

An EfficientNetB3 Approach for Retinal Disease Classification with XAI and Online Interface

Sadikur Rahman Sadik

Department of Internet of Things and Robotics Engineering, University of Frontier Technology, Bangladesh, Kaliakair, Gazipur-1750, Dhaka, Bangladesh
E-mail: sadikurrahmanssc@gmail.com
ORCID iD: <https://orcid.org/0009-0007-8558-1156>

Md. Abdul Halim Khan

Department of Internet of Things and Robotics Engineering, University of Frontier Technology, Bangladesh, Kaliakair, Gazipur-1750, Dhaka, Bangladesh
E-mail: mahknkn@gmail.com
ORCID iD: <https://orcid.org/0000-0003-3698-610X>

Samsuddin Ahmed

Department of Internet of Things and Robotics Engineering, University of Frontier Technology, Bangladesh, Kaliakair, Gazipur-1750, Dhaka, Bangladesh
E-mail: samsuddin0001@uftb.ac.bd
ORCID iD: <https://orcid.org/0000-0002-1267-0533>

Atiqur Rahman

Department of Internet of Things and Robotics Engineering, University of Frontier Technology, Bangladesh, Kaliakair, Gazipur-1750, Dhaka, Bangladesh
Department of Computer Science and Engineering, Daffodil International University, Dhaka 1216, Bangladesh
E-mail: 1801025@iot.uftb.ac.bd, atiqurrahmannayeem@gmail.com
ORCID iD: <https://orcid.org/0009-0009-9306-6976>

Sadia Enam

Department of Internet of Things and Robotics Engineering, University of Frontier Technology, Bangladesh, Kaliakair, Gazipur-1750, Dhaka, Bangladesh
E-mail: sadia0001@uftb.ac.bd
ORCID iD: <https://orcid.org/0009-0001-3424-8042>

Md. Toukir Ahmed*

Department of Internet of Things and Robotics Engineering, University of Frontier Technology, Bangladesh, Kaliakair, Gazipur-1750, Dhaka, Bangladesh
E-mail: toukircse14@gmail.com, toukir0001@uftb.ac.bd
ORCID iD: <https://orcid.org/0009-0002-3816-8602>
*Corresponding author

Received: January 13, 2026; Revised: April 10, 2026; Accepted: June 21, 2026; Published: August 8, 2026

Abstract: In real-world clinical settings, the growing number of patients and the shortage of experienced ophthalmologists make early and accurate diagnosis of retinal diseases increasingly challenging. Cataracts, diabetic retinopathy, and glaucoma are some of the most common causes of lifelong blindness around the world. This is why there is a need for automated diagnostic systems that can accurately diagnose and interpret clinical data. The major goal of this work is to find out if a deep learning architecture based on EfficientNetB3 and Explainable Artificial Intelligence (XAI) can accurately classify multiple types of retinal diseases while still being clear to doctors. The proposed system categorizes retinal fundus images into four groups: cataract, diabetic retinopathy, glaucoma, and normal. The dataset consisted of a balanced and publicly accessible collection of 4,217 retinal fundus pictures, processed using standard preprocessing techniques to enhance their generalizability. We chose EfficientNetB3 as the main architecture since it is better at extracting features, and we compared it to the standard convolutional neural network baselines to show how

useful it is. The proposed model achieved an accuracy of 97%, outperforming ResNet50 (91%) and VGG16 (87%). The high precision, recall, and F1-scores (0.94 – 1.00), the Cohen's kappa of 0.95, and the low logarithmic loss of 0.10 all point to reliable predictions. The receiver operating characteristic (ROC) analysis yielded an AUC of 1.00 across all disease classes. Gradient-weighted Class Activation Mapping (Grad-CAM) was utilized to address the interpretability deficit in deep learning-based medical systems and to pinpoint clinically significant retinal regions that influence model predictions. The results indicate that employing XAI alongside EfficientNetB3 enhances both diagnostic precision and interpretability, hence validating its suitability as a transparent decision-support system for the automated screening of retinal disorders.

Index Terms: Explainable AI, Deep Learning, Eye Disease Diagnosis, Medical Imaging, Ophthalmology.

1. Introduction

In recent years, the application of deep learning (DL) models to the interpretation of medical images has grown significantly. These models are particularly useful for diagnosing and treating a wide range of disorders. Deep learning has revolutionized computer vision (CV), enabling unprecedented advancements in disease identification and categorization. However, a significant challenge remains: medical practitioners often require assistance in interpreting the predictions of these algorithms due to their complexity. Retinal image analysis is a critical component of computer vision with applications in medical diagnosis, surveillance, biometrics, and human-computer interaction. Convolutional Neural Networks (CNNs) have long been the state-of-the-art approach for object detection and have fundamentally transformed computer vision. Recently, there has been significant interest in novel transformer architectures due to their ability to model complex relationships in input data using self-attention mechanisms. By effectively extracting retinal features and learning robust representations, these models have achieved state-of-the-art performance in various computer vision tasks, including retinal disease classification. While these results are promising, the lack of interpretability in conventional deep learning models may lead to misdiagnoses and inappropriate treatment plans, potentially compromising patient safety. The World Health Organization (WHO) reports that a significant number of individuals worldwide suffer from vision impairment. Age-related macular degeneration (AMD), diabetic retinopathy (DR), glaucoma, and cataracts are among the most common causes. Due to population growth, aging, and increasing diabetes prevalence, the incidence of these eye diseases is expected to rise globally. This study addresses the interpretability challenge by integrating Explainable AI (XAI) methods into a novel system for detecting and classifying retinal diseases. XAI provides medical professionals with deeper insights into DL model predictions, thereby fostering trust in the system. The proposed approach involves training the model on extensive medical image datasets to accurately detect a wide range of retinal diseases. To optimize accuracy, techniques such as data augmentation and careful hyperparameter tuning are employed. The use of XAI methods enhances model transparency, enabling medical professionals to understand the rationale behind specific predictions. The results demonstrate that this method outperforms traditional models in terms of accuracy, interpretability, and robustness. This represents a novel contribution to the field of medical imaging. The integration of the proposed model and XAI approaches has the potential to significantly impact medical imaging, contributing to more efficient and effective clinical decision-making. This technique addresses the growing reliance on medical imaging in healthcare and the need for diagnostic models that are both accurate and interpretable. This study advances medical image analysis by overcoming the limitations of traditional deep learning models. It also lays the groundwork for future research into how to use XAI approaches in healthcare systems. The results suggest that improved early diagnosis of retinal diseases could enhance patient outcomes and reduce the global prevalence of vision loss.

This research primarily aims to enhance the diagnosis and interpretation of eye diseases using deep learning and Explainable AI methods. The specific objectives of the study are:

- To develop a deep learning model that accurately detects and classifies eye diseases using a large set of medical imaging data.
- To integrate XAI techniques (such as SHAP and LIME) into the model to improve its interpretability, helping medical practitioners understand the rationale behind the predictions.
- To evaluate the effectiveness of the proposed approach in terms of accuracy, interpretability, and robustness when applied to real-world medical data.

