IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Mathematical Modelling of Engineering Problems
Vol. 12, No. 12, December, 2025, pp. 4426-4434
Journal homepage: http://iieta.org/journals/mmep

Enhancing the Performance of Multi-Class Classification Systems for Retinal Diseases Using Deep Learning Models

Noor Mowafeq Allaya*, Ula Tarik Salim

Department of Computer Engineering, University of Mosul, Mosul 41002, Iraq

Corresponding Author Email: noor.mowafeq@uomosul.edu.iq

https://doi.org/10.18280/mmep.121231

Received: 4 November 2025
Revised: 9 December 2025
Accepted: 16 December 2025
Available online: 31 December 2025

Keywords:
retinal disease classification, deep learning, ensemble learning, medical image analysis, explainable AI, Grad-CAM

ABSTRACT

Early and accurate detection of retinal diseases is essential to prevent avoidable vision loss. However, manual assessment of fundus images is time-consuming and can vary across clinicians. Deep convolutional neural networks (DCNNs) have improved automated screening, but many models remain computationally demanding and provide limited interpretability. This study proposes a hybrid ensemble framework for multi-class retinal disease classification that balances accuracy, efficiency, and explainability. InceptionV3 and DenseNet121 were used as feature extractors on a public Eye Diseases Classification dataset comprising four categories: normal, cataract, glaucoma, and diabetic retinopathy. The extracted deep features were fused and classified using several ensemble strategies, including hard voting, soft voting, stacking, bagging with Random Forest, and gradient boosting. Performance was evaluated using accuracy, precision, recall, and F1-score, together with training time. DenseNet121 achieved higher accuracy than InceptionV3 while requiring shorter training time. Ensemble learning further improved performance. Bagging with Random Forest reached 99.4% accuracy, and the optimized boosting model achieved 100% accuracy on the held-out test set. Model interpretability was examined using Grad-CAM, which highlighted clinically plausible regions such as the optic disc, macula, and lesion areas. Although the results are promising, external validation on additional datasets is required before clinical deployment.

1. INTRODUCTION

Retinal diseases are a major cause of visual impairment and blindness worldwide, affecting individuals across all age groups. Visual signals are captured by the retina and transmitted to the brain through the optic nerve [1, 2]. Several conditions, including cataract, glaucoma, diabetic retinopathy, and age-related macular degeneration, can lead to severe and irreversible vision loss if diagnosis or treatment is delayed [3]. The World Health Organization estimates that 2.2 billion people experience vision impairment, and a substantial proportion of cases could be prevented or treated through timely diagnosis and intervention [4]. Cataract and diabetic retinopathy account for a large number of preventable cases, and glaucoma remains a major cause of permanent vision loss [5].

Manual screening of retinal fundus images is labor-intensive and depends on expert interpretation, which may lead to variability and delayed clinical decisions [6, 7]. Recent advances in artificial intelligence have enabled automated analysis of ophthalmic images using deep learning, particularly deep convolutional neural networks (DCNNs) [8-11]. Despite their success, conventional end-to-end deep models often require substantial computational resources and may generalize poorly across datasets, especially when image acquisition conditions differ. In addition, many DCNNs provide limited transparency, which can reduce clinical trust [12].

Ensemble learning is a well-established strategy to improve robustness and generalization by combining multiple models. Methods such as hard voting, soft voting, stacking, bagging, and boosting can reduce variance and mitigate the limitations of individual classifiers [13-15]. Nevertheless, many existing retinal classification studies focus primarily on maximizing accuracy, while offering limited analysis of computational cost and limited interpretability beyond qualitative examples [16].

To address these limitations, this work makes the following contributions:

(1) A hybrid framework is proposed that combines two complementary CNN backbones (InceptionV3 and DenseNet121) with multiple ensemble learning strategies (hard voting, soft voting, stacking, bagging, and gradient boosting) for four-class retinal disease classification.

(2) Single-CNN and ensemble configurations are compared using predictive metrics and training time to assess practical deployability.

