	0.9994	0.9979
	5	0.9838	0.9824	0.9988	0.9960
	Avg $\pm$ SD	0.9850 $\pm$ 0.0038	0.9905 $\pm$ 0.0050	0.9992 $\pm$ 0.0004	0.9994 $\pm$ 0.0015
MobileNetV2	1	0.9954	0.9948	0.9999	0.9996
	2	0.9899	0.9886	0.9996	0.9985
	3	0.9873	0.9855	0.9996	0.9984
	4	0.9873	0.9855	0.9996	0.9984
	5	0.9863	0.9850	0.9998	0.9992
	Avg $\pm$ SD	0.9905 $\pm$ 0.0035	0.9859 $\pm$ 0.0031	0.9997 $\pm$ 0.0001	0.9992 $\pm$ 0.0005
ShuffleNetV2	1	0.9817	0.9790	0.9986	0.9942
	2	0.9853	0.9833	0.9996	0.9982
	3	0.9782	0.9758	0.9990	0.9963
	4	0.9883	0.9868	0.9997	0.9989
	5	0.9726	0.9704	0.9991	0.9966
	Avg $\pm$ SD	0.9812 $\pm$ 0.0055	0.9831 $\pm$ 0.0062	0.9992 $\pm$ 0.0004	0.9978 $\pm$ 0.0015
SqueezeNet1	1	0.9878	0.9871	0.9997	0.9988
	2	0.9817	0.9810	0.9995	0.9980
	3	0.9878	0.9864	0.9996	0.9983
	4	0.9848	0.9830	0.9993	0.9975
	5	0.9894	0.9883	0.9998	0.9993
	Avg $\pm$ SD	0.9863 $\pm$ 0.0027	0.9852 $\pm$ 0.0032	0.9996 $\pm$ 0.0002	0.9984 $\pm$ 0.0006
MobiShuffNetV2	1	0.9787	0.9764	0.9996	0.9983
	2	0.9990	0.9989	1.0000	1.0000
	3	0.9995	0.9996	1.0000	1.0000
	4	1.0000	1.0000	1.0000	1.0000
	5	1.0000	1.0000	1.0000	1.0000
	Avg $\pm$ SD	0.9954 $\pm$ 0.0084	0.9950 $\pm$ 0.0080	0.9999 $\pm$ 0.0001	0.9997 $\pm$ 0.0007

These findings show that the accuracy is not only high but also consistent across the folds and this also confirms the robustness and reliability of the model. Comparatively, VGG19 achieved a mean accuracy of 0.9850 with a standard deviation of 0.0038, while MobileNetV2 achieved a mean accuracy of 0.9905 with a standard deviation of 0.0035. Although these models performed

well, they consistently scored lower than MobiShuffNetV2, particularly in terms of AUC and PR AUC, where MobiShuffNetV2 achieved nearly perfect scores.

The consistent high performance of MobiShuffNetV2 across all folds demonstrates its superior ability to generalize to new, unseen data. This makes it the most

suitable model for classifying eye diseases, offering excellent classification accuracy and robustness for real-world deployment.

B. Proposed MobiShuffNetV2: Performance Visualization with Learning Curve, Confusion Matrix, ROC, Accuracy, PR Curve, and AUC

1) Learning curve analysis

To determine how effective and stable MobiShuffNetV2 is during training, we investigated its learning curve over 30 epochs. Fig. 7 below shows the learning curve of MobiShuffNetV2 including the training and validation losses, and the accuracy of the model across the 30-epoch training period. To begin with, training loss starts at 0.4389 during epoch 1 and steadily declines to the

final value of 0.0008 after epoch 30, which indicates an effective learning process and convergence of the model. The validation loss, starting at 0.1488 at epoch 1, also shows a steady decrease, as it reaches 0.0202 at epoch 30, another evidence of the ability of the model to generalize to unseen data. On the aspect of accuracy, the training accuracy increases to 0.8438 initially and the validation accuracy increases to 0.9949 after epoch number 30, indicating that the model has improved to be more accurate on the training and validation data. Notably, the loss to validation is also very stable during the training process, and there are no severe fluctuations or distortions noticed. This stability in behaviour on the training as well as on the validation set suggests that the model is learning efficiently without being excessively unstable.

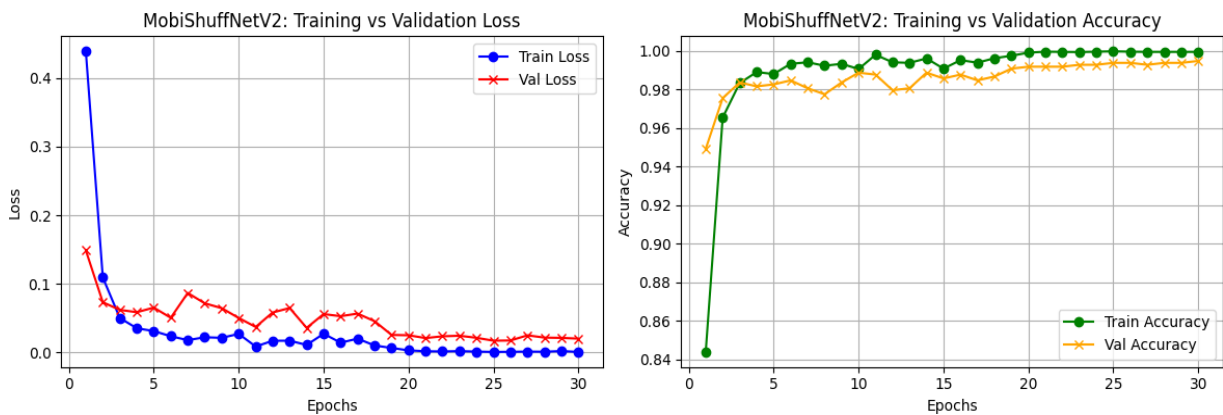


Fig. 7. Learning curve of MobiShuffNetV2 showing training and validation loss and accuracy across epochs. The plot helps visualize model convergence and stability of performance.