Following the introduction, Section 2 provides a detailed literature review on recent works performed in the areas of medical image processing and eye disease classification, which describe existing methodologies and gaps in the literature. Section 3 describes the methods used in this study, including the design of the deep learning model, integration of Explainable AI (XAI) techniques, and data augmentation strategies to enhance performance and interpretability. The implementation and experimental setup are described in Section 4, including the hardware configuration, the model training and testing procedures, and the interpretation of the results. Performance evaluations

and comparisons with existing models are presented in Section 5, and the discussion of the results is included in this section. The final section, Section 6, concludes the paper by reviewing the main findings and discussing future work in eye disease detection and the application of XAI for healthcare improvement.

2. Literature Review

Mishra et al. [1] proposed machine learning as a means for the early detection of various eye disorders, such as diabetic retinopathy, glaucoma, and cataracts. Their CNN based model was 91% accurate in classifying eye images. To improve health-assistive usability, they further used XAI methods and developed a user-friendly intuitive GUI for real-time diagnostics. In another study, Sharma Gupta et al. [2] utilized a deep learning technique based on the VGG19 model to classify healthy and unhealthy retinal images. The technique was geared toward eye diseases such as diabetic retinopathy, glaucoma and cataracts. They achieved an accuracy of 95% using a pre-trained CNN. Transfer learning is a viable method for improving early detection in medical imaging, and especially in the field of ophthalmology, researchers reported. Aya A. Abd El-Khalek et al. [3] designed a computer-aided diagnostic (CAD) system for classifying stages of age-related macular degeneration (AMD) using fundus images. In local and global texture analysis, GLCM and GLRLM feature extraction methods were used in the proposed system. Photos were then classified into the appropriate category (geographic atrophy (GA), intermediate AMD, wet AMD or normal) using a maximum weighted majority voting method. It did so with 96.85% accuracy. This work demonstrated that the fused dataset of handwritten features with ensemble models could assist physicians in the rapid and precise diagnosis of AMD. Nouf Badah et al. [4] presented an automated approach to glaucoma detection using classical machine learning and deep learning methods. They evaluated the outputs of many popular classifiers, including SVM, KNN, NB, MLP, DT, and RF, in order to find the optimum method for reliable diagnosis. The CNN outperformed other models, with an accuracy exceeding 84% in their results. RF and MLP also performed similarly, achieving an accuracy of 77%. In this work, it is demonstrated the effectiveness of Deep learning methods, especially CNNs, in detecting eye diseases from medical image databases. Guo et al. [5] proposed a machine-based system for the bulk screening and grading of cataracts. A strategy using enhanced texture features and LDA achieved a 84.8% success rate on a clinical database. Yang et al. developed a three-step strategy for the effective automatic detection of cataracts, using a top-hat and bottom-hat modification to emphasize the differences between foreground and background. They used both brightness and texture as attributes and BBNN for classification of cataracts into mild, medium, and severe classes. Wiwik Fuadah et al. [6] built a computer-aided classification system for cataracts using feature extraction based on fundus images, wavelet transform, and sketch-based methods. They achieved 90.9% accuracy by using wavelet transform-based characteristics. However, KNN was able to distinguish normal from cataract cases with 94.5% accuracy, but it required a different methodology using statistical texture classification and human segmentation. Yang et al. [7] then proposed an ensemble learning model to detect and grade cataracts. Cataracts were detected with an accuracy of 93.2% and graded with 84.5% accuracy using this method. Caixinha et al. [8] proposed a method for the real-time diagnosis of cataracts using ultrasonic imaging and machine-learning algorithms. They evaluated different classifiers (SVM, CatBoost, etc.) It is infallible though costly and invasive. An SVM classifier was used to classify fundus images in a different study and the model achieved an accuracy of 93.33% in detecting the severity of cataracts. A smartphone-based app for self-screening for cataracts was also developed that used texture analysis and was expected to be 85% accurate. Jagadale et al. [9] used Hough circle detection to determine the center and diameter of the lens. With the help of an SVM classifier they successfully found cataract with a 90.25% accuracy. Sigit et al. [13] employed a single-layer perceptron to detect cataracts using Android smartphones with 85% accuracy. An advanced approach employs hierarchical feature extraction and neural networks to simplify cataract grading. The detection and grading accuracies were 94.83% and 85.98%, respectively [10]. Cataracts have been commonly detected by DL methods recently. Gao et al. proposed a deep learning-driven approach for grading nuclear cataracts. Ran et al. recommended a shallow CNN for cataract detection, achieving an accuracy of 93.52%. A six-level cataract grading method with 90.69% accuracy was suggested in [11], based on combining Random Forests and Deep CNNs. An automated cataract detection was developed using transfer learning on a CNN (accuracy: 92.91%). Jun et al. [12] described a cataract grading system that employed a Tournament-based RankedCNN and binaryCNN model. Hossain et al. proposed an automated detection method based on DCNNs using ResNet, achieving an accuracy of 95.77%. Recently, Zhang et al. proposed an attention-based multi-model ensemble method for detecting cataracts in ultrasound images. This approach had three distinct categorization networks plus an object recognition network. However, the model struggled due to limited training data and an inability to distinguish between various eye diseases. Recently, Pratap and Kokil [14] investigated the localization of cataracts using CNNs with various support vector configurations in noisy environments. Their results displayed the robustness of the models against noise and emphasized that the cataract detection models should be more accurate, yet less complex.

3. Materials and Methodology

3.1. Dataset Description

We used a publicly available and well-balanced retinal fundus image dataset comprising 4,217 images, collected from multiple sources including IDRiD, Ocular Disease Intelligent Recognition (ODIR), and the High-Resolution

Fundus (HRF) database. The dataset consists of 1,038 cataract images, 1,098 diabetic retinopathy images, 1,007 glaucoma images, and 1,074 normal images, ensuring class balance. This dataset, which features well-balanced classes, is available on Kaggle [15]. Scaling and normalization were the most critical operations in image preprocessing for analysis. Resizing all images to 224×224 pixels provides the neural network with a consistent input, facilitating model optimization. Normalization stabilizes the data distribution and accelerates convergence by scaling pixel values to the range $[0,1]$. Data augmentation enhances the model's generalization capability. For example, horizontal flipping and a custom preprocessing function were employed. Horizontal flipping increases data diversity and helps the model learn orientation-invariant features. The custom preprocessing function also allows for easy updates to the image pipeline for future optimization. Fig. 1 shows the pie chart of the class distribution in the dataset.

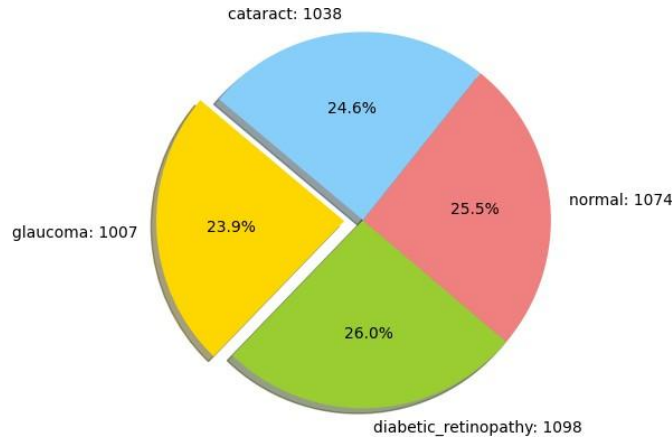


Fig.1. Class distribution in dataset

A. Data Prepossessing Batch Normalization

Batch normalization [16] is performed following the EfficientNetB3 feature extractor to normalize the 1536-dimensional feature vector. It normalizes the features and consequently adjusts their mean and variance for each batch, minimizing internal covariate shift. That serves to stabilize and accelerate training, enables the use of larger learning rates, and increases the flow of gradients. Batch normalization also introduces two learnable parameters: scale and shift, which allow the model to learn the optimal feature distributions. In general, it improves the stability of the training and perform regularization effect in addition to dropout.

