EAI Endorsed Transactions
on AI and Robotics
Research Article EAI.EU

Explainable AI Based Deep Ensemble Convolutional Learning for Multi-Categorical Ocular Disease Prediction

Abu Kowshir Bitto[1] Rezwana Karim[1], Mst Halema Begum[2], Md Fokrul Islam Khan[2], Md. Maruf Hassan[3], Abdul Kadar Muhammad Masum[3*]

[1]Department of Software Engineering, Daffodil International University, Dhaka, Bangladesh
[2]Department of Management Information Systems, International American University, Los Angeles, United States
[3]Department of Computer Science and Engineering, Southeast University, Dhaka, Bangladesh

Abstract

Diseases of the eye such as diabetic retinopathy, glaucoma, and cataract remain among the leading causes of blindness and vision impairment worldwide. Diagnosis in its early stages followed by early treatment is crucial to preventing permanent loss of vision. Recent advances in Artificial Intelligence (AI), particularly Transfer Learning and Explainable AI (XAI), have proven highly promising in automating the identification of retinal pathologies from medical images. In this paper, we propose an ensemble deep learning approach that integrates four pre-trained convolutional neural networks, i.e., VGG16, MobileNet, DenseNet, and InceptionV3, to classify retinal images into four categories: diabetic retinopathy, glaucoma, cataracts, and normal. The ensemble method leverages the power of multiple models to improve classification accuracy. Additionally, Explainable AI techniques are applied to make the model more interpretable, with visual explanations and insights into AI system decision-making and thereby establishing clinical trust and reliability. The system is evaluated on a new benchmarked eye disease dataset used from Hugging Face, and the results in terms of accuracy and model transparency are encouraging. This research contributes towards developing reliable, explainable, and efficient AI-driven diagnostic systems to assist healthcare professionals in the early detection and management of eye diseases

Keywords: Eye Disease, Deep Ensemble Learning, Transfer Learning, Explainable AI, Ocular Disease.

Received on 03 May 2025, accepted on 05 July 2025, published on 28 July 2025

doi: 10.4108/airo.9234

1. Introduction

The eye is one of the most important sensory organs of the human body. In the contemporary period, ocular disease has emerged as a significant worldwide health concern [1], affecting the productivity and autonomy of individuals [2]. The standard visual acuity is established at 20/20; conversely, measurements less than 20/40 or 20/60 represent low vision capabilities [3]. Many individuals globally suffer from retinal diseases, which, if not diagnosed and treated early, may result in blindness [2].

Eye disease includes several conditions such as diabetic retinopathy, diabetic macular edema, glaucoma, and cataracts [4][5]. One of the leading causes of blindness among working-age adults is diabetic retinopathy [6]. Glaucoma is the most common cause of blindness globally, and it contributes to a high percentage of vision loss cases. Glaucoma results from optic nerve damage, which is usually accompanied by increased intraocular pressure. Glaucoma, if not detected and treated early, can result in irreversible

*Corresponding author. Email: akmmasum@yahoo.com

blindness. In 2020, an estimated 79 million individuals worldwide were affected by glaucoma, with its incidence increasing at a fast rate, especially in urban areas [7][8]. Cataracts cause opacification of the lens in the eye, which in turn leads to a gradual loss of visual acuity. Over half of the population older than 65 years suffer from age-related cataracts, which occur due to accumulation of proteins in the lens. Surgery to replace the opaque lens with an artificial one has shown effectiveness; however, the rising incidence of cataracts, particularly diabetes-related cataracts, remains a significant public health concern [9] - [13].

Artificial intelligence (AI), specifically through Transfer Learning and Explainable AI (XAI), has become a promising means for transforming the diagnosis of eye diseases. Artificial intelligence algorithms have shown proficiency in detecting a variety of retinal and eye conditions, such as diabetic retinopathy, glaucoma, and cataracts, from high-resolution digital images of the eye [14][15]. Transfer Learning entails utilizing pre-trained models, i.e., Convolutional Neural Networks (CNNs), in detecting new images corresponding to medical conditions, thereby enhancing the accuracy and effectiveness of disease detection [16]- [18]. In parallel, Explainable AI (XAI) enhances model interpretability and transparency, providing clinicians with clearer insights into the decision-making algorithms utilized by AI tools, an element critical to establishing clinical trust and adoption. Such innovations through artificial intelligence allow for faster and more accurate diagnoses, thus improving patient outcomes and assisting clinicians in making informed decisions.

In current context, diabetic retinopathy, glaucoma, and cataracts are three of the globe's most common causes of blindness and visual impairment, impacting millions of people and placing a significant burden on healthcare systems. Existing diagnostic techniques are based largely on ophthalmologists' subjective analysis of retinal photographs, which is not only time-consuming and prone to human error but also of limited availability especially in rural or disadvantaged areas where trained ophthalmologists are scarce. With the increased incidence of eye diseases and the need for increased speed and accuracy in diagnosis, artificial intelligence (AI) stands out as a revolutionary solution. By automating the interpretation of retinal scans, AI has the potential to greatly speed up the diagnostic process, improve precision, and aid timely interventions before irreversible vision impairment sets in. To overcome these problems, this study aims to develop a robust AI-driven system for the diagnosis and classification of severe eye diseases i.e., diabetic retinopathy, glaucoma, and cataracts using transfer learning techniques on retinal scans. Moreover, to make the model outputs clear and trustworthy, explainable AI (XAI) techniques will be integrated, enabling healthcare professionals to understand and validate the AI-generated diagnoses with confidence. The model will be validated for performance using the standard performance metrics like accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC) to determine its clinic value for early and accurate eye disease diagnosis.

This research explores the use of Transfer Learning and Explainable AI (XAI) to classify retinal images with focus on diseases like diabetic retinopathy, glaucoma, cataracts, and normal eyes with a new benchmark dataset from Hugging Face. The aim is to develop an accurate, interpretable, and clinically relevant AI system to help the early diagnosis of these ocular conditions. The system will help in providing an accelerated, more accurate diagnosis, allowing healthcare professionals to take prompt action and enhance patient care.

2. Related Work

2.1. Deep learning for Eye Disease

Deep learning has emerged as a revolutionary approach for the detection and classification of many eye diseases using high-powered neural network models for enhancing diagnostic yield as well as ophthalmic functioning. The exponential rise in the use of these technologies, particularly in the evaluation of fundus images, demonstrates their capacity to identify an extensive spectrum of ocular diseases with high sensitivity and specificity.

Cen et al., [20] who came up with an automated system capable of detecting up to 39 different fundus diseases and conditions, with their accuracy rates on par with expert retina specialists. This work shows the capabilities of deep learning models in advanced diagnostic applications. On the same lines, Kumar and Dhanalakshmi [21] came up with a new architecture, i.e., EYE-YOLO, based on utilizing multi-spatial pyramid pooling for improving detection of eye disease from fundus images since it represents innovation in deep learning approaches towards fulfilling ocular health-specific demands. Shibata et al. [22] highlighted the efficacy of the ResNet architecture in accurately diagnosing glaucoma from fundus photography, emphasizing its role in enhancing the diagnostic processes, especially in challenging cases like highly myopic eyes. Aurangzeb et al., [23] used advanced machine learning techniques are critical in streamlining early detection and diagnosis across various ocular conditions, including diabetic retinopathy and glaucoma. Sarhan et al. [24] also reinforces the extensive utilization of machine learning in ophthalmic data processing, affirming its significance in image analysis for eye disease diagnosis. Deep learning models have shown remarkable capabilities in feature extraction and pattern recognition directly from medical images. Alam et al., [25] who reported a high sensitivity of 94.84% in identifying retinopathies through their AI-based tool. Albalawi et al. [26] shows the integration of dynamic Swin transformers in developing a classification system for multiple retinal diseases, showcasing the importance of utilizing contemporary architectures to enhance diagnostic precision.