2) Exploring misclassification patterns of MobiShuffNetV2 through confusion matrix analysis

To gain a more refined understanding of the classification relationships and potential limitations associated with MobiShuffNetV2, we analyzed its confusion matrix. The confusion matrix of the proposed Mo-biShuffNetV2 model is presented in Fig. 8, which demonstrates the model's ability to differentiate between various types of ocular diseases. Like the capacity of a learning curve to visualize model convergence, the confusion matrix gives a more extensive description of classification performance per category. The diagonal entries correspond to the correctly classified instances per class: Eyelid (210), normal (261), cataract (219), uveitis (157), and conjunctivitis (141). In both the Normal and Uveitis classes, the model achieves perfect precision and recall, scoring 1.0000 in both cases, and thus effectively detects all examples of the two classes. The Eyelid, Cataract, and Conjunctivitis classes are also showing improved performance, with precision and recall approaching unity, thereby strengthening the model's ability to identify these ophthalmic conditions with high accuracy. Overall, the confusion matrix shows that the proposed MobiShuffNetV2 model performs exceptionally well, with only minor misclassifications, indicating its strong ability to generalize and accurately detect eye diseases across the dataset.

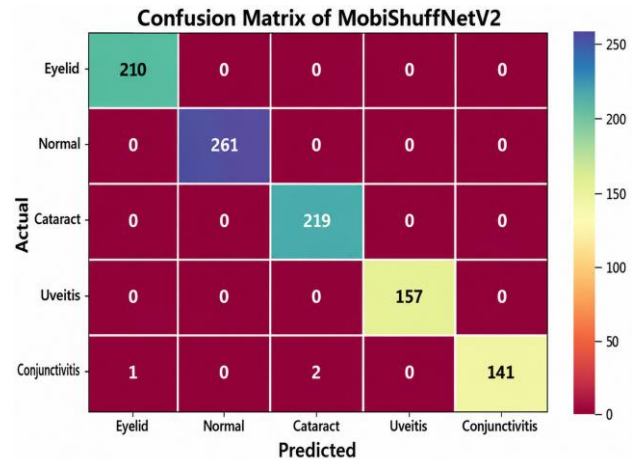


Fig. 8. Confusion matrix showing the performance of the model in classifying various categories.

3) Proposed MobileShuffNetV2 ROC and precision-recall curve analysis

To further assess the discriminatory performance of MobiShuffNetV2 in the five classes of eye disease, we analyzed its ROC and Precision-Recall curves. Fig. 9 shows the ROC and Precision-Recall curve of the proposed model MobiShuffNetV2 in the five-class eye diagnosis classification task. According to the ROC curves, the model has an AUC of 1.0000 across all classes, indicating that it perfectly discriminates between all five

classes. The Precision-Recall curves also suggest that the model is exact and recalls with most PR AUCs being nearly equal to 1 across all classes; only a slight decrease is observed in Conjunctivitis (0.9999). These results indicate that MobiShuffNetV2 offers a good trade-off

between sensitivity and specificity, resulting in reduced false positives and false negatives. The robust and reliable performance makes it a strong candidate for deployment in automated clinical diagnostic systems for detecting ophthalmic diseases.