B. Transfer Learning

The model applies transfer learning [17] by using EfficientNetB3, which has been pre-trained over the ImageNet dataset, as a feature extractor. Rather than starting from scratch, it uses pre-trained weights which have already learned general image patterns. The base EfficientNetB3 extracts meaningful features from the retinal images, and only the classification layers affixed are trained on the respective dataset. This method not only accelerates training, but also allows greater accuracy, especially with limited data.

C. Regularization

To prevent overfitting, we also apply regularization techniques [18], using weight decay L2 and regularization L1 over the biases. The L2 version of the regularizer will discourage very large weights, and the L1 version of the regularizer will help with the selection of features (promoting sparsity). These techniques help to handle model complexity, which allows models to generalize better to unseen data.

D. Dropout

The model also comprises a 0.45 rate dropout layer after a 256 dense layer. Dropout [19] happens during training time by deactivating randomly 45% of the neurons in this layer. By making it more difficult for neurons to learn from one another, this prevents the model from overfitting. Through this, the model learns more general and stable features. Dropout is disabled during inference such that all the neurons work together to make the final prediction.

Table 1. Architecture of EfficientNetB3 model

Layer Type	Output Shape	Parameters / Details	Notes
Input	(224, 224, 3)	—	Input image size and channels
EfficientNetB3 (base)	(None, 1536)	Pretrained on ImageNet	Main feature extractor, convolutional backbone with global max pooling
BatchNormalization	(None, 1536)	~6,144 (depends on shape)	Normalizes output for stability
Dense	(None, 256)	~393,472 (weights +	Fully connected layer with L1 & L2 regularization and

		bias)	ReLU activation
Dropout (0.45)	(None, 256)	0	Regularization to reduce overfitting
Dense (Softmax)	(None, 4)	~1,028	Output layer for 4-class classification

3.2. Proposed Machine Learning Model

The image classifier model is derived from the EfficientNetB3 architecture, a cutting-edge efficient CNN that scales network breadth, depth, and resolution in an efficient manner. RGB images of size 224×224 pixels are fed into the model as input, in adhering to EfficientNetB3 input expectations. The model basis lies in EfficientNetB3 that has been pretrained over the ImageNet dataset with removal of classification heads (include-top=False). The base framework of this model transforms input images through a series of convolution layers and Mobile Inverted Bottleneck Convolution (MBConv) blocks to extract complex hierarchical feature representations at diversified scales. Subsequently, a global max pooling layer squeezes spatial feature maps to a settled size of 1536-dimensional feature vector, encapsulating predominant information of input image. In an effort to add stability to training and to boost overall performance of model, it has a Batch Normalization layer in place succeeding to that of EfficientNetB3 basis. After this layer, a densely connected layer of 256 neurons has been included with L1 and L2 regularization in place to avoid overfitting as well as model generalization. The layer has an activation function for ReLU which renders the model non-linear. There has been utilization of Dropout layer with rate 0.45 for regularization through dropout of neurons randomly during their training and hence their inter- dependencies. The output layer has a Dense layer with neurons equal to target classes (four in this case), with softmax activation function for classifying to create a class of probability distribution. There has been usage of Adamax optimizer with an initialization of learning rate of 0.001 for multi-class classification problem through categorical crossentropy loss, and monitored through accuracy metrics while it trains.

3.3. Model Selection Rationale

The deep learning architectures evaluated in this study were selected to represent diverse design strategies and performance characteristics in image classification. EfficientNetB3 was chosen for its compound scaling approach that balances accuracy and efficiency. ResNet50 and DenseNet121 were included due to their proven ability to learn deep and discriminative features in medical imaging tasks. MobileNetV2 was selected to assess performance under computational constraints, making it suitable for deployment in resource-limited settings. VGG16, although an older architecture, was included as a classical baseline to provide a comparative reference against more recent models. The previously included Recurrent Neural Network (RNN) was removed, as RNNs are not well-suited for static image classification tasks. This selection enables a comprehensive comparison of accuracy, efficiency, and architectural complexity across different CNN designs.

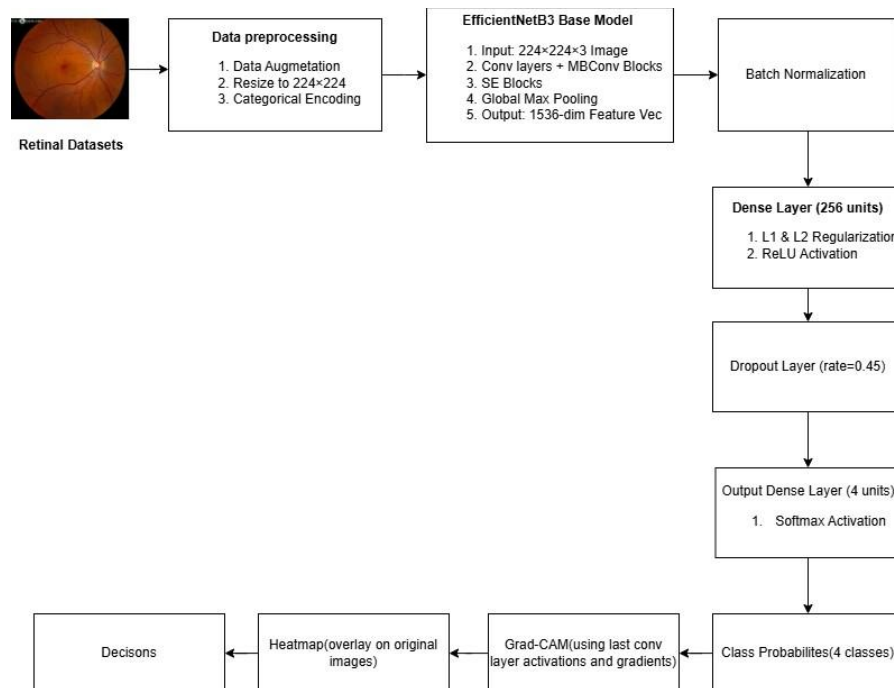


Fig.2. Model diagram of the proposed EfficientNetB3 architecture employed in this study for eye disease detection and classification

Fig. 2. delineates overall design as well as explainability pipeline of our classification system for retinal images. For the retinal datasets, images have to undergo data preparation with procedures for data augmentation to expand

training data, make all the photos the same size, 224×224 pixels, which is the right size for the model, and categorical encoding of class labels. The preprocessed images are fed into the base model of EfficientNetB3, which consists of a convolutional neural network architecture of convolutional layers and MBConv or Mobile Inverted Bottleneck Convolution and Squeeze- and-Excitation (SE) blocks. The model extracts a 1536-dimensional feature vector via global max pooling. The feature vector undergoes Batch Normalization to normalize it and stabilize and speed up training through normalization of layer inputs. Normalized features are fed into a Dense layer of 256 units with L1 and L2 regularization to avoid overfitting and ReLU activation for the introduction of non-linearity. There is also a Dropout layer of 0.45 to avoid overfitting through dropout of neurons at random during training. There's a final Dense layer of four units that employ a softmax activation function to give class probabilities for the four retinal diseases. Grad-CAM for interpretability enhancement analyzes the predictions of the model and computes gradients with respect to activations of the terminal convolutional layer to generate heatmaps. The heatmaps are overlaid over source images in an endeavor to highlight through visual means key areas that are most important in terms of making the categorization decisions.