2.2. Transfer learning for Eye Disease

Transfer learning has been shown to be a critical technique in the identification and classification of eye diseases,

particularly where insufficient significant labeled data is available. Using pre-trained models on large datasets, researchers can fine-tune these models to improve their performance on tasks in ophthalmic research.

Madduri and Rao [27] created a VGG16-SGD model using transfer learning to address diabetic eye disease with an accuracy of 96.49% for multi-class classification in the fundus imaging scenario. They provided a hybrid R-CNN + LSTM model methodology, which could classify different retina disorders successfully, proving the successful application of transfer learning in the field. Mahmood et al. [28] in which transfer learning using ResNet50 yielded high accuracies (95.8%) in cataract detection, highlighting the potential of pre-trained models to generalize across different pathologies. Chen et al. [29] pointed out the effective use of deep transfer learning for the diagnosis of some ocular diseases by the analysis of fundus images, indicating that pre-trained deep Convolutional Neural Networks (CNNs) enable models to obtain excellent performance metrics despite the constraints of small datasets. Aranha et al. [30] corroborated this opinion, asserting that disease classification is enhanced by pre-trained CNNs, and therefore they are suitable for detection of different ocular diseases from poor-quality fundus images.

2.3. Deep Ensemble for Eye Disease

Deep ensemble learning has become a powerful approach in eye disease classification and diagnosis, taking advantage of the strengths of various models to boost diagnostic precision and robustness. Ensemble techniques, by aggregating the predictions from various deep learning models, have the potential to enhance the detection of diseases such as diabetic retinopathy (DR), age-related macular degeneration (AMD), and glaucoma.

Sanamdikar et al. [31] proposed a deep ensemble learning method with vessel segmentation for the purpose of improving diabetic retinopathy classification accuracy from retinal images. They utilized the Canny edge detection operator to segment the retinal images prior to utilizing an ensemble of models that were specifically developed for identifying unique features pertinent to different classes of diabetic retinopathy. Desiani et al. [31] showed that an ensemble learning approach with weighted voting, based on multiple convolutional neural network (CNN) models, achieved competent diabetic retinopathy classification with an accuracy rate of 87%. The study emphasized the value of integrating heterogeneous models to take advantage of their complementary strengths in feature extraction and representation processes. Wali et al. [32] also provided additional proof of this concept in their research on the classification of optical coherence tomography (OCT) images. They integrated the merits of several deep learning models, such as DenseNet121 and InceptionV3, for the classification of retinal images into specific categories like choroidal neovascularization and diabetic macular edema. Their ensemble model exhibited improved accuracy and robustness in ocular disease detection.

3. Methodology

Our proposed framework primarily uses an ensemble of transfer learning models to predict the classification of eye disease. Our step-by-step procedure of the paper shows in the Figure 1:

3.1. Dataset

We use new benchmark dataset which no other research yet used named NKS_EYE_DISEASE_CLASSIFICATION [33] dataset consists of labeled eye disease images, categorized into four classes: diabetic retinopathy, cataract, glaucoma, and normal. It contains 4,217 images which are divided into training and testing sets, allowing for multi-class classification tasks. This dataset is primarily designed for developing machine learning models focused on diagnosing eye diseases from retinal images.

3.2. Data Preprocessing

- **Image Representation:** In our preprocessing step for ensuring consistency across all input images, each sample was resized to 224×224 pixels using bilinear interpolation. This image resizing ensures that all images have a uniform size, which is required for transfer learning models such as Mobilenet, DenseNet. Without resizing, the model would not be able to process images of varying resolutions, leading to inconsistencies in feature extraction. This preprocessing ensures that the model effectively extracts key features related to eye diseases, improving classification accuracy.
- **Label Encoding:** The labels for eye diseases classification in this dataset are usually categorical form (e.g., "Glaucoma", "Diabetic Retinopathy", "Cataract"," Normal"), whereas deep learning models need numerical inputs. To convert categorical disease labels into a numerical format we applied Label Encoding. The transformation was fit on the training set and then applied to the test set to ensure consistency. This encoding allowed the model to process class labels effectively and make predictions that could be mapped back to the original disease names.
- **SMOTE (Synthetic Minority Over-sampling Technique):** For reducing the issue of class imbalance problem in our dataset, we applied SMOTE technique to generate synthetic samples for underrepresented disease categories. Since SMOTE operates on tabular data, we initially flattened the image matrices into 2D feature representations. After applying SMOTE, the dataset was reshaped back into its original (224, 224, 3) format to maintain compatibility with deep learning models. This balancing technique prevents the model from biasing toward majority classes, ensuring fair representation across different eye disease classification. The effectiveness of this approach was validated by analyzing class distributions before and after SMOTE

application. The class distribution before and after applying SMOTE, where the data is balanced for class 0, 1, 2, and 3 are showing Table-1 whereas 0 as Cataract, 1 as Diabetic Retinopathy, 2 Glaucoma and 3 is Normal. The pie chart shows in Figure 2 of SMOTE analysis displays the class distribution of the eye disease dataset before and after balancing. The class distribution was not balanced prior to using SMOTE, with the red slice indicating 23.8% for glaucoma, green 24.7% for cataract, orange 25.3% for normal, and blue 26.3% for diabetic retinopathy. With the use of SMOTE, the distribution was made more balanced because each class was now equally distributed at 25.0% as seen by the equal-sized portions of red, green, orange, and blue.

- **Splitting data & Normalization:** After completing several preprocessing steps, now we split the dataset into training and validation sets using an 80-20 ratio, allowing for performance monitoring and hyperparameter tuning during training. Then we normalized all image pixel values to a [0, 1] range by dividing by 255.0, which accelerates model convergence and stabilizes the gradient updates. Furthermore, the class labels were transformed into one-hot encoded vectors to facilitate multi-class classification, where each label is represented as a binary vector, enabling the model to compute class-wise prediction probabilities.

3.3. Model Description

In our proposed framework we choose Transfer learning model over training deep learning models from scratch for eye disease classification due to its ability to handle limited datasets, prevent overfitting, and improve classification accuracy. Medical image datasets are often small and difficult to annotate, making it challenging for traditional deep learning models to generalize well. By utilizing pre-trained architectures such as ResNet, EfficientNet, and DenseNet, enables the extraction of high-level visual features without requiring extensive labeled data. This approach reduces computational costs, accelerates training, and enhances model generalization. Additionally, transfer learning allows fine-tuning of specific layers, ensuring the model effectively adapts to domain-specific features in medical imaging, ultimately leading to more accurate and reliable disease detection. In our study, we choose six pre-trained models for eye disease classification. VGG16, DenseNet121, MobileNetV2, EfficientNetB0, ResNet50, and InceptionV3. Each of these models has distinct architectural characteristics that contribute to their performance in feature extraction and classification. We trained and tested our models on the NKS_EYE_DISEASE dataset. Each model was initialized with pre-trained ImageNet weights and fine-tuned by adding a Global Average Pooling (GAP) layer, a fully connected dense layer with 256 neurons and ReLU activation, followed by a final SoftMax classification layer. The models were compiled using the Adam optimizer and the categorical cross-entropy loss function, ensuring stable weight updates and efficient learning during backpropagation. During training, the models fit on the training data for 10 epochs with a batch size of 32, while validation data is used to monitor the performance and prevent overfitting. After training, each model is evaluated on the test set to assess its classification accuracy and performance metrics. The loss function used in this training process is categorical cross-entropy, which is designed for multi-class classification tasks. It measures the difference between the true class labels and the predicted probability distribution generated by the SoftMax function.