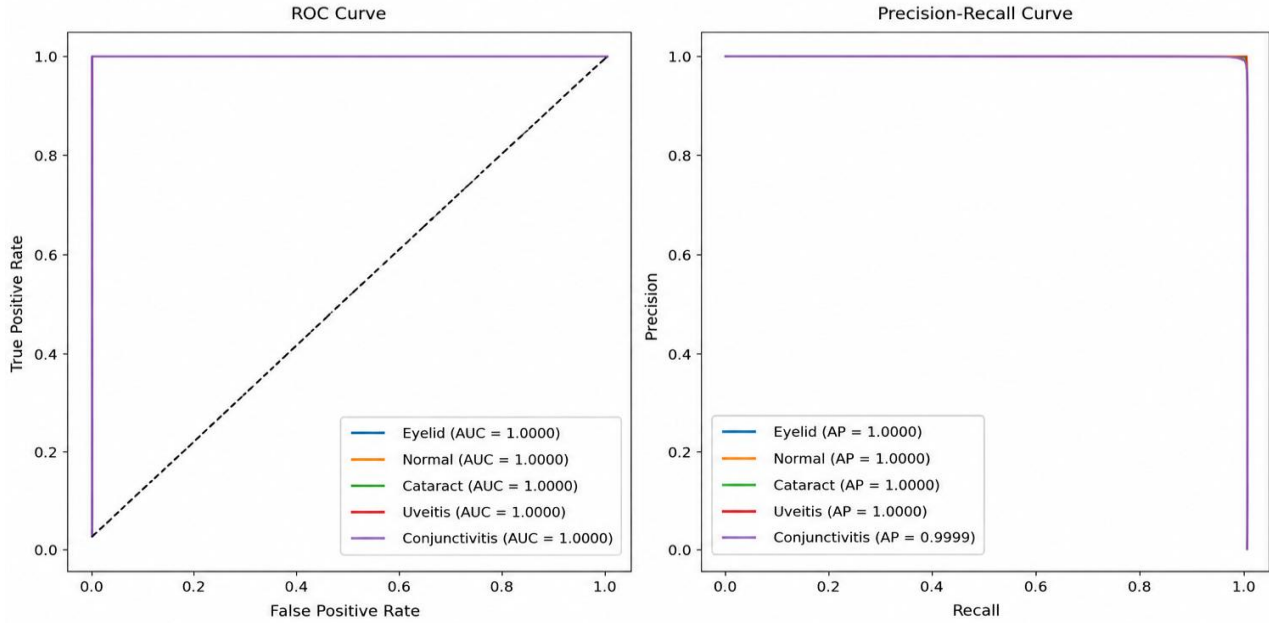


Fig. 9. ROC curve (left) and Precision-Recall curve (right) of the proposed MobiShuffNetV2.

C. Model Efficiency: Training and Inference Time Comparison

In addition to accuracy, the efficiency of a model is a highly crucial factor for its practical deployment, specifically in real-time and resource-constrained environments. Table XV compares the training and inference times for all evaluated models. Among all models, VGG19 has the longest training time at 1265.85 s, which is expected given its architectural complexity. In contrast, SqueezeNet1 is the fastest with a training time of just 286.83 s. However, when it comes to inference time, VGG19 again takes the longest total inference time at 3.4276 s on the test set. In addition to this, MobileNet V2, ShuffleNet V2 and SqueezeNet1 have lower in-inference times per sample with a range of 0.001396 to 0.001437 s. However, MobiShuffNetV2 has a marginally higher inference time of 0.001634 s, which can be described by its more complex design. However, MobiShuffNetV2 still maintains an effective balance between precision and speed, making it a viable solution for real-world applications.

TABLE XV. MODEL TRAINING AND INFERENCE TIME COMPARISON

Model	Training Time (s)	Total Inference Time (s)	Inference Time per Sample (s)
VGG19	1265.85	3.4276	0.003459
MobileNetV2	441.63	1.3847	0.001397
ShuffleNetV2	430.90	1.4236	0.001437
SqueezeNet1	286.83	1.3838	0.001396
MobiShuffNetV2	1018.31	1.6198	0.001634

D. Model Reliability Metrics and Confidence Intervals

To test the robustness and consistency of the proposed architecture against other state-of-the-art models, several reliability metrics were considered as shown in Table XVI. The Cohen's kappa value is highest for MobiShuffNetV2, with a value of 0.9962, indicating strong agreement between the predicted and actual labels. This statistic is used to evaluate inter-rater agreement while correcting for chance agreement. The higher the kappa value, the more consistent and reliable the model's predictions are. Therefore, MobiShuffNetV2 is the most reliable model for eye disease classification. In addition, MobiShuffNetV2 has the narrowest 95% confidence interval for accuracy, [0.9911, 0.9990], suggesting that it is accurate and repeatable across different subsets of data. The 95% confidence interval indicates the range in which the true accuracy of the model is expected to fall with 95% confidence. However, the Cohen's kappa values for MobileNetV2 and ShuffleNetV2 are 0.9898, with slightly wider confidence intervals of [0.9842, 0.9959], suggesting that they also perform well, although with slightly more variability across data subsets. In contrast, VGG19 and SqueezeNet1 are less reliable, with Cohen's kappa scores of 0.9860 and 0.9821, respectively. Their confidence intervals are wider, [0.9802, 0.9938] and [0.9764, 0.9916], respectively, suggesting that these models are still valid but less stable and precise than MobiShuffNetV2.

TABLE XVI. MODEL PERFORMANCE METRICS

Model	Cohen's Kappa	95% Confidence Interval for Accuracy
VGG19	0.9860	[0.9802, 0.9938]
MobileNet_V2	0.9898	[0.9842, 0.9959]
ShuffleNet_V2	0.9898	[0.9842, 0.9959]
SqueezeNet1	0.9821	[0.9764, 0.9916]
MobiShuffNetV2	0.9962	[0.9911, 0.9990]

E. Visualizing Feature Maps in Dual-Branch MobiShuffNetV2 Using Grad-CAM

Further, Grad-CAM has been implemented for both branches of the proposed MobiShuffNetV2 model, as it cannot be directly applied to the integrated feature space. We select each branch individually in the Grad-CAM analysis process to ensure that both branches sufficiently emphasize the salient regions. As can be seen in Fig. 10,

both branches can detect the affected areas with accuracy, which is practically almost perfect. This observation suggests that the representational capacity is significantly enhanced by the combination of the feature maps from the two branches. The relevant spatial areas are served by each branch separately, and the fusion mechanism ensures that the overall features strengthen the right places, thereby enhancing the overall predictive ability of the model. A later verification of the model by domain experts through manual examination, as supported by the Grad-CAM heatmaps, confirms that the highlighted areas are exact coincidences with the affected areas, thereby verifying the model's accuracy. Ultimately, this improved feature fusion yields a substantial improvement in the overall model accuracy.