4. Implementation and Experimental Setup

4.1. Metrics for Model Evaluation

We utilize the following measures to see how well our proposed model works:

A. Confusion Matrix

The confusion matrix [20] is a chart that shows how many times each category shows up in a dataset. These come in four types: True Positive (TP), False Negative (FN), False Positive (FP), and True Negative (TN).

Table 2. Confusion matrix displaying the performance of the EfficientNetB3 model

Actual Class	Predicted Class	
	Positive (1)	Negative (0)
Positive	True Positive (TP)	Negative (FN)
Negative	Positive (FP)	True Negative (TN)

B. Accuracy

The accuracy is the number of accurately predicted cases (true positives and true negatives) divided by the total number of examples looked at. It indicates you how accurate a categorization model is in general. The way to figure it out is:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (1)$$

C. Precision

Precision [21] tells us what percentage of the images the model correctly identified as positive out of all the images it considered were positive. You may figure out precision by:

$$Precision = \frac{TP}{TP+FP} \times 100\% \quad (2)$$

D. Recall

The recall [22] is the quantity of samples that were correctly classified positive divided by the total amount of properly categorized positive specimens and incorrectly classified negative samples. It checks to see how well a model can find all the good examples. The model is more sensitive and has fewer false negatives if it has a higher recall. The formula for figuring out recall is:

$$Recall = \frac{TP}{TP+FN} \times 100\% \quad (3)$$

E. F1 Score

The F1 score [23] is only one number that finds the harmonic mean of accuracy and recall. It seeks a balance between precision and recall, which tells you how precise a model is, particularly when the classes are not the same size. You can find out the F1 score by:

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \times 100\% \quad (4)$$

F. Kappa Score

The kappa score [24], often called Cohen's kappa, shows how much two raters or classification techniques agree with one other, taking into account the likelihood that they may agree by chance. It can be anywhere from -1 to 1, with 1 being complete agreement, 0 meaning no better than random chance, and negative values meaning disagreement. Kappa score can be considered as:

$$Kappa = \frac{p_o - p_e}{1 - p_e} \quad (5)$$

G. ROC Curve

The ROC curve [25] shows the rate of true positives and false positives for differing levels of threshold to assess the effectiveness of a binary classifier. The AUC (Area under Curve) of this plot gives a single measure that shows how effective in class separation a model is. Larger AUC shows higher discrimination.

H. Logarithmic Loss

Logarithmic loss, or log loss, measures how well a classification model's predicted probabilities match the actual labels. It penalizes false predictions more when the model is confident but wrong, producing a lower loss for accurate and confident predictions. Lower log loss values indicate better performance of the model [26]. Log loss can be calculated as:

$$LogLoss = -\frac{1}{N} \sum_{i=0}^N \sum_{j=1}^M x_{ij} \times \log(p_{ij}) \quad (6)$$

4.3. Interpreting the Model Performance

Explainable AI The goal of using Explainable AI [27] to make machine learning algorithms more clear and understandable for human users, fostering trust in AI systems by clarifying their managerial processes. This is of particular significance in sectors such as autonomous transportation and healthcare, where AI decisions can matter a great deal. XAI enables users to make sense of how an AI model functions, establishes transparency in explaining its predictions, increases its trustworthiness, and uncovers biases to enable fair decision-making. A crucial methodology employed in XAI is Grad-CAM that enhances CNNs' interpretability. Grad-CAM makes a visual explanation by showing which aspects of the image are crucial for the model's choice. It is to find the gradients of the target class score in relation to the feature maps, and then average over the plane of the feature maps to obtain the weight of them. The end product is a heat map which gives a visual indication of the important regions of the image. Furthermore, in medical AI, for example, in predicting eye diseases, the use of Grad-CAM was very useful. It assists the clinicians in knowing which parts of an eye scan image affect the diagnostic decisions of the model keeping medically meaningful features first. This interpretability increases confidence in AI diagnosis, and validates the efficacy of the model, which helps ensuring proper patient care.

4.4. Experimental Setup

To demonstrate our proposed model, we have to apply some setups which are given below:

A. Platform

The experiments were conducted in a Python 3.10 environment using Keras 2.4 and TensorFlow GPU 1.8. The hardware included a Intel Core i7 v8 computer running Windows 11 @3.10GHz CPU, 12GB RAM, and an NVIDIA Geforce GPU.

B. Training and Validation

We trained an EfficientNetB3 model to categorize patches utilizing various hyperparameters. The model underwent training for 20 epochs, with a sample size of 10 and a learning rate of 0.001. Training was ended if no progress was observed after three reductions in the learning rate. If the validation loss did not decrease for one epoch, the learning rate was halved. The data was divided into three categories: training (80%), validation (10%), and testing (10%). No cross-validation was conducted, and the Adamax optimizer was utilized with its default parameters.

C. Dataset Separation

The dataset was divided into training, validation, and test sets in an 80-10-10 ratio after file locations and associated labels were defined. We developed data generators for training, validation, and testing. TensorFlow classes took care of data augmentation and preprocessing. The training data was mixed up, but the test data was not.

5. Result Analysis and Discussion

5.1. Performance Evaluation of Models

Based on performance metrics, results are analyzed below:

A. Accuracy

Table 3. shows that EfficientNetB3 achieved the highest accuracy at 97.00%, followed by MobileNet-V2 at 95.00%, and DenseNet121 at 94.00%. ResNet50 has accuracy of 91.00%, while VGG16 had the lowest accuracy at 87.06%.

Table 3. Comparison of classification accuracy across different deep learning models. EfficientNetB3 exhibit the highest performance

Model Name	Accuracy (%)
EfficientNetB3	97.00
ResNet50	91.00
VGG16	87.06
MobileNet-V2	95.00
DenseNet121	94.00

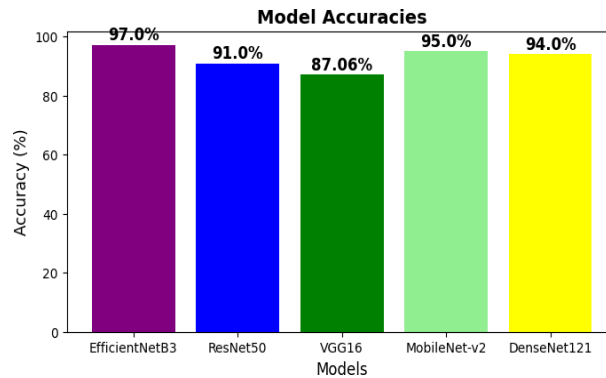


Fig.3. Comparison of classification accuracies across different models, with EfficientNetB3 achieving the highest accuracy

B. Precision

Table 4 displays precision scores for different eye conditions. DenseNet121 and EfficientNetB3 excelled in precision for cataract (97% and 96%, respectively). For diabetic retinopathy, EfficientNetB3, ResNet50, and DenseNet121 all achieved perfect scores. DenseNet121 led for glaucoma (97%), and EfficientNetB3 had the highest precision for normal conditions (94%).