- **VGG16:** VGG16 is a deep CNN architecture consisting of 16 layers, primarily composed of 3×3 convolutional filters and max pooling layers. It follows a simple yet powerful sequential design, making it effective in extracting hierarchical visual features. Despite its large number of parameters, VGG16 is widely used for medical image classification due to its ability to capture fine-grained details in images.
- **DenseNet121:** DenseNet121 is a densely connected convolutional network that enhances gradient flow and feature reuse by connecting each layer to all subsequent layers. This architecture improves learning efficiency, reduces the number of parameters, and captures complex patterns within eye disease images. DenseNet's ability to extract detailed features makes it particularly useful for medical image analysis.
- **MobileNetV2:** MobileNetV2 is a lightweight deep learning model optimized for efficiency with depthwise separable convolutions and inverted residual blocks. It is computationally less expensive compared to other architectures, making it well-suited for real-time medical applications and deployment in resource-constrained environments such as mobile or edge devices.
- **InceptionV3:** InceptionV3 is a deep CNN architecture that employs parallel convolutional filters of different sizes within the same layer. This multi-scale feature extraction helps capture both fine and coarse details in images, making it particularly effective for medical imaging tasks where intricate patterns must be identified for accurate disease classification.

3.4. Proposed Deep Ensemble Model

Our proposed system is designed to facilitate automatic eye disease detection and classification through an ensemble of multiple CNN-based transfer learning models. The purpose of the ensemble system is to improve classification accuracy by leveraging the strengths of various deep learning models while preserving model interpretability and clinical reliability. A diagrammatic sketch of the proposed system is shown in Figure 3.

From the Figure 3 we can see that the initial step is the collection of retinal fundus images, which are the inputs to the system. To make them processable by deep learning models, a sequence of data preprocessing steps is carried out. These encompass image resizing to fit standard input sizes accommodating the pre-trained model requirements, and

label encoding to transform categorical disease labels into numerical format. SMOTE is also applied to address class imbalance in the dataset through the creation of synthetic samples for under-sampled classes, thereby accommodating equitable model training and enhancement of predictive capability with respect to minority classes. Following preprocessing, the images are input in parallel into four pre-trained transfer learning architectures: VGG16, DenseNet121, MobileNetV2, and InceptionV3. These architectures are pre-trained on large image datasets before being fine-tuned on the target retinal image dataset for learning task-specific features. These architectures are chosen due to their complementary strengths in learning rich, discriminative, and hierarchical visual features from medical images. Within every model, images undergo feature extraction by convolutional and pooling layers, followed by several fully connected layers meticulously crafted for disease prediction. The prediction units generally consist of a feature extraction layer, one or several dense (fully connected) layers, an activation function and a final output layer predicting the probability distribution across the potential disease classes. After individual predictions are obtained from each model, an ensemble method is used to decide the ultimate disease label. A majority-voting ensemble technique is utilized, in which the predicted labels of the models are counted and the class receiving the most votes is chosen as the final prediction. This approach to fusion exploits the varied decision-making tendencies of the individual classifiers and hence decreases the possibility of incorrect or biased predictions by any one classifier and enhances the overall strength of the system. After procuring probability scores for each of the classes from the VGG16, DenseNet121, MobileNetV2, and InceptionV3 models, we take on a probability averaging ensemble approach to determine the ultimate class label.

3.5. Performance Evaluation

We evaluated performance using various metrics to identify the most effective classifier for detecting eye diseases. Performance indicators, expressed as percentages (%), were calculated using Eqs. (1–3).

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total\ Number\ of\ Images} \times 100\% \quad (1)$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \times 100\% \quad (2)$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \times 100\% \quad (3)$$

Additionally, we generated a confusion matrix, AUC-ROC, train, and validation graph for each model to assess their performance comprehensively.

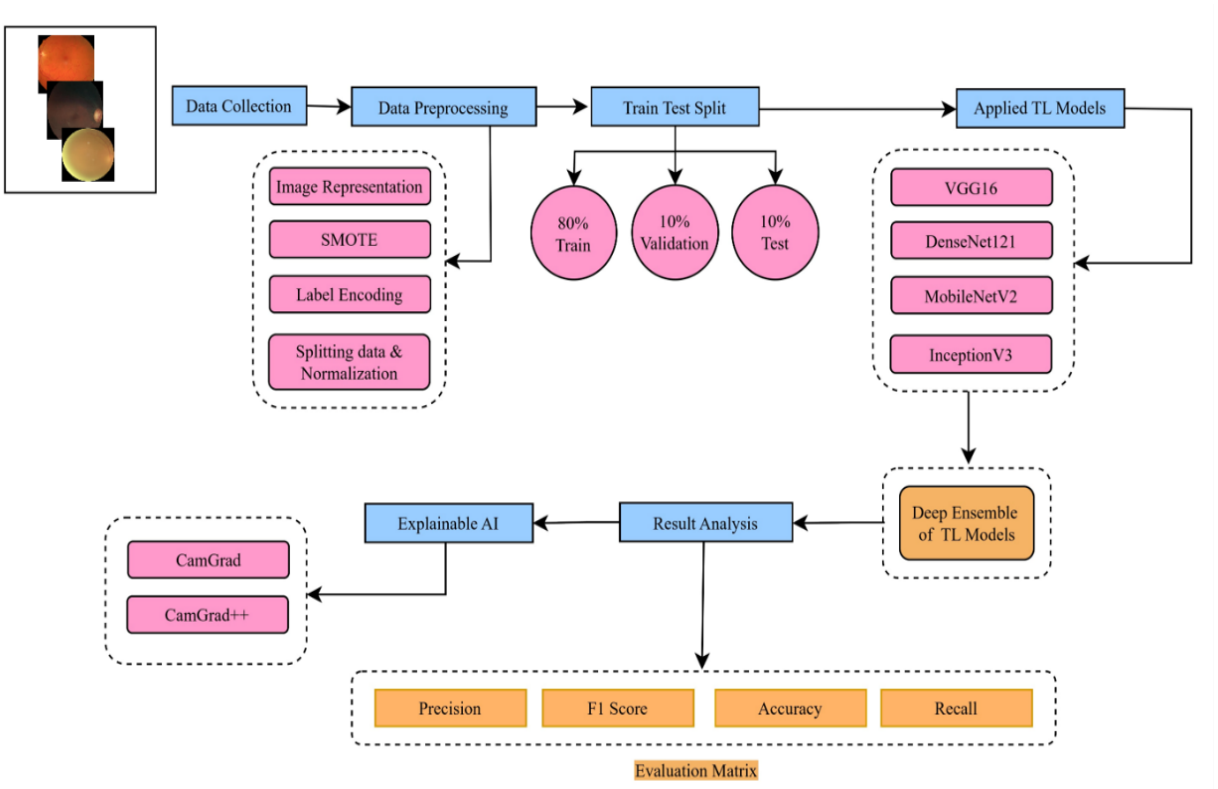

Figure 1. Step by Step Procedure Diagram of Workflow

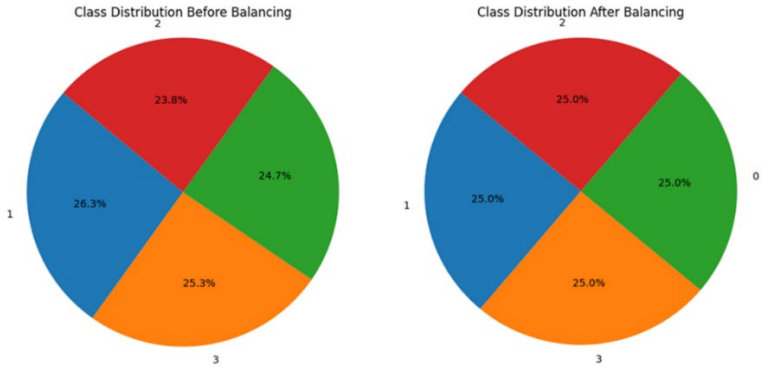

Figure 2. Class Distribution of Eye Disease Dataset Before and After SMOTE Balancing