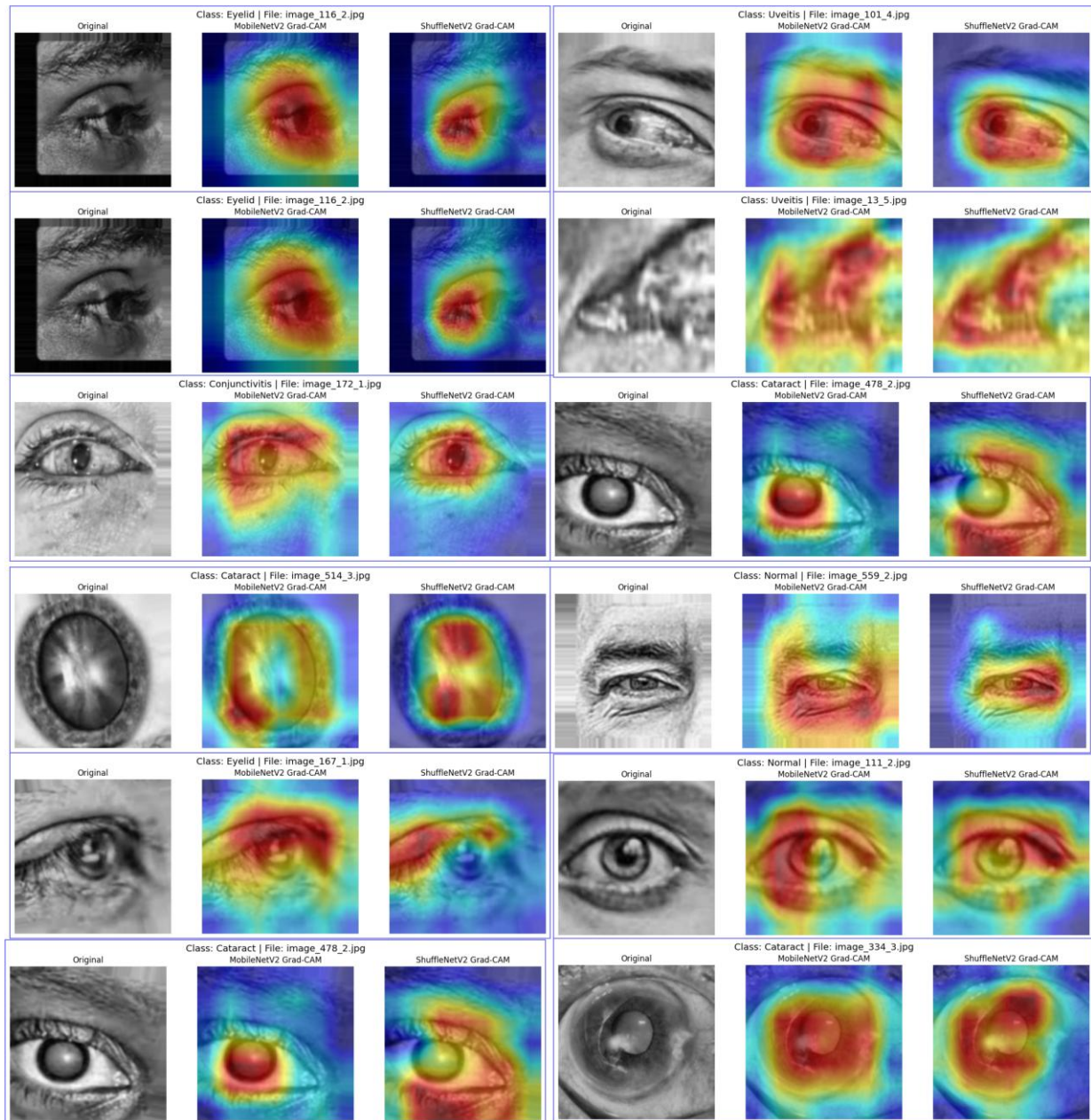


Fig. 10. Grad-CAM heatmaps for dual-branch MobiShuffNetV2 model.

F. Perturbation Robustness and Resource-Efficiency Analysis of the Proposed Model

For further consideration of the practical behavior of the proposed model, perturbation testing and resource-usage profiling were conducted. The input-level distortion assessment of robustness is summarized in Table XVII, and the five distortions used were low light, high light, blur, Cutout, and CutMix. These tests are relevant because eye images captured in real environments can be corrupted by illumination changes, loss of partial information, visual noise, and/or local occlusion. The proposed model performed well under low-lighting conditions, with an accuracy of 96.97%, suggesting that the learned representations are stable under variations in image brightness. Likewise, in the presence of Cutout and CutMix, the model's accuracy reached 98.79% and 97.02%, respectively, demonstrating that the classifier can maintain decision consistency even in the presence of masking or patch-level variations. However, the accuracy decreased under high lighting and blur, reaching 89.00% and 77.40%, respectively. This implies that the model is more sensitive to overexposed images and blurred visual structures, where important disease-related patterns may become less distinguishable. Therefore, the model shows a degree of robustness under various perturbations, but blur handling and illumination normalization should be further improved in future work.

TABLE XVII. PERTURBATION-BASED ROBUSTNESS PERFORMANCE OF THE PROPOSED MODEL

Perturbation Condition	Accuracy (%)
Low lighting	96.97
High lighting	89.00
Blur	77.40
Cutout	98.79
CutMix	97.02

TABLE XVIII. RESOURCE-USAGE PROFILE OF THE PROPOSED MODEL AT THE END OF THE FULL PIPELINE

Resource Metric	Value
RAM used	2.63 GB
Total RAM	31.35 GB
RAM usage	10.50%
GPU name	Tesla T4
GPU allocated	0.20 GB
GPU reserved	1.84 GB
GPU maximum used	3.08 GB

Resource profiling was also conducted to assess the computational efficiency of the proposed model during full pipeline execution. As presented in Table XVIII, the model used only 2.63 GB RAM out of 31.35 GB, corresponding to 10.50% total RAM usage. On a Tesla T4 GPU, the allocated GPU memory was 0.20 GB, while the maximum GPU memory used was 3.08 GB. These results indicate that the proposed model maintains a low memory footprint while preserving strong classification and robustness performance. Hence, the model can be considered resource-efficient for experimental and prototype-level ocular disease classification, although

further testing on mobile, embedded, and multi-device environments is still needed before making stronger deployment claims.