(3) Grad-CAM is used to generate visual explanations and to examine whether the learned attention patterns align with clinically meaningful retinal structures.

The overall objective is to deliver a screening system that is accurate, computationally efficient, and interpretable.

2. RELATED WORK

Several studies have investigated deep learning for automated retinal disease classification. In a previous study [17], a CNN-based system was developed for multi-class diagnosis of diabetic eye disease using fundus images and achieved 81.33% accuracy. The authors emphasized that while deep learning performs well in binary settings, multi-class discrimination remains challenging.

In a previous study [18], pre-trained CNNs and optimization strategies were evaluated for mild and multi-class diabetic eye disease classification. VGG16 achieved 88.3% accuracy for multi-class classification and 85.95% accuracy for mild multi-class classification. VisionDeep-AI [19] combined vessel segmentation and disease classification using a bi-directional feature pyramid network and a U-Net-based architecture. The reported vessel segmentation accuracy reached 97.73%, while classification accuracy was 81.50%.

Other works focused on enhancing preprocessing and feature extraction. Image enhancement and segmentation were used prior to classification, and EfficientNetB7 produced strong results [20]. A custom DCNN achieved high detection rates across cataract, diabetic retinopathy, glaucoma, and normal cases. Nawaz et al. [21] proposed an optimized CNN for large-scale retinal disease classification and reported 95% accuracy on the Eye-Net dataset. A hybrid pipeline used segmentation, SqueezeNet feature extraction, and a stacked sparse autoencoder classifier, achieving 96.3% accuracy [22].

Comparative studies have also been reported. In a previous study [23], multiple CNN architectures were evaluated on the MURED dataset, and the best reported accuracy remained limited, highlighting the dependence on dataset scale and quality. Ensemble-based approaches have shown improved robustness. Multiple CNNs were trained on RFMiD and combined using ensemble strategies, achieving an AUROC of 0.9613 for screening and 0.9295 average AUROC for condition classification [24]. In a previous study [25], ResNet50 and DenseNet121 were applied to multi-class identification of cataract, glaucoma, and diabetic retinopathy. In the study by Khan et al. [26], an ensemble of EfficientNet variants achieved an AUC of 0.973 on RFMiD, with emphasis on efficiency and scalability.

Overall, prior work confirms the effectiveness of DCNNs and ensembles for retinal analysis. However, several limitations persist. Many studies address binary or narrowly defined multi-class problems, which restricts real-world applicability. Some ensemble methods are computationally intensive and provide limited discussion of training cost. Moreover, interpretability is often treated superficially, and few studies analyze whether model attention consistently overlaps with clinically relevant anatomical structures.

This study differs by targeting four common retinal categories in a single unified framework, combining deep feature extraction with classical ensemble learners to balance accuracy and efficiency, and integrating Grad-CAM to examine the clinical plausibility of model attention.

Table 1. Related work analysis

References	Focus Area	Dataset	Model Used	Accuracy (%)	Key Contributions
[17]	CNN-based multi-class classification of DED	Publicly available DED dataset	CNN model	81.33	CNN-based classification of multiple DED categories
[18]	Deep learning-based classification of mild and multi-class DED	Retinal fundus images	VGG16, CNN	88.3	VGG16 achieved high accuracy for mild and multi-class DED
[19]	VisionDeep-AI for vessel segmentation & classification	Large fundus image dataset	Bi-directional feature pyramid, U-Net	97.73	VisionDeep-AI enhances segmentation and classification performance
[20]	Deep learning-based enhancement, segmentation, and classification of DED	DED fundus images	ResNet50, VGG16, Xception, EfficientNetB7	98.33	Custom DCNN model with advanced segmentation techniques
[21]	CNN-based multi-class classification optimizing feature extraction	Eye-Net dataset	Optimized CNN model	95	Memory-efficient CNN model for large-scale classification
[22]	IDL-MRDD: Multi-label classification of retinal diseases	Benchmark multi-retinal disease dataset	SqueezeNet, SSAE	96.3	Hybrid deep learning model for multi-label classification
[23]	Comparison of CNN architectures for retinal disease classification	MURED dataset	Scratch Model, GoogleNet, VGG, ResNet, MobileNet, DenseNet	49.85	Comparison of multiple CNN architectures for disease classification
[24]	Ensemble CNN models for multi-disease detection	RFMiD dataset	Ensemble CNNs	92.95	Ensemble CNNs improve retinal disease classification in RFMiD
[25]	MobileNetV2 for multi-class classification of retinal diseases	Color fundus images	MobileNetV2, ResNet50, DenseNet121	94.23	MobileNetV2 optimized for real-time disease detection
[26]	Multi Retinal Disease Classification Model (MRDCM)	RFMiD dataset (45 disease classes)	EfficientNetB4, EfficientNetV2S	97.3	MRDCM surpasses previous models with ensemble learning