Table 1. Class Distribution Before and After SMOTE

Class	Before SMOTE Count	After SMOTE Count
Cataract	937	997
Diabetic Retinopathy	997	997
Glaucoma	902	902
Normal	959	959

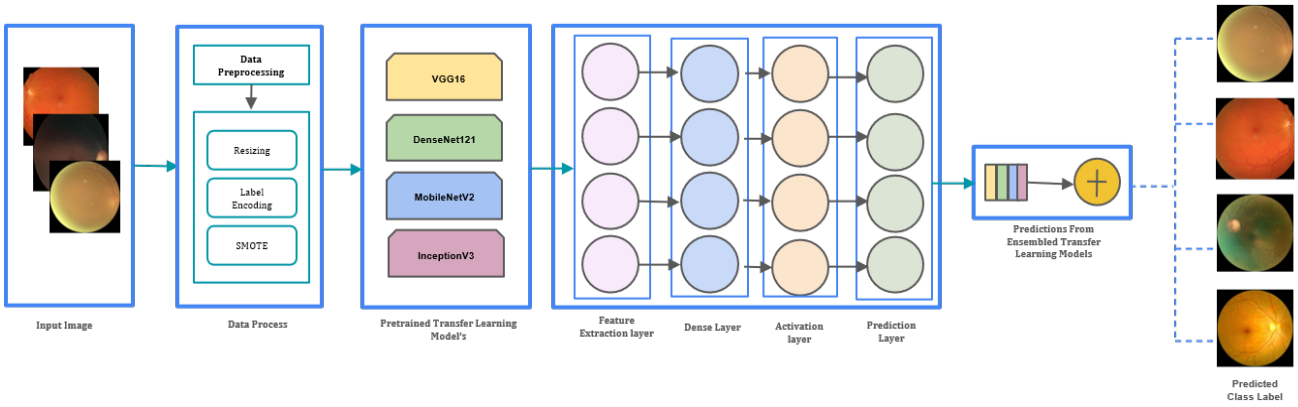

Figure 3. Proposed Framework of Ensemble Model

4. Result and Discussion

The experiments in this study were conducted using multiple pre-trained deep learning models for the classification task. The models used include VGG16, DenseNet121, MobileNetV2 and InceptionV3. All models were implemented using Python 3.12 and TensorFlow 2.12. Training was conducted on an CPU with 16 GB RAM. The coding and execution were performed in Google Colab, which provided a cloud-based platform to efficiently handle the computational requirements. The models were first initialized with ImageNet weights and fine-tuned on the NKS_EYE_DISEASE dataset. For model evaluation, various metrics were used, including accuracy, classification reports, confusion matrices, and Receiver Operating Characteristic (ROC) curves. The confusion matrix was calculated to assess true positive, false positive, true negative, and false negative counts for each class. To assess the ensemble model's performance, an average prediction was computed across the models. The accuracy of the ensemble model was evaluated, providing a comparative insight into the collective performance of the models. Finally, the results from the individual models were plotted in an accuracy comparison chart to visualize the overall performance of each model and the ensemble. The complete analysis was done using Python libraries, including TensorFlow, Keras, and Scikit-Learn.

We present the detailed evaluation of the performance of the proposed ensemble model, along with several individual models. The individual models tested were VGG16, DenseNet121, MobileNetV2, and InceptionV3, each showcasing different levels of performance in terms of accuracy. VGG16 achieved 82.27%, DenseNet121 led with the highest accuracy of 87.91%, MobileNetV2 attained 86.01%, and InceptionV3 recorded 82.22%. Following the

evaluation of these individual models, we proceeded with the ensemble model, which integrated the outputs of the models. The ensemble model demonstrated a significant improvement in performance, achieving an accuracy of 90%.

Table 2 gives a comparative evaluation of applied deep learning models VGG16, DenseNet121, MobileNetV2, InceptionV3, and a custom Deep Ensemble for classifying four eye disease classes: Cataract, Diabetic Retinopathy, Glaucoma, and Normal. The Deep Ensemble model achieved the highest combined accuracy of 90%, far outperforming the single models through pooling their prediction capability. It was highly consistent in all classes, with very high precision and recall for Diabetic Retinopathy (F1-score: 1.00) and satisfactory performance on Normal and Cataract too. The top-performing single model was DenseNet121, which achieved 87.91% accuracy and satisfactory F1-scores on most classes, demonstrating its robustness and efficiency in deep feature extraction. MobileNetV2, while lightweight, also worked efficiently with 86.01% accuracy and can be employed in low-resource environments like mobile platforms or edge computing in telemedicine applications. However, while moderate accuracies were achieved by VGG16 (82.27%) and InceptionV3 (82.22%), both were poor in detecting Glaucoma, which appears to be the most difficult class for all models due to its subtle and overlapping nature.

From Table 3 and Figure 4 we can see the True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) by class is given for five deep learning models: VGG16, DenseNet121, MobileNetV2, InceptionV3, and a Deep Ensemble model. These numbers give further insight into the capability of each model to distinguish between the four classes of eye diseases: Cataract, Diabetic Retinopathy, Glaucoma, and Normal. The Deep Ensemble model consistently demonstrates outstanding classification ability, especially for Diabetic Retinopathy, where it correctly classifies all positive samples (TP = 101, FN = 0) with only one false positive, demonstrating nearly perfect sensitivity and specificity. For Normal images, it is extremely sensitive (TP = 106) with a relatively low FN (9), which is good identification. DenseNet121 also performs well in all classes with similar TP and zero FN values—highlighting its strength, especially for Normal and Cataract detection. VGG16 and InceptionV3 are poorer with classification errors. VGG16 suffers especially with Glaucoma, with 41 false negatives and 23 false positives, reflecting inability to detect the fine characteristics of this disease. Similarly, InceptionV3 works poorly on Diabetic Retinopathy (TP = 83, FN = 18) and Glaucoma (TP = 71, FN = 34) and is therefore less precise for those classes. MobileNetV2, even though computationally more economical, is volatile: it works well on Cataract and Diabetic Retinopathy but gives higher values of FP for Glaucoma and Normal, which indicates certain trade-off between speed and accuracy.

Figures 5 present the ROC and Precision-Recall curves for all our models, illustrating their class-wise performance in distinguishing between categories. The performance of the VGG16 model, which achieves an AUC of 1.00 for Class 1, with a Precision-Recall AUC of 1.00, demonstrating perfect classification for this category. DenseNet121 and MobileNetV2, both attaining an AUC of 1.00 for Class 1 and a Precision-Recall AUC of 1.00, indicating consistently high performance. InceptionV3, which achieves a slightly lower AUC of 0.98 for Class 1 and 0.99 for Class 0, with corresponding Precision-Recall AUC values of 0.94 for Class 1 and 0.97 for Class 0, reflecting minor variations in the precision-recall trade-offs. The performance of the ensemble model, which integrates predictions from all architectures and maintains an AUC of 1.00 for Class 1, with a Precision-Recall AUC of 1.00, ensuring robust classification performance. The ensemble model effectively preserves the strengths of individual networks while balancing generalization across multiple classes.