Table 4. Performance comparison of models based on precision scores for cataract, diabetic retinopathy, glaucoma, and normal classes.

Model	Cataract	Diabetic Retinopathy	Glaucoma	Normal
EfficientNetB3	0.96	1.00	0.97	0.94
ResNet50	0.97	0.99	0.85	0.89
VGG16	0.89	0.95	0.86	0.86
MobileNet-V2	0.86	0.88	0.85	0.85
DenseNet121	0.93	0.94	0.93	0.87

C. Recall

Table 5 shows recall scores where EfficientNetB3 led across all classes with scores of 0.99 for cataract, 1.00 for diabetic retinopathy, 0.93 for glaucoma, and 0.95 for normal. DenseNet121 and ResNet50 also performed well, with high recall across various conditions. VGG16 had a lower recall for Glaucoma.

Table 5. Recall scores by model and class.

Model	Cataract	Diabetic Retinopathy	Glaucoma	Normal
EfficientNetB3	0.99	1.00	0.93	0.95
ResNet50	0.94	0.99	0.88	0.88
VGG16	0.95	1.00	0.75	0.89

MobileNet-V2	0.92	0.88	0.89	0.90
DenseNet121	0.88	0.93	0.85	0.94

D. F1 Score

Table 6 indicates F1 scores where EfficientNetB3 consistently performed best, especially in cataract (0.98) and diabetic retinopathy (1.00). DenseNet121 showed strong results for Normal (0.91). VGG16 and MobileNetV2 showed varied results across classes.

Table 6. F1 Scores by model and class.

Model	Cataract	Diabetic Retinopathy	Glaucoma	Normal
EfficientNetB3	0.98	1.00	0.95	0.95
ResNet50	0.95	0.99	0.86	0.88
VGG16	0.92	0.97	0.79	0.87
MobileNet-V2	0.94	0.93	0.88	0.88
DenseNet121	0.93	0.85	0.88	0.91

5.2. Analysis of Test and Validation Curves

For this experiment, the base model EfficientNetB3 is augmented with a Batch Normalization layer, a Dense layer with 256 units for regularization, and a Dropout layer with a rate of 0.45. The Dense layers use the ReLU activation function to introduce non-linearity. The dataset was split into 80% for training, 10% for validation, and 10% for testing. The Adamax optimizer, with a learning rate initialized at 0.001, was used to train the model. If the validation accuracy does not improve for a specified number of epochs (patience), a callback function reduces the learning rate by a factor of 0.5. The training batch size is 32, and the training runs for a maximum of 20 epochs. Additionally, the pretrained EfficientNetB3 layers were fine-tuned to enhance performance on the target dataset. The left plot shows that both the training loss and the validation loss decrease steadily. This indicates that the model is effectively minimizing prediction errors. The minimum validation loss occurred at epoch 19, indicating optimal generalization to new observations. The training accuracy (right plot) continued to improve, rising above 99% in the final epochs. Validation accuracy peaked at approximately epoch 15, reaching about 95%. This indicates that the model fitted the validation set optimally before slight overfitting began. Loss indicates the deviation of the model's predicted probabilities from the actual labels, while accuracy indicates the proportion of correct predictions. These metrics were calculated after each epoch for both the training and validation sets.

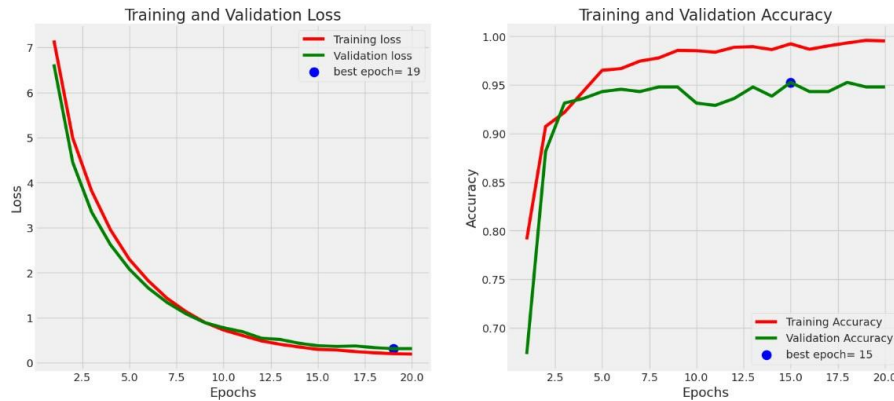


Fig. 4. Test and validation curves depicting the model's performance over training epochs, showing the accuracy and loss trends for both the training and validation datasets.

A. ROC and AUC Analysis

Fig. 5 shows the AUC scores and ROC curves for the multi-class classification of eye diseases. EfficientNetB3 demonstrated exceptional performance with an AUC score of 1.00 for all classes: cataract, diabetic retinopathy, glaucoma, and normal. This indicates excellent discrimination between classes.

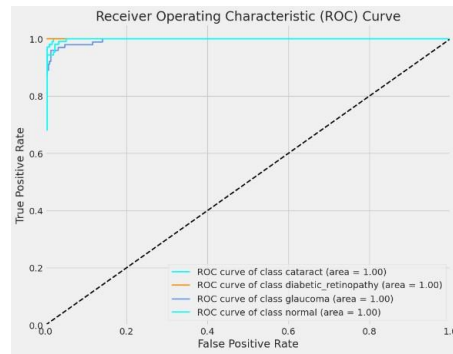


Fig. 5. ROC curve and AUC of the EfficientNetB3 model, demonstrating its excellent classification performance with an AUC value of 1.00.

B. Logarithmic Loss

Table 7 presents the log loss for each model. EfficientNetB3 achieved the lowest log loss of 0.10, indicating superior prediction accuracy and fewer errors. ResNet50 followed with a log loss of 0.21. VGG16, MobileNetV2, and DenseNet121 had higher log losses of 0.26, 0.59, and 0.56, respectively, with MobileNetV2 and DenseNet121 being less reliable compared to the top models.

Table 7. Logarithmic loss of different models.

Model Name	Logarithmic Loss EfficientNetB3
ResNet50	0.10
VGG16	0.21
DenseNet121	0.56

C. Kappa Score Analysis

Table 8 shows Cohen's Kappa scores. The maximum Kappa score of 0.95 was attained by EfficientNetB3, indicating the most favorable agreement between predicted and actual classifications. ResNet50 had a Kappa score of 0.88, demonstrating good predictive ability, although slightly lower than that of EfficientNetB3. VGG16 scored 0.86, while MobileNetV2 and DenseNet121 had the lowest Kappa Scores of 0.61, indicating poorer performance.

Table 8. Kappa scores of different models.