Table 1 summarizes representative studies, highlighting datasets, model choices, reported accuracy, and main contributions.

The reviewed literature indicates that high performance is achievable; however, generalization across datasets, computational efficiency, and explainability remain open challenges.

3. PROPOSED METHODOLOGY

Figure 1 illustrates the overall workflow of the proposed approach. The pipeline includes dataset preparation, image preprocessing, deep feature extraction using two CNN backbones, feature fusion, ensemble classification, and interpretability analysis using Grad-CAM.

3.1 Dataset overview

The Eye Diseases Classification dataset contains 4,217 color fundus images distributed across four classes: 1,074 normal, 1,038 cataract, 1,007 glaucoma, and 1,098 diabetic retinopathy [27]. The class distribution is approximately balanced. Therefore, no over-sampling, under-sampling, or class-weighting strategy was applied. Instead, a stratified split was used to preserve class proportions.

The dataset was partitioned into 80% training data (3,373 images) and 20% held-out test data (844 images) using stratification at the image level. During training, 10% of the training set was further reserved for validation to support early stopping and hyperparameter selection. All reported results correspond to the held-out test set, which was not used during training or model selection (Table 2).

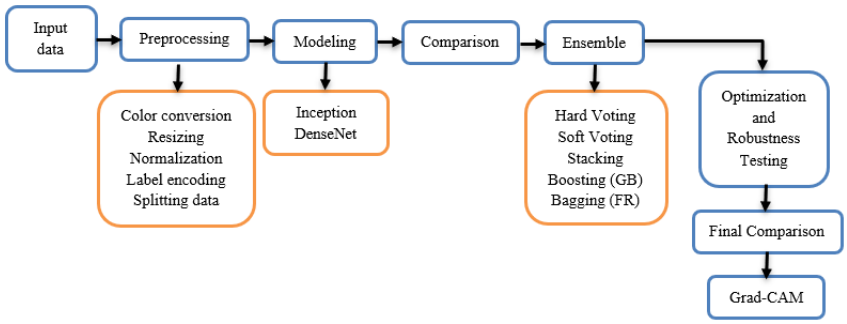

Figure 1. Proposed scheme

Table 2. Dataset overview

Category	Number of Images	Description
Normal	1,074	Retinal images from healthy individuals, used as a control group.
Cataract	1,038	Images depicting cataract-affected eyes, characterized by lens clouding.
Glaucoma	1,007	Retinal images showing signs of glaucoma, a disease that damages the optic nerve.
Diabetic Retinopathy	1,098	Images illustrating diabetic retinopathy, a complication of diabetes affecting blood vessels in the retina.