Figures 6 shows the training and validation loss, along with accuracy curves, for the VGG16, DenseNet121, MobileNetV2, and InceptionV3 models over multiple epochs. VGG16 model demonstrates effective learning, with a progressive decrease in training loss and a similar downward trend in validation loss, suggesting good generalization. The accuracy curve shows a steady increase, with validation accuracy stabilizing at approximately 0.85, reflecting strong generalization capability. DenseNet121, where training loss consistently declines, while validation loss exhibits minor fluctuations before stabilizing. The accuracy plot reveals a significant improvement, with training accuracy reaching approximately 0.92 and validation accuracy stabilizing around 0.88, indicating strong model performance with slight overfitting. MobileNetV2 and InceptionV3 models, showing a steady decline in both training and validation loss. While MobileNetV2's validation loss stabilizes after a few epochs, InceptionV3 exhibits minor fluctuations, suggesting variations in learning stability. The accuracy trends reveal a consistent increase in training accuracy for both models, but the slight gap between training and validation accuracy in InceptionV3 suggests minor overfitting compared to MobileNetV2. These models exhibit effective learning, with DenseNet121 achieving the highest performance while InceptionV3 shows some instability in validation trends.

In medical image analysis, particularly in critical functions like classification of eye diseases, Explainable AI (XAI) stands out as the bridge between the deep learning black-box and clinical trust. While our deep learning model based on ensembling demonstrated superior accuracy for classifying retinal conditions such as cataract, diabetic retinopathy, glaucoma, and normal ones, clinical use requires something beyond performance measures—it requires transparency in decision-making. Misclassifications between visually similar diseases underscore the need for interpretability; XAI techniques such as Grad-CAM can be employed to visualize what regions of the retinal image had the greatest impact on a prediction. Not only does this enable verification and understanding of the thought process behind the model by ophthalmologists, but it also facilitates error analysis, supporting researchers in improving feature representations and reducing diagnostic uncertainty. Also, in real-world clinical scenarios where patient outcomes and trust are paramount, explanations of automatic decisions promote clinician confidence and adhere to regulatory standards for

AI-assisted diagnosis. To gain a clearer understanding of the misclassification behaviour of our proposed ensemble of transfer learning models for eye disease classification, we incorporate the Explainable AI and misclassified samples presented in Figures 7. The data set has four classes: Cataract, Diabetic Retinopathy, Glaucoma, and Normal. These misclassification trends indicate feature similarity between disease categories, which may have contributed to the model's confusion. The fundus images in Figure 7 illustrate specific misclassified cases. We can see through Explainable AI that which region effecting to misclassified, a close examination of these samples reveals that the visual overlap in retinal structures across different diseases makes it challenging for the model to distinguish them accurately.

Table 2. Performance Calculation of Class-wise Precision, Recall, F1-score, and Overall Accuracy for Individual and Ensemble Models

Models	Class	Precision	Recall	F1-score	Accuracy
VGG16	Cataract	0.83	0.87	0.85	82.27%
	Diabetic Retinopathy	0.98	0.99	0.99	
	Glaucoma	0.74	0.61	0.67	
	Normal	0.75	0.83	0.79	
DenseNet121	Cataract	0.84	0.91	0.88	87.91%
	Diabetic Retinopathy	0.99	0.97	0.98	
	Glaucoma	0.84	0.77	0.81	
	Normal	0.85	0.87	0.86	
MobileNetV2	Cataract	0.99	0.81	0.89	86.01%
	Diabetic Retinopathy	0.97	0.98	0.98	
	Glaucoma	0.74	0.80	0.77	
	Normal	0.80	0.85	0.82	
InceptionV3	Cataract	0.85	0.91	0.88	82.22%
	Diabetic Retinopathy	0.91	0.82	0.86	
	Glaucoma	0.83	0.87	0.85	
	Normal	0.98	0.99	0.99	
Deep Ensemble	Cataract	0.74	0.61	0.67	90%
	Diabetic Retinopathy	0.75	0.83	0.79	
	Glaucoma	0.84	0.91	0.88	
	Normal	0.99	0.97	0.98	

Table 3. Performance Comparison of True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) values for each class across all models.

Models	Class	TP	FP	TN	FN
VGG16	Cataract	88	18	303	13
	Diabetic Retinopathy	100	2	319	1
	Glaucoma	64	23	294	41
	Normal	95	32	275	20
DenseNet121	Cataract	92	17	304	9
	Diabetic Retinopathy	98	1	320	3
	Glaucoma	81	15	302	24
	Normal	100	18	289	15
MobileNetV2	Cataract	82	1	320	19
	Diabetic Retinopathy	99	3	318	2
	Glaucoma	84	30	287	21
	Normal	98	25	282	17
InceptionV3	Cataract	92	16	305	9
	Diabetic Retinopathy	83	8	313	18
	Glaucoma	71	16	301	34
	Normal	101	35	272	14
Deep Ensemble	Cataract	90	9	312	11
	Diabetic Retinopathy	101	1	320	0
	Glaucoma	80	14	303	25
	Normal	106	21	286	9

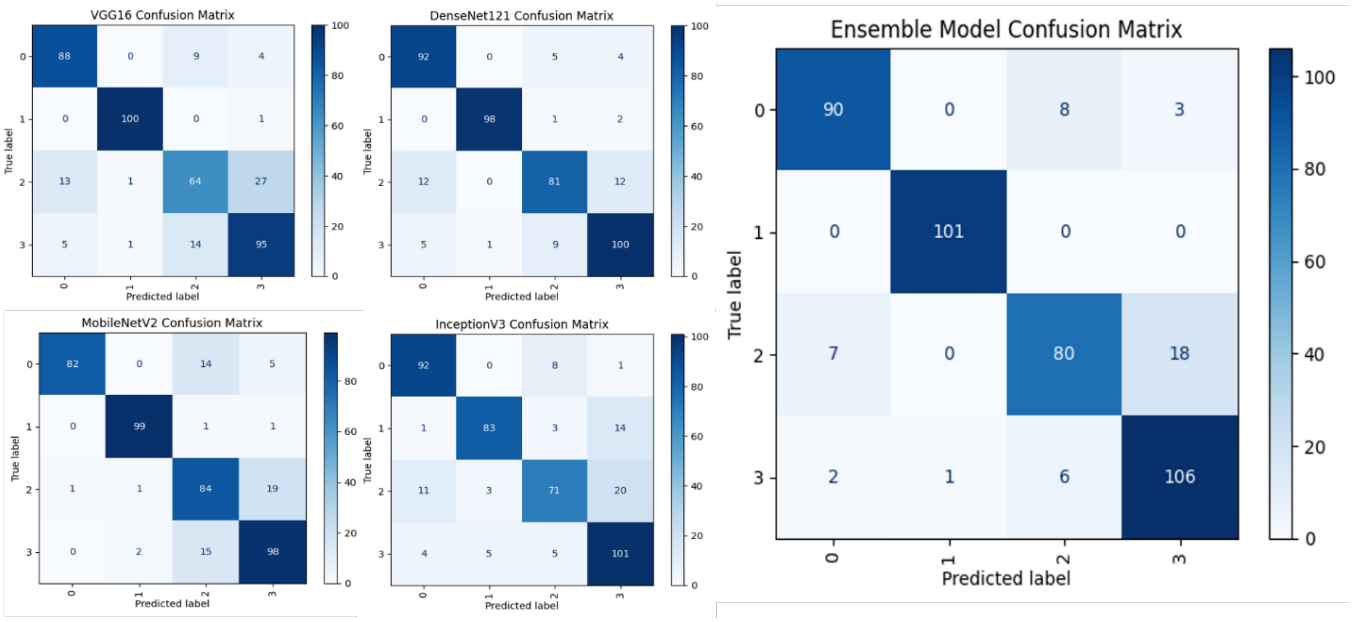

Figure 4. Confusion matrices for individual models and the ensemble model in multi-class eye disease classification.