G. Discussion on Generalization Potential and Knowledge-Gap Contribution

The proposed MobiShuffNetV2 framework was built on five eye diseases: Normal, Cataract, Conjunctivitis, Uveitis, and Eyelid disease. Thus, the diagnostic use of its current scope should be taken up within these disease categories, and not as a general eye disease detection system. However, the framework does offer an important basis for the further development of simple, based in the eye, screening for other conditions. MobileNetV2 and ShuffleNetV2 enable the model to learn complementary visual representations, such as opacity, redness, eyelid abnormalities, structural changes, and inflammation-related texture abnormalities. If sufficient labeled samples of other visually observable anterior eye conditions are available for retraining or fine-tuning, these learned representations could be transferred to them.

This study aims to address the primary knowledge gap of efficient deep learning frameworks specialized for simple eye photographs. Existing studies on eye disease detection are mainly based on fundus or Optical Coherence Tomography (OCT) images, which require specialized imaging devices and clinical facilities. This work, on the other hand, investigates whether simple eye images can be used for early multi-class eye disease detection with a simple hybrid architecture. This work provides a valuable reference for low-cost pre-screening of ocular diseases due to its accessible image acquisition, targeted preprocessing, class balancing, hybrid model design, Grad-CAM interpretability, robustness testing, and resource profiling.

The suggested framework can be expanded in two primary ways for other diseases. First, the final classification layer can be adjusted and optimized with a larger amount of data that includes new disease classes. Second, open-set or out-of-distribution detection can be added to enable the model to label uncertain or unseen instances as “out-of-distribution” instead of assigning them to one of the five classes. Diseases with subtle visual patterns or posterior-segment involvement would be particularly meaningful for such extensions, where the initial external eye photographs may not provide sufficient diagnostic information. Hence, although the present work addresses a critical methodological gap by demonstrating the efficiency of ocular-photography-based classification, future studies should be conducted on larger multi-center and multi-device datasets that include more disease classes and expert annotations.

H. Comparative Analysis of Eye Disease Detection Approaches

As shown in Table XIX, fundus-based and OCT-based approaches achieve high accuracies ranging from 87.9% to above 99%, primarily by utilizing advanced convolutional neural network settings, attention mechanisms, and ensembles. These methods exhibit high predictive performance but require specialized and

expensive imaging equipment, as well as substantial preprocessing pipelines and multistage modeling processes, which limit their applicability in rural or resource-constrained settings. Additionally, most studies

are based on single-source data, have limited external validation, and lack interpretability, all of which are barriers to implementing them in routine clinical practice.

TABLE XIX. COMPARISON OF EYE DISEASE DETECTION APPROACHES: FUNDUS-BASED, OCT-BASED, AND SIMPLE EYE IMAGE-BASED

Category	Study/Citation	Method/Model	Accuracy (%)	Contribution
Fundus-based	[21]	Lightweight CNN, SMOTE	87.9	Efficient multiclass retinal screening with class balancing, lightweight model
	[22]	EfficientNetB3, transfer learning	95.12	Robust retinal disease classification using transfer learning
	[23]	RetinaDNet, ensemble of 6 backbones	99.2	Ensemble-based dual-branch fundus and vessel segmentation for high accuracy
	[24]	MobileNet + DenseNet121 + PCA	98.5	Hybrid architecture combining CNN features with PCA for improved classification
	[25]	Pretrained networks (DenseNet-121)	96.2	Evaluated multiple pre-trained networks for retinal disease detection
OCT-based	[35]	VGG + ResNet, transfer learning	99.7	High accuracy three-class and six-class retinal OCT classification
	[36]	Ensemble binary classifiers + YOLOv3	>99	Multi-stage ensemble approach with pathology detection and Grad-CAM interpretability
	[37]	RETFound (pretrained foundation model)	AUC 0.80	Large-scale pretraining on OCT scans for feature-rich representations
	[37]	OCTNet, attention DenseNet	97.92	Attention-based network emphasizing relevant OCT features
	[44]	FN-OCT, Inception fusion + attention	98.7	Fusion of multiple CNN variants with attention for improved OCT classification
Simple Eye Images	This Study	MobiShuffNetV2	99.70	Lightweight hybrid model for 5-class eye disease detection using simple eye photographs

In comparison, our simple eye image-based strategy, which focuses on the lightweight MobiShuffNetV2, achieves 99.70% (internal dataset) and 97.96% (external independent test set) accuracy, demonstrating high generalization with a real-time and efficient model. By integrating explainable AI (Grad-CAM), MobiShuffNetV2 produces interpretable heatmaps that highlight disease-affected regions, enhancing clinical trust. Its lightweight architecture and fast inference enable deployment as a web application, providing practical usability in real-world settings. The small size and fast inference of the architecture make it easily deployable as a web application, making its use in a real-world context a viable endeavor. This method requires no specialized imaging hardware, which makes early eye-disease screening affordable in underserved groups. Mo-biShuffNetV2, therefore, sets a new standard in pre-screening devices based on basic ocular images, thus filling a significant gap in accessibility and scalability.