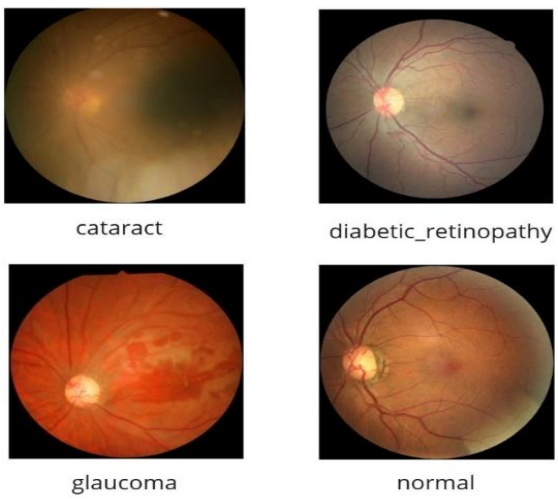

Figure 2. Retinal images from the dataset

Representative examples from each class are shown in Figure 2. Normal images present typical retinal anatomy without visible pathology. Cataract images often exhibit reduced contrast due to lens opacity. Glaucoma images are characterized by changes around the optic nerve head, including increased cupping. Diabetic retinopathy images may contain microaneurysms, hemorrhages, and exudates.

3.2 Preprocessing

All images were resized to 224 × 224 pixels to ensure consistent input dimensions. Pixel intensities were normalized to the [0,1] range to stabilize training. The images were kept in RGB format because color information is clinically relevant in fundus imaging.

To reduce overfitting and improve robustness, data augmentation was applied during training. The augmentation included random horizontal flips, small rotations (±10°), random zooming (up to 10%), and mild brightness and contrast variations. After preprocessing, the training set contained 3,373 images and the held-out test set contained 844 images [28].

Deep learning-based modeling: In this study, two ImageNet-pretrained CNN architectures were used as feature extractors: InceptionV3 and DenseNet121. For each backbone, the original classification head was removed and replaced by a GAP layer, followed by a fully connected layer with 256 units (ReLU) and dropout (0.5). A final softmax layer with four outputs was used during CNN training to support

supervised feature learning. Deep features were extracted from the penultimate layer and later used to train ensemble classifiers.

- **Inception**

InceptionV3 employs parallel convolutional operations with multiple kernel sizes within Inception modules, allowing multi-scale feature extraction. This design improves representational capacity while controlling computational cost through factorized convolutions and 1 × 1 bottleneck layers. Figure 3 shows the InceptionV3 architecture used in this work [29, 30].

- **DenseNet**

DenseNet121 introduces dense connectivity, where each layer receives feature maps from all preceding layers and passes its own feature maps to all subsequent layers. This design strengthens gradient flow, encourages feature reuse, and reduces redundant computation. DenseNet121 is therefore well suited for medical image classification tasks that require fine-grained feature representation. Figure 4 shows the DenseNet121 architecture [31].

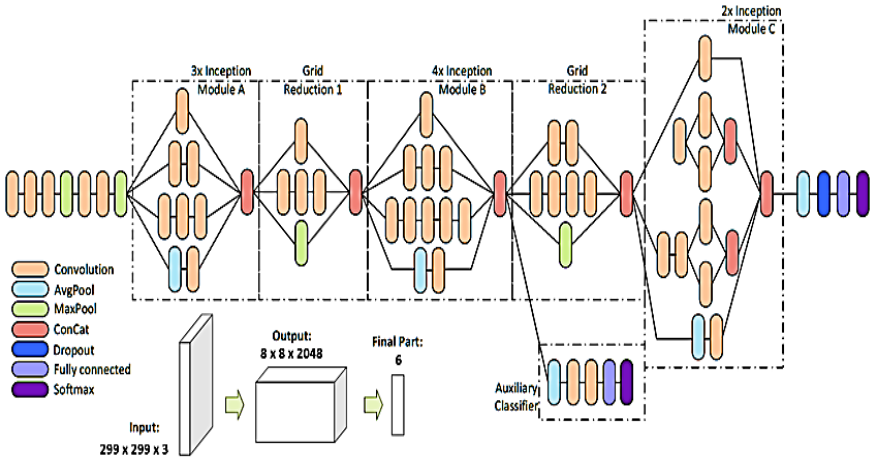

Figure 3. Inception model architecture [29]