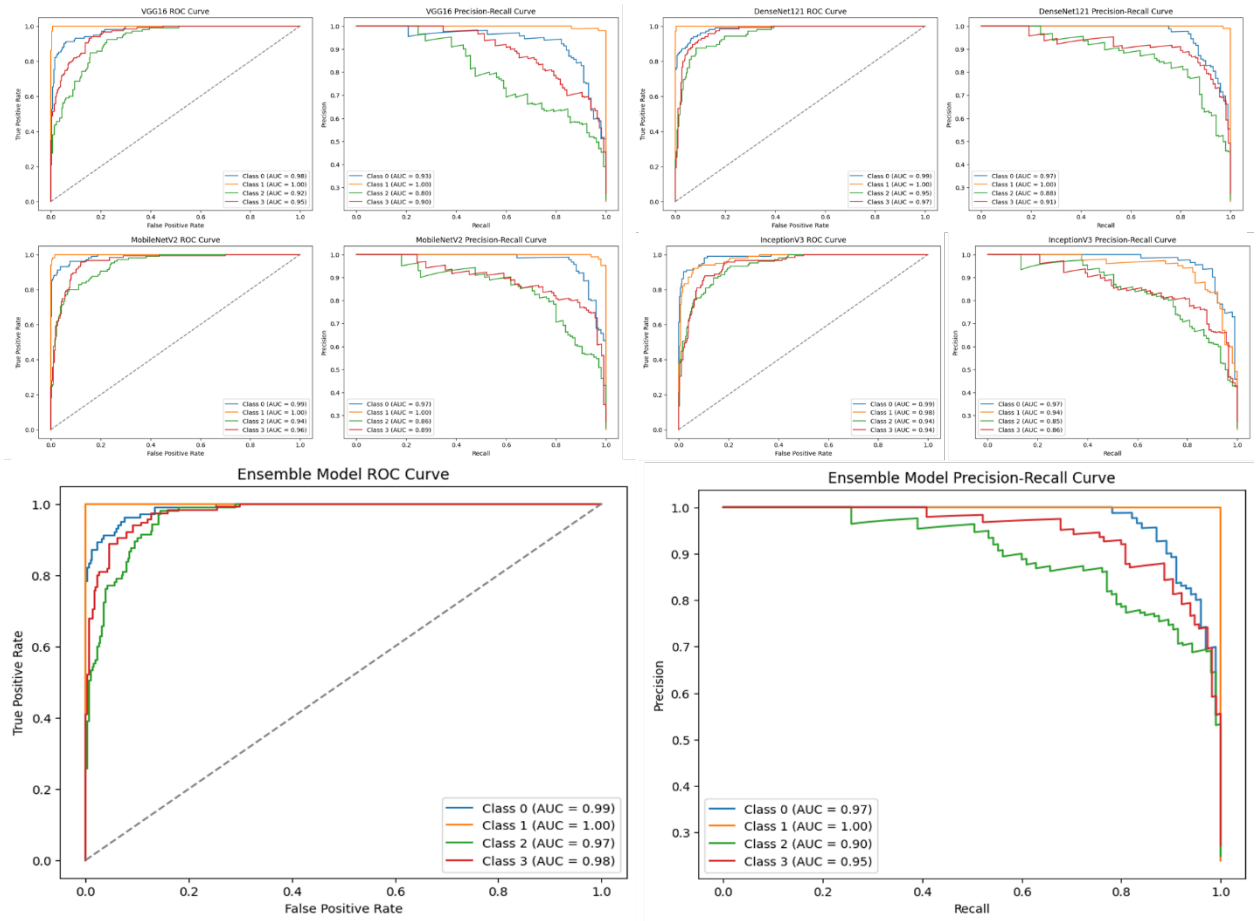

Figure 5. ROC and Precision-Recall curves for individual models and the proposed ensemble model.

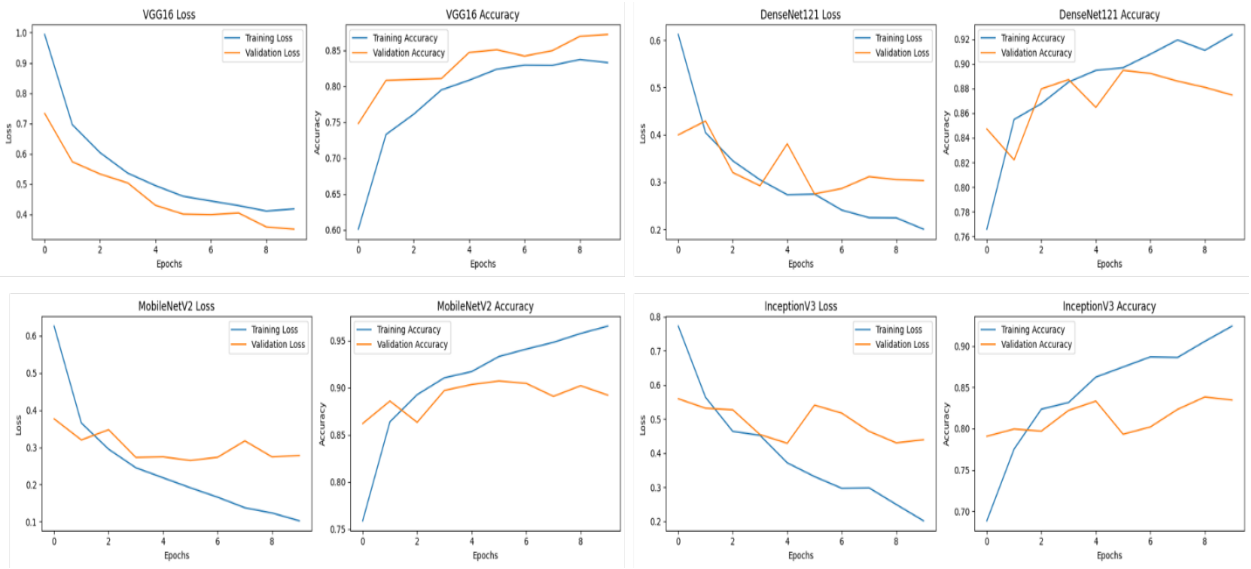

Figure 6. Training and validation loss and accuracy curves for the four individual transfer learning models.

The comparative study presents in Table 4 that ensemble models always perform better than standalone deep learning architecture in eye diseases detection problems. Studies [34] and [36], conducted on diabetic retinopathy using individual or simple models, achieved moderate accuracy of around 84%. Conversely, studies [35] and [37] utilized ensemble CNN and achieved higher accuracy (88.1% and 88.59%) with the combination strength of different models. The proposed study also follows this trend by achieving the highest accuracy (90%) for several eye diseases diabetic retinopathy, glaucoma, and cataracts demonstrating that an ensemble model well designed can generalize very well across a broad spectrum of ocular conditions and improve diagnostic performance. Furthermore, this work contributes uniquely by applying Explainable AI (XAI) techniques to provide further transparency, generalizability, and acceptability to the decisions of the model. In providing explainable insights into diagnostics, not only does the model achieve superior accuracy, but it also gains confidence from clinicians, hence becoming deployable in real-world clinical settings.

Cataract being misclassified as Glaucoma suggests that the model may struggle to differentiate between lens opacity and optic nerve abnormalities, as both conditions can exhibit similar retinal brightness and contrast patterns. Glaucoma being confused with Normal eyes may result from early-stage glaucoma cases lacking prominent symptoms, leading the model to classify them as healthy. Normal eyes being misclassified as Glaucoma could be attributed to minor retinal variations or image artifacts that resemble pathological features. These findings highlight the need for enhanced feature extraction techniques to improve differentiation between disease classes. Possible solutions include incorporating attention mechanisms to focus on key pathological features, multi-scale feature fusion to capture fine-grained differences, and data augmentation to improve generalization. Future work can also explore multi-label classification approaches to account for overlapping disease characteristics, further refining the model's performance in real-world clinical applications.