Model Name	Kappa Score
EfficientNetB3	0.95
ResNet50	0.88
VGG16	0.86
MobileNet-V2	0.61
DenseNet121	0.61

D. Confusion Matrix Analysis

Fig. 6. Confusion matrix of the EfficientNetB3 model evaluated on a test set comprising 422 retinal images randomly selected from the dataset. As depicted in Figure 6, the confusion matrix shows high values along the diagonal for all four classes: cataract, diabetic retinopathy, glaucoma, and normal; indicating strong classification accuracy across categories. The results demonstrate that the EfficientNetB3-based model effectively distinguishes between different retinal disease types, confirming its robustness and reliability on unseen data. The breakdown is as follows:

- Cataract: 103 correct predictions, with 1 misclassified as normal.
- Diabetic retinopathy: perfect accuracy with 110 correct predictions and no misclassifications.
- Glaucoma: 94 correct predictions, with 2 instances misclassified as cataract and 5 as normal.
- Normal: 102 correct predictions, with 2 instances misclassified as cataract and 3 as glaucoma.

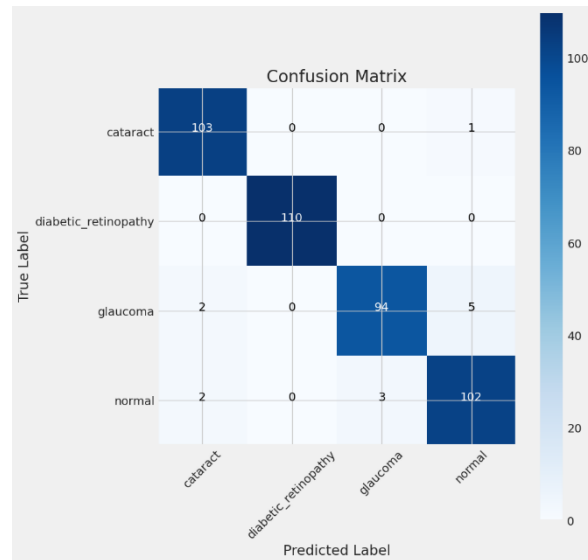


Fig.6. Confusion matrix

The darker shades on the matrix indicate a higher number of correct predictions, while lighter shades represent misclassifications. EfficientNetB3 performs exceptionally well, especially for diabetic retinopathy, but there is some room for improvement in distinguishing glaucoma from cataract and normal classes.

E. Grad-CAM Visualization

Fig. 7 depicts a Grad-CAM heat map superimposed on a retinal fundus photograph. The heat map helps visualize which image regions are most significant for the model's predictions:

- Optic disc: The bright yellow-white area represents the optic disc, a key landmark where the optic nerve exits the eye.
- Retinal blood vessels: The red and purple lines radiating from the optic disc are blood vessels supplying the retina.
- Central macula: The hyperpigmented central macula, appearing more red and yellow, may indicate increased metabolic activity or a pathological process.

The Grad-CAM method was used to interpret the model's predictions. For Glaucoma, the heatmaps highlighted the optic disc, which is a key feature in identifying glaucoma. For Diabetic Retinopathy, the model focused on areas with microaneurysms, which are characteristic signs of the disease. The Grad-CAM visualizations did not highlight irrelevant areas such as the background or eyelashes, ensuring that the model's attention was directed towards clinically significant regions.

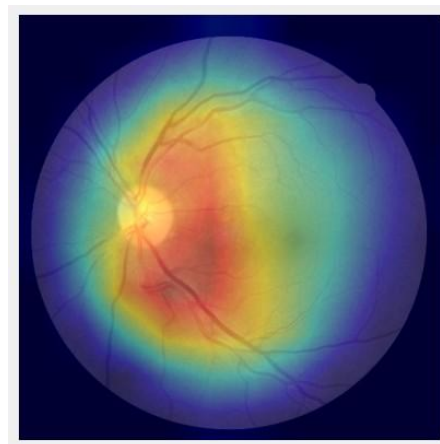


Fig.7. Grad-CAM visualization showing the regions of interest in the eye images that contributed most to the EfficientNetB3 model's predictions

F. Web Interface Visualization

The visualization system was implemented using Flask on the backend. The entire visualization system was developed using VS Code. A local website, accessible via an IP address, provides an interface for eye disease

classification. Upon accessing the IP address, the website opens. Users can upload images for classification in the designated area. After uploading an image, the website displays the classification result. These functions are powered by the EfficientNetB3 model, which performs the image classification.

```
INFO:werkzeug:WARNING: This is a development server.
* Running on http://127.0.0.1:5000
INFO:werkzeug:Press CTRL+C to quit
INFO:werkzeug: * Restarting with stat
```

Fig.8. Local website's IP address

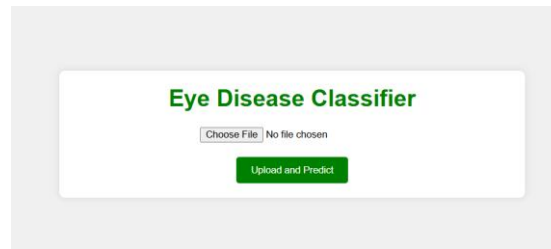


Fig.9. Web page design showcasing the user interface for the eye disease detection system

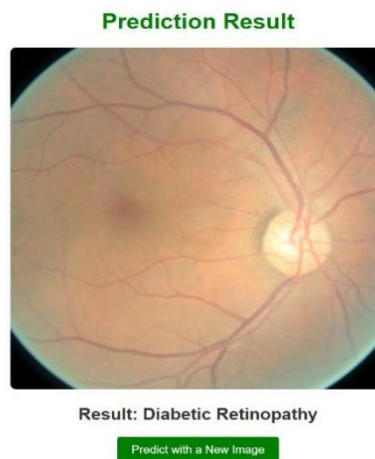


Fig.10. Prediction results of the EfficientNetB3 model, showcasing the model's classification outcomes for different eye diseases based on the input medical images

5.3. Discussion

The EfficientNetB3-based model outperformed other DL models in classifying retinal diseases because it employs a compound scaling approach that achieves an optimal compromise among network depth, width, and input resolution. We downsampled images to 224×224 to match the input size of the pretrained model. We also used ImageNet weights to reduce convergence time and enhance feature learning. The model performed well and was stable with a learning rate of 0.001, the Adamax optimizer, and training for 20 epochs. Regularization ($L1=0.006$, $L2=0.016$), dropout (0.45), and batch normalization effectively reduced overfitting and enhanced training robustness. Although the model performed well, it remains sensitive to noise and variations in images. Grad-CAM interpretability of it yet has to be proven to make sure that it has clinical utility. In times to come, various big-data clinical databases from other hospitals will be utilized to challenge the model to function under varying demographics, modalities of imaging, and settings. It is also important to see how well performs with poor or anomalous pictures of eyes. Also, engaging ophthalmologists and other doctors in user testing will allow us to ascertain how informative and descriptive Grad-CAM-based explanations are. Individual input can make it simpler for model to understand over time in making it simpler to use in clinic.