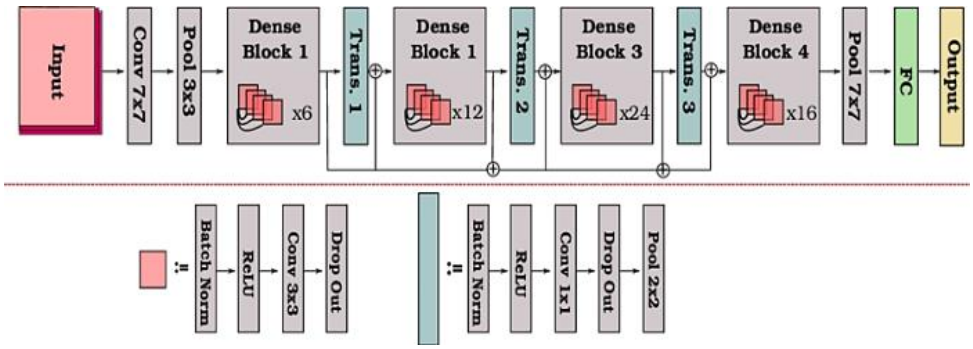

Figure 4. DenseNet121 model architecture [32]

The DenseNet121 is a DCNN that improves feature information sharing and gradient path with deep network connectivity. DenseNet layers connect to every one of the preceding layers, accepting their feature maps for processing until they propagate the results to all subsequent layers, which reduces processing redundancy [32]. The model design utilizes four dense blocks with a layer and includes transition layers containing 1 × 1 convolutions together with 2 × 2 average pooling elements. The first step combines a 7 × 7 convolution filter with a 3 × 3 max pooling operation to handle the input before the dense blocks receive it.

The distribution of 3 × 3 convolutional layers with BatchNorm and ReLU activation functions enables each dense block to improve stability and convergence. The model includes dropout layers, which help prevent overfitting, while global average pooling (GAP) reduces the feature dimensions before its output gets sent to a fully connected (FC) layer with a softmax classifier for prediction purposes. The results showed DenseNet121 as an efficient and better feature propagation, so it's an optimal choice for medical image classification, giving high accuracy and computational efficiency performance.

3.3 Ensemble learning

To improve classification robustness and generalization, ensemble learning methods were applied to the fused deep features extracted from InceptionV3 and DenseNet121. Ensemble learning combines multiple predictors to reduce variance and improve stability, which is particularly beneficial in medical imaging tasks. The following ensemble strategies were evaluated.

- **Hard Voting**

Hard voting assigns the final class label based on the majority vote across base classifiers [33]. Each classifier produces a discrete class prediction, and the class with the highest vote count is selected.

- **Soft Voting**

Soft voting aggregates predicted class probabilities instead of discrete labels. The probability distributions produced by base models are averaged, and the class with the highest mean probability is selected [34, 35]. Soft voting typically improves

performance when probability estimates are well calibrated.

- **Stacking**

Stacking trains multiple base models and uses their outputs as inputs to a meta-learner. The meta-learner learns how to optimally combine base predictions, often improving generalization compared with simple voting [36, 37].

- **Bagging (Random Forest)**

Bagging trains multiple models on bootstrap samples of the training data and aggregates their predictions by majority vote [38]. Random Forest is a commonly used bagging model based on decision trees. It reduces variance, is robust to noise, and can model complex decision boundaries.

3.4 Interpretability using Grad-CAM

Grad-CAM was applied to the last convolutional layers of the trained CNNs to generate class-specific heatmaps. For glaucoma, activations were concentrated around the optic nerve head and cup-to-disc region, which are clinically relevant for assessing glaucomatous damage. For diabetic retinopathy, the highlighted regions often corresponded to lesion patterns near the posterior pole, including areas consistent with microaneurysms and exudates. In normal images, activation tended to focus on anatomical landmarks such as the optic disc and major vessels without emphasizing irrelevant background regions.

For cataract images, attention patterns were more diffuse. This behavior is expected because cataract primarily affects the lens and can reduce global image contrast rather than producing localized retinal lesions. Overall, Grad-CAM results suggested that the models relied on clinically plausible cues. However, this analysis remained qualitative, and future work should include expert evaluation to quantify agreement between explanations and clinical criteria [39].