To investigate deeper the and check accurate prediction trends in our models, we utilize Explainable AI (XAI) with our Ensemble Model as well shows in Figure 8. The results indicate that regions around the optic disc, retinal vessels, and macular area are important in Glaucoma and Normal case predictions and therefore their susceptibility for accurate prediction. The top row shows segmentation maps which accentuate regions the model considers significant using techniques such as Grad-CAM. Color intensity (black to cyan/blue) illustrates various activation levels for our disease. The below row overlays these heatmaps on the retinal images, visually demonstrating the specific areas (e.g., lesions, optic disc) that the model was focusing on while making the accurate prediction. This is done to validate if the model is learning clinically relevant features and helps facilitate explainability and transparency in AI-driven diagnosis.

5. Conclusion

In this study, we proposed an ensemble-based deep learning approach for automated eye disease classification, leveraging four state-of-the-art transfer learning models: DenseNet121, MobileNetV2, VGG16, and InceptionV3. Experimental results showed that the ensemble model outperformed individual models in terms of classification accuracy, robustness, and generalization. Comparative analysis of training and validation curves, along with confusion matrix evaluations, highlighted its effectiveness in distinguishing between eye disease classes—cataract, diabetic retinopathy, glaucoma, and normal cases. However, misclassifications were noted, particularly between visually similar diseases such as cataract and glaucoma, underscoring the challenge of feature overlap in retinal images. To enhance model performance, we plan to incorporate attention

mechanisms and advanced feature fusion strategies. Despite strong results, the model has limitations, including reliance on a single publicly available dataset that may not reflect real-world diversity, and the risk of overfitting due to the depth of the ensemble architecture. Future work will address these issues by integrating multi-institutional datasets, applying cross-dataset validation, and exploring semi-supervised and self-supervised learning methods. Additionally, incorporating explainable AI (XAI) techniques will improve interpretability and foster trust among clinicians. Ultimately, we aim to deploy the proposed model in clinical environments and mobile-based screening applications to support early detection and diagnosis of eye diseases, enabling timely medical interventions.

Table 4. Comparative Analysis with Similar Context

Reference Study	Context	Disease	Best Model	Accuracy
[34]	Diabetic Retinopathy Detection	Diabetic Retinopathy	DLM2	84.19%
[35]	Glaucoma Stage Classification	Glaucoma	Ensemble of CNNs	88.1%
[36]	Diabetic Retinopathy Grading	Diabetic Retinopathy	Ensemble-based architecture	84%
[37]	Retinal Disease Detection	Multiple Retinal Diseases	Ensemble Learning	88.59%
This Study	Eye Disease Detection	Diabetic Retinopathy, Glaucoma, Cataracts, Normal	Deep Ensemble Model	90%

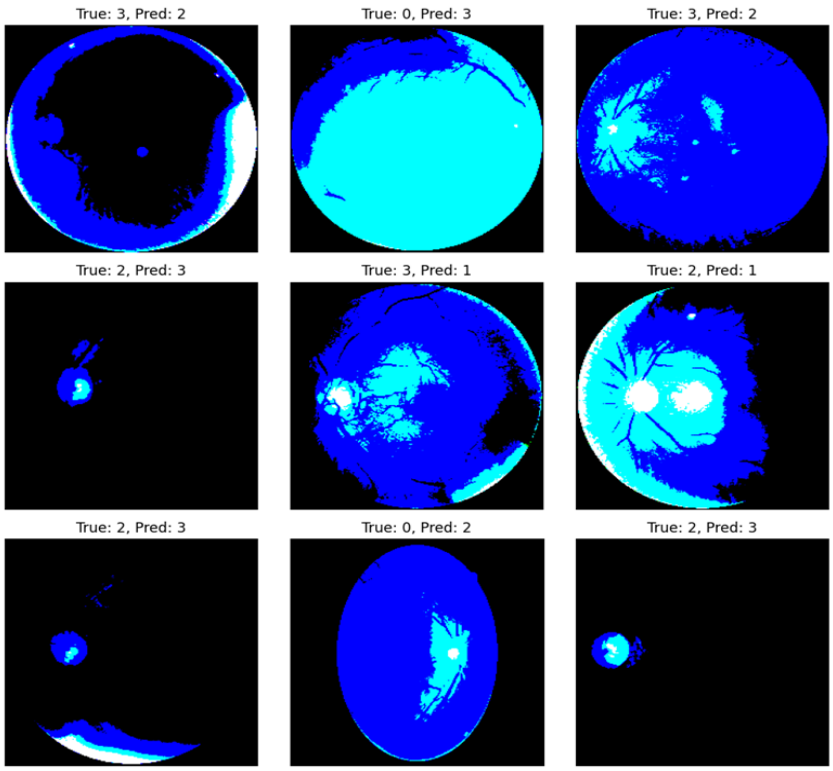

Figure 7. Misclassifications with Explainable AI for which Region Confusing Model

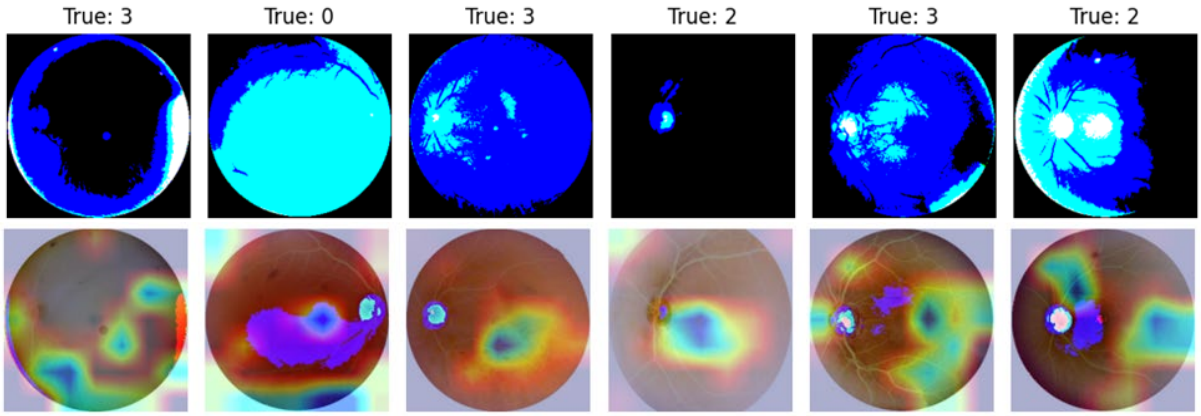

Figure 8. Explainable AI Visualization of Regions for Accurate Prediction

References

[1] Bitto AK, Mahmud I. Multi categorical of common eye disease detect using convolutional neural network: a transfer learning approach. Bulletin of Electrical Engineering and Informatics. 2022 Aug 1;11(4):2378-87.

[2] Niloy GM, Bitto AK, Biplob KB, Sammak MH, Das A, Hridoy GG. MobileNet-Eye: An Efficient Transfer Learning for Eye Disease Classification. In2024 International Conference on Advances in Computing, Communication, Electrical, and Smart Systems (iCACCESS) 2024 Mar 8 (pp. 01-05). IEEE.

[3] Bock R, Meier J, Nyúl LG, Hornegger J, Michelson G. Glaucoma risk index: automated glaucoma detection from color fundus images. Medical image analysis. 2010 Jun 1;14(3):471-81.

[4] Sarki R, Ahmed K, Wang H, Zhang Y. Automatic detection of diabetic eye disease through deep learning using fundus images: a survey. IEEE access. 2020 Aug 10;8:151133-49.