I. Interactive Web Application for Eye Disease Detection

The proposed MobiShuffNetV2 model has been deployed in a real-time environment using the Hugging Face Gradio platform, enabling interactive use by both general users and rural communities. The application allows users to upload an ophthalmic image and obtain predictions in approximately 3 s, demonstrating its efficiency for practical deployment.

Visualizations, such as Grad-CAM heatmaps, allow users to see which parts of the image influenced the model's decision, thereby building confidence in the model's decisions among users. As illustrated in Fig. 11, the tool can simultaneously identify images from various sets, including Cataract, Uveitis, Eyelid, and Conjunctivitis. Additionally, the highlighted areas can be seen as those that attract the most attention for the classifiers.

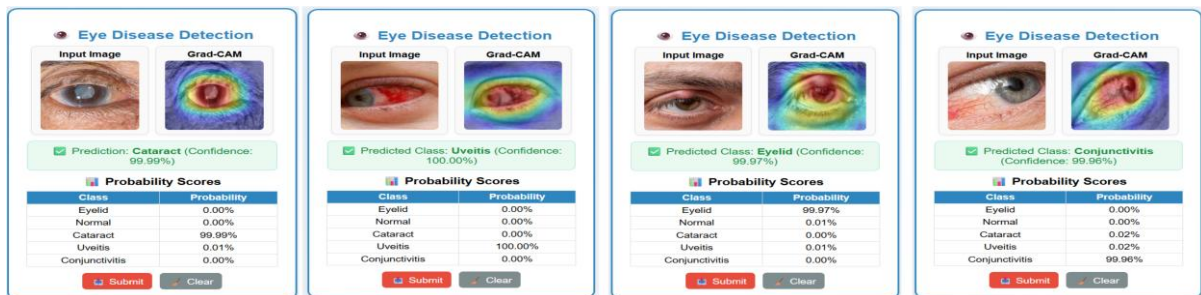


Fig. 11. Interactive web application displaying input image, Grad-CAM visualization, and predicted class probabilities.

Moreover, every tested image within the platform is completely new, thus the model's generalization capacity to unseen examples is validated. This realization serves as solid proof of concept for MobiShuffNetV2 as a simple, reliable, and easy-to-understand first-line screening tool with capabilities such as early detection of ophthalmic diseases, alleviation of clinicians' workload, fast decision-making support, and accessibility in underdeveloped areas. MobiShuffNetV2, which utilizes a basic eye camera to enable real-time screening, is a perfect example of how deep learning can be easily integrated into clinical practice without taking the long road.

V. CONCLUSION

This work introduces a classification system for human eyes that displays photographic images of the eye for five different conditions (Normal, Cataract, Conjunctivitis, Uveitis, and eyelid disorders) using MobiShuffNetV2, a lightweight hybrid deep learning model. The model achieved a testing accuracy of 99.70% (95% CI: [0.9911, 0.9990]). Additionally, both its precision and recall were at maximum against the Normal and Uveitis categories, and its Cohen's Kappa of 0.9962 indicates a very high level of agreement. The usage of cross-validation results (99.54% $\pm$ 0.84% accuracy) proved the power of the model, and testing on another set of 296 images gave the accuracy of 97.96%, which points out the model's high generalization capability. Grad-CAM visualizations enabled the provision of interpretable heatmaps that accurately outlined disease-affected areas, thereby enhancing clinical trust and confidence. The efficient architecture, with a size of 157.34 MB and a per-sample inference time of 0.0016 s, is the advantage that allows the model to be deployed in real-time on mobile and edge devices. The actual deployment of a web app provides significant support for the use of this model in settings where resources are limited.

The small dataset size is the primary limitation of the present study, which occasionally leads to false classifications, particularly between visually similar classes such as Uveitis and Conjunctivitis. In addition, due to the unavailability of another publicly accessible and clinically validated eye disease dataset with similar class distributions, external validation could not be performed in this study. Consequently, the generalization capability of the proposed framework across different clinical environments remains a potential limitation. Also, the current model is more suitable as a pre-screening tool than a clinical diagnostic tool. Future work will therefore focus on enlarging and diversifying the dataset through additional clinical image acquisition and advanced augmentation strategies. Further improvements in preprocessing techniques and domain-specific enhancement methods will also be explored to improve model robustness. Moreover, the proposed framework will be piloted in real rural hospital settings to evaluate its practical usability and clinical reliability in real-world scenarios. These improvements are expected to enhance the model's generalization capability, reliability, and acceptance among healthcare practitioners, making

MobiShuffNetV2 a scalable and practical solution for early ocular disease detection in remote and underserved regions.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research is funded by Woosong University Academic Research 2026.