3.5 Training configuration and hyperparameters

InceptionV3 and DenseNet121 were initialized with ImageNet pretrained weights. Input images were resized to 224 × 224 pixels and normalized to [0,1]. Data augmentation was applied online during training.

Both CNNs were trained using the Adam optimizer with an initial learning rate of 1×10^{-4}, a batch size of 32, and categorical cross-entropy loss. Training was performed for up to 50 epochs. Early stopping monitored validation loss with a patience of 7 epochs, and the learning rate was reduced on plateau (factor 0.5, patience 3).

Ensemble models (Random Forest bagging and gradient boosting) were trained using deep features extracted from the penultimate layer of the CNNs. Hyperparameters such as the number of trees, maximum depth, and learning rate were tuned using grid search on the training and validation sets. All experiments were implemented in Python using TensorFlow/Keras and scikit-learn and were executed on a GPU-enabled workstation.

4. EVALUATION METRICS

Model performance was assessed using accuracy, precision, recall, and F1-score. These metrics were computed on the held-out test set.

4.1 Accuracy

Accuracy measures the proportion of correctly classified samples among all samples [40].

4.2 Precision

Precision measures the reliability of positive predictions and is defined as the ratio of true positives to the total number of predicted positives [41].

4.3 Recall

Recall measures the ability to identify actual positive samples and is defined as the ratio of true positives to the total number of actual positives [42].

4.4 F1-score

The F1-score is the harmonic mean of precision and recall and provides a balanced measure when class-wise performance is important [43].

5. EXPERIMENTAL RESULTS

5.1 Inception results

Table 3 reports class-wise precision, recall, and F1-score for InceptionV3. The model achieved an overall accuracy of 0.85 on the test set. Performance was highest for cataract and diabetic retinopathy, while glaucoma exhibited lower recall, suggesting that some glaucoma cases were misclassified.

Table 3. Classification of the inception model

Class	Precision	Recall	F1-score
Cataract	0.89	0.94	0.91
Diabetic Retinopathy	0.89	0.93	0.91
Glaucoma	0.83	0.75	0.79
Normal	0.79	0.79	0.79
Accuracy		0.85	
Macro Average	0.85	0.85	0.85

Figure 5 shows the learning curves for training and validation. Training accuracy increased steadily and approached saturation. Validation accuracy stabilized after the initial epochs, indicating reasonable generalization. Training and validation loss decreased overall, although minor fluctuations were observed in validation loss, which may indicate residual sensitivity to sample variability.

Figure 6 presents the ROC curves. The AUC values were high across classes, reaching 0.99 for cataract and diabetic retinopathy and 0.96 for glaucoma and normal. These results indicate strong separability, although classification errors remained more frequent for glaucoma.

5.2 DenseNet results

Table 4 summarizes performance for DenseNet121. The model achieved 0.91 test accuracy. Diabetic retinopathy yielded the highest performance, with an F1-score of 0.99. Cataract also achieved a strong performance, while glaucoma remained the most challenging class due to lower recall (0.81).