6. Conclusions and Future Works

Our work focused on employing explainable AI and machine learning in diagnosing eye disorders. The highest-performing model was the EfficientNetB3 model. With an overall accuracy of 97.003%, EfficientNetB3 significantly outperformed other models in precisely classifying four separate eye disorders. In each class, the model reliably possessed excellent F1 scores, precision, and recall. F1 values, for example, ranged from 0.95 to 1.00, precision levels ranged from 0.94 to 1.00, and recall levels ranged from 0.93 to 1.00 (Tables 4, 5, and 6). The model is better at accurately identifying and discriminating between cataract, diabetic retinopathy, glaucoma, and normal situations than ResNet50,

VGG16, MobileNetV2, and DenseNet121. We better comprehended the decision-making process of the model through the application of explainable AI methods such as Grad-CAM, an indispensable tool in medicine. Knowing how a model came to its conclusion is just as important in clinical practice as knowing what that conclusion is. On the whole, results show that ophthalmic patients can make better and more confident choices through employing the best of machine learning models with explainable AI that do not only give extremely correct diagnosis but in addition make it understandable how predictions are arrived at.

Furthermore, the proposed system shows strong potential for real-world deployment in large-scale screening programs and telemedicine applications, particularly in remote and resource-limited regions where access to ophthalmologists is restricted. By providing automated disease classification along with visual explanations, the system can function as an effective clinical decision-support tool, assisting ophthalmologists in early diagnosis, prioritization, and remote patient monitoring. Overall, this approach contributes toward improving accessibility, efficiency, and reliability in ophthalmic healthcare delivery.

In the future, we will work on combining 3D retinal analysis with OCT data, making the software more accessible throughout the world by adding support for several languages and localised interfaces, and balancing datasets with synthetic data created by GANs. Time-series analysis can be used to keep track of how a disease is getting worse so that treatment can start early. A cloud-based, scalable deployment architecture will cut down on hardware expenses, while integrating EHRs will add patient data to the system. Finally, making sure that bias and fairness are present in all groups will be very important.

All the Declarations and Statements

Author Contributions Statement

Sadikur Rahman Sadik – Data Curation, handled data acquisition, dataset preprocessing, and implemented the research model. Model Training, Validation, and Performance Evaluation: Led the model training process, validated results using standard metrics, and benchmarked performance against existing methods.

Md. Abdul Halim Khan – Visualization, Conceptualization, Methodology, Writing – Drafted the initial manuscript, contributed to the literature survey, and documented the technical background of the study.

Samsuddin Ahmed – Supervision: Proposed research ideas, constructed the overall framework, and supervised project execution. Writing – Review and Editing, and Project Management: Reviewed and edited the manuscript, ensuring clarity and coherence and helped coordinate project milestones and deadlines.

Atiqur Rahman – Formal Analysis, Visualization, and Statistical Analysis: Performed in-depth analysis of experimental results, prepared performance charts, ensured the statistical robustness of the evaluation and helped coordinate project milestones and deadlines.

Sadia Enam – Formal Analysis, Visualization, and Statistical Analysis: Performed in-depth analysis of experimental results, prepared performance charts, ensured the statistical robustness of the evaluation and helped coordinate project milestones and deadlines.

Md. Toukir Ahmed – Supervision: Proposed research ideas, constructed the overall framework, and supervised project execution. Writing – Review and Editing, and Project Management: Reviewed and edited the manuscript, ensuring clarity and coherence, and helped coordinate project milestones and deadlines.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement

The authors declare no conflicts of interest.

Funding Declaration

We received no external funding for this research.

Data Availability Statement

All data utilized in this study were obtained from publicly available sources. No new datasets were generated or analyzed.

Ethical Declarations

The authors should declare the following: Statements of ethical approval for studies involving human subjects and/or animals.

Acknowledgments

First of all, we want to thank Allah for giving us the power, understanding, and will to start and finish this research. We are very grateful to our families for the help and encouragement they gave us when we were doing our research. We want to

thank our valued supervisors for all the real help, motivation, and constructive criticism they gave us. This research would not have been accomplished without their great supervision and help that can't be replaced. We are very grateful to them for letting us collaborate with them. We also want to thank the University of Frontier Technology, Bangladesh, for giving us the resources we needed and making this study space possible.

Declaration of Generative AI in Scholarly Writing

During the preparation of this work, the authors used different AI tools to improve the language and readability. After using these tools/services, the authors reviewed and edited the content as needed and took full responsibility for the content of the publication.

Abbreviations

The following abbreviations are used in this manuscript:

ML – Machine Learning
 AI – Artificial Intelligence
 XAI – Explainable Artificial Intelligence
 CNN – Convolutional Neural Network
 Grad-CAM – Gradient-weighted Class Activation Mapping
 TP – True Positive
 FP – False Positive
 TN – True Negative
 FN – False Negative
 ROC – Receiver Operating Characteristic
 AUC – Area Under Curve
 CAD – Computer -Aided Design
 SHAP – Shapely Additive explanations
 LIME – Local Interpretable Model-agnostic Explanations
 MBConv – Mobile Inverted Bottleneck Convolution
 DL – Deep Learning
 ReLU – Rectified Linear Unit

References

- [1] A. Mishra, N. Choubey, P. Jumde, A. Samiya, A. Garg, and A. Kale, "Early eye disease detection using machine learning," in *2025 IEEE 14th International Conference on Communication Systems and Network Technologies (CSNT)*. IEEE, 2025, pp. 134–139.
- [2] A. Sharma and S. Gupta, "AI-driven eye disease detection: A VGG19-based approach to retinal image classification," in *2025 6th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI)*, 2025, pp. 1325–1329.
- [3] A. A. A. El-Khalek, H. M. Balaha, A. Mahmoud, N. S. Alghamdi, M. Ghazal, A. T. Khalil, M. E. A. Abo-Elsoud, and A. El-Baz, "A novel machine learning-based classification framework for age-related macular degeneration (AMD) diagnosis from fundus images," in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, 2024, pp. 1–4.
- [4] N. Badah, A. Algefes, A. AlArjani, and R. Mokni, "Automatic eye disease detection using machine learning and deep learning models," in *Pervasive Computing and Social Networking: Proceedings of ICPCSN 2022*. Springer, 2022, pp. 773–787.
- [5] L. Guo, J.-J. Yang, L. Peng, J. Li, and Q. Liang, "A computer-aided healthcare system for cataract classification and grading based on fundus image analysis," *Computers in Industry*, vol. 69, pp. 72–80, 2015.
- [6] Y. N. Fuadah, A. W. Setiawan, and T. Mengko, "Performing high accuracy of the system for cataract detection using statistical texture analysis and k-nearest neighbor," in *2015 International Seminar on Intelligent Technology and Its Applications (ISITIA)*. IEEE, 2015, pp. 85–88.
- [7] J.-J. Yang, J. Li, R. Shen, Y. Zeng, J. He, J. Bi, Y. Li, Q. Zhang, L. Peng, and Q. Wang, "Exploiting ensemble learning for automatic cataract detection and grading," *Computer Methods and Programs in Biomedicine*, vol. 124, pp. 45–57, 2016.
- [8] M. Caixinha, J. Amaro, M. Santos, F. Perdigão, M. Gomes, and J. Santos, "In-vivo automatic nuclear cataract detection and classification in an animal model by ultrasounds," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 11, pp. 2326–2335, 2016.
- [9] A. Jagadale, S. Sonavane, and D. Jadav, "Computer aided system for early detection of nuclear cataract using circle hough transform," in *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*. IEEE, 2019, pp. 1009–1012.
- [10] L. Cao, H. Li, Y. Zhang, L. Zhang, and L. Xu, "Hierarchical method for cataract grading based on retinal images using improved haar wavelet," *Information Fusion*, vol. 53, pp. 196–208, 2020.
- [11] M. S. Junayed, M. B. Islam, A. Sadeghzadeh, and S. Rahman, "Cataractnet: An automated cataract detection system using deep learning for fundus images," *IEEE Access*, vol. 9, pp. 128799–128808, 2021.
- [12] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-CAM: Why did you say that?" arXiv preprint arXiv:1611.07450, 2016.
- [13] H. Zhang and C.-J. Wang, "Understanding regularization in deep learning," *arXiv preprint arXiv:2106.13259*, 2021.