REFERENCES

- [1] S. S. Patel, N. V. Saraiya, H. H. Tessler, and D. A. Goldstein, "Mycobacterial ocular inflammation: Delay in diagnosis and other factors impacting morbidity," *JAMA Ophthalmology*, vol. 131, no. 6, pp. 752–758, 2013.
- [2] S. N. Nova, M. S. Rahman, and C. Chakraborty, "Patients' health surveillance model using IoT and 6G technology," in *Green Technological Innovation for Sustainable Smart Societies: Post Pandemic Era*, pp. 191–209, Springer, 2021.
- [3] D. Etya'ale, "Blindness and vision impairment," in *Global health essentials*, Springer, 2023, pp. 209–213.
- [4] A. Berni, C. Foti, L. Ulla *et al.*, "Choriocapillaris in age-related macular degeneration: A systematic review of optical coherence tomography angiography-based assessments and challenges in standardization," *American Journal of Ophthalmology*, vol. 277, pp. 482–496, 2025.
- [5] M. N. Absur, "Anomaly detection in biomedical data and image using various shallow and deep learning algorithms," in *Proc. Data Intelligence and Cognitive Informatics: Proceedings of ICDICI 2021*, Springer, 2022, pp. 45–58.
- [6] A. Saini, K. Guleria, and S. Sharma, "An efficient deep learning model for eye disease classification," in *Proc. 2023 International Research Conference on Smart Computing and Systems Engineering (SCSE)*, vol. 6, IEEE, 2023, pp. 1–6.
- [7] A. Bali and V. Mansotra, "Analysis of deep learning techniques for pre-diction of eye diseases: A systematic review," *Archives of Computational Methods in Engineering*, vol. 31, no. 1, pp. 487–520, 2024.
- [8] M. Kulyabin, A. Zhdanov, A. Nikiforova *et al.*, "OCTDL: Optical coherence tomography dataset for image-based deep learning methods," *Scientific Data*, vol. 11, no. 1, 365, 2024.
- [9] X. Chen, J. Wang, T. Qian *et al.*, "An artificial intelligence cloud platform for oct-based retinal anomalies screening system in real clinical environments," *NPJ Digital Medicine*, vol. 8, no. 1, 559, 2025.
- [10] I. Khalil, A. Mehmood, H. Kim, and J. Kim, "OCTNET: A modified multi-scale attention feature fusion network with inceptionv3 for retinal oct image classification," *Mathematics*, vol. 12, no. 19, 3003, 2024.
- [11] E. Sherman, L. M. Niziol, P. M. Hicks *et al.*, "A screening strategy to mitigate vision impairment by engaging adults who underuse eye care services," *JAMA Ophthalmology*, vol. 142, no. 10, pp. 909–916, 2024.
- [12] V. Umaefulam, S. Safi, G. Lingham, I. Gordon, A. Mueller, N. S. Krishnam, V. L. A. Carneiro, M. Yu, J. R. Evans, and S. Keel, "Approaches for delivery of refractive and optical care services in community and primary care settings," *The Cochrane Database of Systematic Reviews*, vol. 2024, no. 5, 2024, CD016043.
- [13] A. Santone, M. Cesarelli, E. Colasuonno, V. Bevilacqua, and F. Mercaldo, "A method for ocular disease diagnosis through visual prediction explainability," *Electronics*, vol. 13, no. 14, 2706, 2024.
- [14] S. Abbas, A. Qaisar, M. S. Farooq, M. Saleem, M. Ahmad, and M. A. Khan, "Smart vision transparency: Efficient ocular disease prediction model using explainable artificial intelligence," *Sensors*, vol. 24, no. 20, 6618, 2024.
- [15] G. Clare, J. H. Kempen, and C. Pavésio, "Infectious eye disease in the 21st century—An overview," *Eye*, vol. 38, no. 11, pp. 2014–2027, 2024.
- [16] S. Toumaj, A. Heidari, and N. Jafari Navimipour, "Leveraging explain-able artificial intelligence for transparent and trustworthy