[5] N. E. institute, "Facts About Diabetic Eye Disease," https://nei.nih.gov/health/diabetic/retinopathy, 2015, [Online; Last Reviewed: September-2015].

[6] Prasad K, Sajith PS, Neema M, Madhu L, Priya PN. Multiple eye disease detection using Deep Neural Network. InTENCON 2019-2019 IEEE Region 10 Conference (TENCON) 2019 Oct 17 (pp. 2148-2153). IEEE.

[7] Abbas Q. Glaucoma-deep: detection of glaucoma eye disease on retinal fundus images using deep learning. International Journal of Advanced Computer Science and Applications. 2017;8(6).

[8] Dervisevic E, Pavljasevic S, Dervisevic A, Kasumovic SS. Challenges in early glaucoma detection. Medical Archives. 2016 May 31;70(3):203.

[9] Cvekl A, Vijg J. Aging of the eye: Lessons from cataracts and age-related macular degeneration. Ageing Research Reviews. 2024 Jul 6:102407.

[10] Bloemendal H, de Jong W, Jaenicke R, Lubsen NH, Slingsby C, Tardieu A. Ageing and vision: structure, stability and function of lens crystallins. Progress in biophysics and molecular biology. 2004 Nov 1;86(3):407-85.

[11] Roskamp KW, Paulson CN, Brubaker WD, Martin RW. Function and aggregation in structural eye lens crystallins. Accounts of chemical research. 2020 Apr 9;53(4):863-74.

[12] Lam D, Rao SK, Ratra V, Liu Y, Mitchell P, King J, Tassignon MJ, Jonas J, Pang CP, Chang DF. Cataract. Nature reviews Disease primers. 2015 Jun 11;1(1):1-5.

[13] Mrugacz M, Pony-Uram M, Bryl A, Zorena K. Current approach to the pathogenesis of diabetic cataracts. International Journal of Molecular Sciences. 2023 Mar 28;24(7):6317.

[14] Moraru AD, Costin D, Moraru RL, Branisteanu DC. Artificial intelligence and deep learning in ophthalmology-present and future. Experimental and therapeutic medicine. 2020 Oct;20(4):3469-73.

[15] Khan NC, Perera C, Dow ER, Chen KM, Mahajan VB, Mruthyunjaya P, Do DV, Leng T, Myung D. Predicting systemic health features from retinal fundus images using transfer-learning-based artificial intelligence models. Diagnostics. 2022 Jul 14;12(7):1714.

[16] Han D, Liu Q, Fan W. A new image classification method using CNN transfer learning and web data augmentation. Expert systems with applications. 2018 Apr 1;95:43-56.

[17] Morid MA, Borjali A, Del Fiol G. A scoping review of transfer learning research on medical image analysis using ImageNet. Computers in biology and medicine. 2021 Jan 1;128:104115.

[18] ul Hassan M, Al-Awady AA, Ahmed N, Saeed M, Alqahtani J, Alahmari AM, Javed MW. A transfer learning enabled approach for ocular disease detection and classification. Health Information Science and Systems. 2024 Jun 11;12(1):36.

[19] Chayan TI, Islam A, Rahman E, Reza MT, Apon TS, Alam MG. Explainable AI based glaucoma detection using transfer learning and LIME. In2022 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE) 2022 Dec 18 (pp. 1-6). IEEE.

[20] Cen LP, Ji J, Lin JW, Ju ST, Lin HJ, Li TP, Wang Y, Yang JF, Liu YF, Tan S, Tan L. Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. Nature communications. 2021 Aug 10;12(1):4828.

[21] Kumar A, Dhanalakshmi R. EYE-YOLO: a multi-spatial pyramid pooling and Focal-EIOU loss inspired tiny YOLOv7 for fundus eye disease detection. International Journal of Intelligent Computing and Cybernetics. 2024 Jul 17;17(3):503-22.

[22] Shibata N, Tanito M, Mitsuhashi K, Fujino Y, Matsuura M, Murata H, Asaoka R. Development of a deep residual learning algorithm to screen for glaucoma from fundus photography. Scientific reports. 2018 Oct 2;8(1):14665.

[23] Aurangzeb K, Alharthi RS, Haider SI, Alhussein M. Systematic development of AI-enabled diagnostic systems for glaucoma and diabetic retinopathy. IEEE Access. 2023 Sep 20;11:105069-81.

[24] Sarhan MH, Nasseri MA, Zapp D, Maier M, Lohmann CP, Navab N, Eslami A. Machine learning techniques for ophthalmic data processing: a review. IEEE Journal of Biomedical and Health Informatics. 2020 Jul 28;24(12):3338-50.

[25] Alam M, Le D, Lim JI, Chan RV, Yao X. Supervised machine learning based multi-task artificial intelligence classification of retinopathies. Journal of clinical medicine. 2019 Jun 18;8(6):872.

[26] AlBalawi T, Aldajani MB, Abbas Q, Daadaa Y. IoT-Opthom-CAD: IoT-Enabled Classification System of Multiclass Retinal Eye Diseases Using Dynamic Swin Transformers and Explainable Artificial Intelligence. International Journal of Advanced Computer Science & Applications. 2024 Jul 1;15(7).

[27] Madduri VK, Rao BS. Detection and diagnosis of diabetic eye diseases using two phase transfer learning approach. PeerJ Computer Science. 2024 Sep 19;10:e2135.

[28] Mahmood SS, Chaabouni S, Fakhfakh A. Improving Automated Detection of Cataract Disease through Transfer Learning using ResNet50. Engineering, Technology & Applied Science Research. 2024 Oct 9;14(5):17541-7.

[29] Guo C, Yu M, Li J. Prediction of different eye diseases based on fundus photography via deep transfer learning. Journal of Clinical Medicine. 2021 Nov 23;10(23):5481.

[30] Aranha, G. D., Fernandes, R. A., & Morales, P. H. (2023). Deep transfer learning strategy to diagnose eye-related conditions and diseases: an approach based on low-quality fundus images. *IEEE Access*, *11*, 37403-37411.

[31] Sanamdikar ST, Shelke MV, Rothe JP. Enhanced Classification of Diabetic Retinopathy via Vessel Segmentation: A Deep Ensemble Learning Approach. Journal homepage: http://iieta. org/journals/isi. 2023 Oct 1;28(5):1377-86.

[32] Desiani A, Primartha R, Hanum H, Dewi SR, Suprihatin B, Al-Filambany MG, Suedarmin M. Weighted Voting Ensemble Learning of CNN Architectures for Diabetic Retinopathy Classification. Jurnal Infotel. 2024 Feb 19;16(1):136-55.

[33] 1. Kumar SN. Nks9/NKS_EYE_DISEASE_CLASSIFICATION · Datasets at hugging face [Internet]. [cited 2025 Jun 17]. Available from: https://huggingface.co/datasets/nks9/NKS_EYE_DISEASE_CLASSIFICATION

[34] Chen PN, Lee CC, Liang CM, Pao SI, Huang KH, Lin KF. General deep learning model for detecting diabetic retinopathy. BMC bioinformatics. 2021 Nov;22:1-5.

[35] Cho H, Hwang YH, Chung JK, Lee KB, Park JS, Kim HG, Jeong JH. Deep learning ensemble method for classifying glaucoma stages using fundus photographs and convolutional neural networks. Current eye research. 2021 Oct 3;46(10):1516-24.

[36] Mehboob A, Akram MU, Alghamdi NS, Abdul Salam A. A deep learning based approach for grading of diabetic retinopathy using large fundus image dataset. Diagnostics. 2022 Dec 7;12(12):3084.

[37] Muchuchuti S, Viriri S. Retinal disease detection using deep learning techniques: a comprehensive review. Journal of Imaging. 2023 Apr 18;9(4):84.