- [14] T. Pratap and P. Kokil, "Computer-aided diagnosis of cataract using deep transfer learning," *Biomedical Signal Processing and Control*, vol. 53, p. 101533, 2019.
- [15] G. V. Doddi, "Eye diseases classification," <https://www.kaggle.com/datasets/gunavenkatdoddi/eye-diseases-classification>, 2022, accessed: 2025-06-21.
- [16] N. Björck, C. P. Gomes, B. Selman, and K. Q. Weinberger, "Understanding batch normalization," *Advances in Neural Information Processing Systems*, vol. 31, 2018.
- [17] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*. IGI Global, 2010, pp. 242–264.
- [18] T. Kotsilieris, I. Anagnostopoulos, and I. E. Livieris, "Regularization techniques for machine learning and their applications," p. 521, 2022.
- [19] P. Baldi and P. Sadowski, "The dropout learning algorithm," *Artificial Intelligence*, vol. 210, pp. 78–122, 2014.
- [20] I. Markoulidakis and G. Markoulidakis, "Probabilistic confusion matrix: A novel method for machine learning algorithm generalized performance analysis," *Technologies*, vol. 12, no. 7, p. 113, 2024.
- [21] A. Bassel, A. Abdulkareem, Z. Alyasseri, N. Sani, and H. Mohammed, "Automatic malignant and benign skin cancer classification using a hybrid deep learning approach. diagnostics 2022, 12, 2472," 2022.
- [22] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006, pp. 233–240.
- [23] R. Diallo, C. Edalo, and O. O. Awe, "Machine learning evaluation of imbalanced health data: A comparative analysis of balanced accuracy, mcc, and f1 score," in *Practical Statistical Learning and Data Science Methods: Case Studies from LISA 2020 Global Network, USA*. Springer, 2024, pp. 283–312.
- [24] B. Sasikala, V. G. Biju, and C. Prashanth, "Kappa and accuracy evaluations of machine learning classifiers," in *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*. IEEE, 2017, pp. 20–23.
- [25] Z. H. Hoo, J. Candlish, and D. Teare, "What is an ROC curve?" pp. 357–359, 2017.
- [26] V. Vovk, "The fundamental nature of the log loss function," *Fields of Logic and Computation II: Essays Dedicated to Yuri Gurevich on the Occasion of His 75th Birthday*, pp. 307–318, 2015.
- [27] O. O. Bifarín and F. M. Fernández, "Automated machine learning and explainable ai (automl-xai) for metabolomics: improving cancer diagnostics," *Journal of the American Society for Mass Spectrometry*, vol. 35, no. 6, pp. 1089–1100, 2024.

Authors' Profiles



Sadikur Rahman Sadik completed his B.Sc. in the Department of IoT and Robotics Engineering at the University of Frontier Technology, Bangladesh. His secondary school certificate (SSC) was obtained from Shamsul Hoque Khan School and College, and his higher secondary certificate (HSC) was obtained from Notre Dame College, Dhaka.



Md. Abdul Halim Khan completed his B.Sc. in the Department of IoT and Robotics Engineering at the University of Frontier Technology, Bangladesh. He is now working as a lecturer in Northern University of Business and Technology Khulna. His research interests include Internet of Things (IoT), Machine Learning, Explainable AI, Robotics, and Embedded Systems.



Samsuddin Ahmed received the B.Sc. degree in Computer Science and Engineering from the University of Chittagong in 2010, and the M.S. degree from the Department of Computer Engineering, Chosun University, South Korea in 2020, under the supervision of Prof. Ho Yub Jung. He is currently an Assistant Professor with the Department of IoT and Robotics Engineering, University of Frontier Technology, Bangladesh. His research interests include Video Coding, Machine Vision, deep learning, data analysis, explainable AI, and the IoT.



Atiqur Rahman graduated from the University of Frontier Technology, Bangladesh, with a B.Sc. degree in IoT and Robotics Engineering. His research interests include Internet of Things (IoT), robotics, artificial intelligence (AI), machine learning (ML), and deep learning, with a particular focus on developing intelligent systems that integrate hardware design, embedded systems, and AI-driven decision-making. He has undertaken various IoT-based projects, demonstrating proficiency in the design and implementation of smart systems. He is currently pursuing an M.Sc. (Engg.) degree in IoT and Robotics Engineering from the same university. He has recently joined Daffodil International University (DIU) as a Lecturer, where he continues to pursue his academic and research interests in emerging technologies.



Sadia Enam obtained her M.Sc. in Electrical and Electronic Engineering (EEE) from Rajshahi University of Engineering & Technology (RUET) in 2022, and her B.Sc. in EEE from RUET in 2016. She worked as a lecturer in the Department of EEE at North Bengal International University (NBIU) from October 14, 2016 to November 25, 2022, and at Bangladesh Army University of Engineering & Technology from December 1, 2022 to November 30, 2023. Since January 15, 2023, she has been a lecturer in the Department of Internet of Things (IoT) and Robotics Engineering (IRE) at the University of Frontier Technology, Bangladesh.



Md. Toukir Ahmed is an Assistant Professor in the Department of Internet of Things and Robotics Engineering (IRE) at the University of Frontier Technology, Bangladesh (UFTB). He is currently pursuing his Ph.D. in Computer Science and Engineering (CSE) at Rajshahi University of Engineering & Technology (RUET). He previously earned both his M.Sc. Engineering and B.Sc. Engineering degrees in CSE from the same university. Before joining UFTB, he worked as a full-time faculty member in the ICT and CSE Departments at the Bangladesh Army University of Science and Technology (BAUST), Saidpur, and also served in the CSE Department at Bangladesh University, Dhaka. His research interests span Traditional Machine Learning, Deep Learning and

Internet of Things. He has authored several journal and conference publications and is passionate about advancing intelligent systems and emerging technologies.

How to cite this paper

Sadikur Rahman Sadik, Md. Abdul Halim Khan, Samsuddin Ahmed, Atiqur Rahman, Sadia Enam, Md. Toukir Ahmed, "An EfficientNetB3 Approach for Retinal Disease Classification with XAI and Online Interface", International Journal of Intelligent Systems and Applications(IJISA), Vol.18, No.4, pp.41-56, 2026. DOI:10.5815/ijisa.2026.04.03