- cancer detection systems,” *Artificial Intelligence in Medicine*, vol. 169, 103243, 2025.
- [17] S. Dabeer, M. M. Khan, and S. Islam, “Cancer diagnosis in histopathological image: CNN based approach,” *Informatics in Medicine Unlocked*, vol. 16, 100231, 2019.
 - [18] R. Abbasi, F. Amin, A. Alabrah, G. S. Choi, S. Khan, M. B. Bin Heyat, M. S. Iqbal, and H. Chen, “Diabetic retinopathy detection using adaptive deep convolutional neural networks on fundus images,” *Scientific Reports*, vol. 15, no. 1, 24647, 2025.
 - [19] S. Toumaj, A. Heidari, R. Shahhosseini *et al.*, “Applications of deep learning in Alzheimer’s disease: A systematic literature review of current trends, methodologies, challenges, innovations, and future directions,” *Artificial Intelligence Review*, vol. 58, 44, 2025.
 - [20] M. Santo, F. Bhoyan, F. Farhad *et al.*, “Maxgrnet: A multi-axis vision transformer with improved generalization for eye disease classification using explainable AI with insertion-deletion operations on fundus images,” *PLOS ONE*, vol. 21, no. 4, e0346329, 2026.
 - [21] A. K. B. Arnob, M. H. R. Chayon, F. Al Farid, M. N. Husen, and F. Ahmed, “A lightweight CNN for multiclass retinal disease screening with explainable AI,” *Journal of Imaging*, vol. 11, no. 8, 2025.
 - [22] R. Alsohemi and S. Dardouri, “Fundus image-based eye disease detection using efficientnetb3 architecture,” *Journal of Imaging*, vol. 11, no. 8, 2025.
 - [23] H. Yu and X. Dong, “Ensemble-based eye disease detection system utilizing fundus and vascular structures,” *Scientific Reports*, vol. 15, 19298, 2025.
 - [24] A. Shamsan, E. M. Senan, and H. S. A. Shatnawi, “Automatic classification of colour fundus images for prediction eye disease types based on hybrid features,” *Diagnostics*, vol. 13, no. 10, 2023.
 - [25] H. Imaduddin and A. R. Sakina, “Eye disease detection using transfer learning based on retinal fundus image data,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 36, no. 1, pp. 509–516, 2024.
 - [26] S. Al-Fahdawi, A. S. Al-Waisy, D. Q. Zeebaree *et al.*, “Fundus-deepnet: Multi-label deep learning classification system for enhanced detection of multiple ocular diseases through data fusion of fundus images,” *Information Fusion*, vol. 102, 102059, 2024.
 - [27] A. R. Wahab Sait, “Artificial intelligence-driven eye disease classification model,” *Applied Sciences*, vol. 13, no. 20, 2023.
 - [28] A. A. Jeny, M. S. Junayed, and M. B. Islam, “Deep neural network-based ensemble model for eye diseases detection and classification,” *Image Analysis and Stereology*, vol. 48, no. 3, pp. 393–405, 2024.
 - [29] T. D. Nguyen, D.-T. Le, J. Bum, S. Kim, S. J. Song, and H. Choo, “Retinal disease diagnosis using deep learning on ultra-wide-field fundus images,” *Diagnostics*, vol. 14, no. 1, 2024.
 - [30] T. Shyamalee, D. Meedeniya, G. Lim, and M. Karunaratne, “Automated tool support for glaucoma identification with explainability using fundus images,” *IEEE Access*, vol. 12, pp. 17290–17307, 2024.
 - [31] M. Vadduri and P. Kuppusamy, “Enhancing ocular healthcare: Deep learning-based multi-class diabetic eye disease segmentation and classification,” *IEEE Access*, vol. 11, pp. 137881–137898, 2023.
 - [32] S. Ejaz, R. Baig, Z. Ashraf, M. M. Alnfai, M. M. Alnahari, and R. M. Alotaibi, “A deep learning framework for the early detection of multi-retinal diseases,” *PLOS ONE*, vol. 19, no. 7, e0307317, 2024.
 - [33] A. A. Abd El-Khalek, H. M. Balaha *et al.*, “A concentrated machine learning-based classification system for Age-related Macular Degeneration (AMD) diagnosis using fundus images,” *Scientific Reports*, vol. 14, no. 1, 2434, 2024.
 - [34] K. Wang, C. Xu, G. Li, Y. Zhang, Y. Zheng, and C. Sun, “Combining convolutional neural networks and self-attention for fundus diseases identification,” *Scientific Reports*, vol. 13, no. 1, 2023.
 - [35] J. Han, S. Choi, J. I. Park, J. S. Hwang, J. M. Han, J. Ko, J. Yoon, and D. D.-J. Hwang, “Detecting macular disease based on optical coherence tomography using a deep convolutional network,” *Journal of Clinical Medicine*, vol. 12, no. 3, 2023.
 - [36] X. Chen, Y. Xue, X. Wu, Y. Zhong, H. Rao, H. Luo, and Z. Weng, “Deep learning-based system for disease screening and pathologic region detection from optical coherence tomography images,” *Translational Vision Science Technology*, vol. 12, pp. 29–29, 2023.
 - [37] K. Du, A. R. Nair, S. Shah *et al.*, “Detection of disease features on retinal OCT scans using RETFound,” *Bioengineering*, vol. 11, no. 12, 2024.
 - [38] L. Inferrera, L. Borsatti, A. Miladinovic *et al.*, “Oct-based deep-learning models for the identification of retinal key signs,” *Scientific Reports*, vol. 13, no. 1, 14628, 2023.
 - [39] W. J. Jaimes, W. J. Arenas, H. J. Navarro, and M. Altuve, “Detection of retinal diseases from OCT images using a VGG16 and transfer learning,” *Discover Applied Sciences*, vol. 7, no. 3, 160, 2025.
 - [40] G. Gencer and K. Gencer, “Advanced retinal disease detection from oct images using a hybrid squeeze and excitation enhanced model,” *PLOS ONE*, vol. 20, pp. 1–21, 2025.
 - [41] Puneet, R. Kumar, and M. Gupta, “Optical coherence tomography image based eye disease detection using deep convolutional neural network,” *Health Information Science and Systems*, vol. 10, no. 1, p. 13, 2022.
 - [42] F. M. Talaat, A. A. A. Ali, R. ElGendy, and M. A. ELShafie, “Deep attention for enhanced oct image analysis in clinical retinal diagnosis,” *Neural Computing and Applications*, vol. 37, no. 2, pp. 1105–1125, 2025.
 - [43] M. Elkholy and M. A. Marzouk, “Deep learning-based classification of eye diseases using convolutional neural network for oct images,” *Frontiers in Computer Science*, vol. 5, 2024.
 - [44] Z. Ai, X. Huang, J. Feng, H. Wang, Y. Tao, F. Zeng, and Y. Lu, “FN-OCT: Disease detection algorithm for retinal optical coherence tomography based on a fusion network,” *Frontiers in Neuroinformatics*, vol. 16, 2022.
 - [45] M. Khalid, M. Z. Sajid, A. Youssef, N. A. Khan, M. F. Hamid, and F. Ab-bas, “Cad-eye: An automated system for multi-eye disease classification using feature fusion with deep learning models and fluorescence imaging for enhanced interpretability,” *Diagnostics*, vol. 14, no. 23, 2024.
 - [46] A. K. Bitto and M. Ahmed, “Image dataset on eye diseases classification (uveitis, conjunctivitis, cataract, eyelid) with symptoms and smote validation,” *Mendeley Data*, vol. 1, 2024.

Copyright © 2026 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).