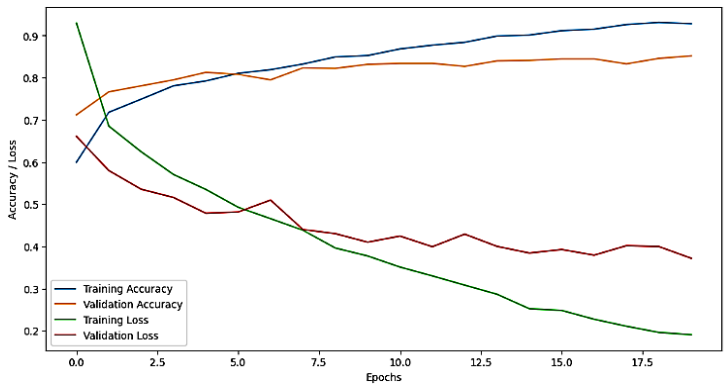

Figure 5. Learning curve of InceptionV3

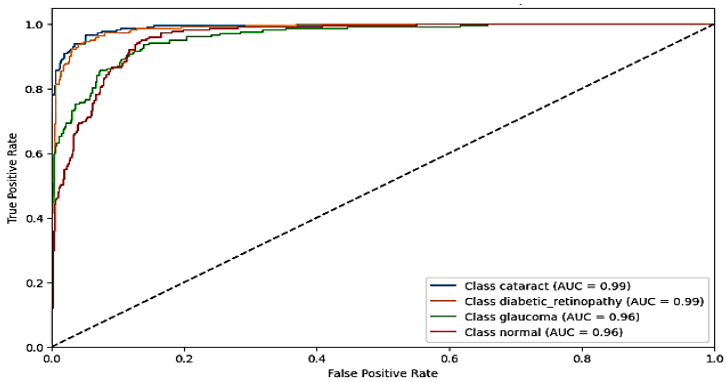

Figure 6. ROC curve of InceptionV3

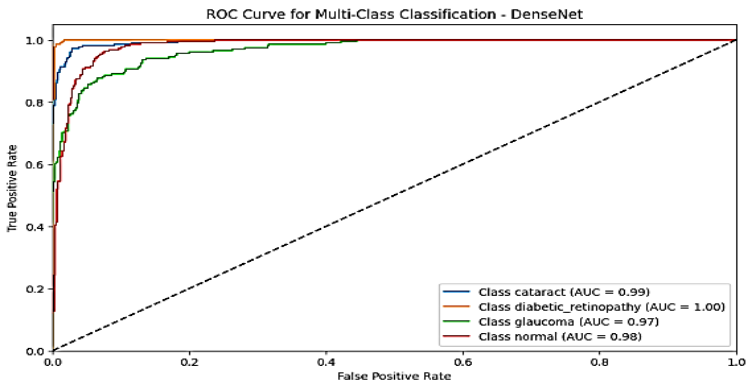

Figure 7. ROC curve of DenseNet

Table 4. Classification of DenseNet

Class	Precision	Recall	F1-score
Cataract	0.89	0.97	0.93
Diabetic Retinopathy	1.00	0.98	0.99
Glaucoma	0.87	0.81	0.84
Normal	0.88	0.89	0.89
Accuracy		0.91	
Macro Average	0.91	0.91	0.91

Figure 7 shows the ROC curves for DenseNet121. AUC values were 1.00 for diabetic retinopathy, 0.99 for cataract, 0.98 for normal, and 0.97 for glaucoma, confirming strong multi-class discrimination.

5.3 Ensemble results

Table 5 compares ensemble strategies using accuracy, precision, recall, and F1-score. Bagging with Random Forest achieved the highest accuracy (0.9941), indicating strong robustness and stability. Stacking, soft voting, and hard voting also improved performance relative to individual CNN models, although with smaller gains.

Table 5. Comparison analysis of learning techniques used

The Model	Accuracy	Precision	Recall	F1-score
Bagging	0.994076	0.994069	0.994076	0.994067
Stacking	0.928910	0.928800	0.928910	0.928442
Soft Voting	0.915877	0.915457	0.915877	0.915097
Hard Voting	0.911137	0.910418	0.911137	0.910225

Table 6 compares training time after applying the training process by the graphics card RTX 4060 Ti and using parallel process for evaluating a short time, also the macro-average accuracy for InceptionV3, DenseNet121, and the combined CNN feature ensemble. DenseNet121 achieved higher macro-average accuracy than InceptionV3 and required shorter

training time. The fused feature ensemble further improved macro-average accuracy with a moderate increase in total computation.

Table 6. Training time with the accuracy comparison

Model	Cataract Acc (%)	Diabetic Retinopathy Acc (%)	Glaucoma Acc (%)	Normal Acc (%)	Macro Avg Accuracy (%)	Total Training Time (min)	Avg. Epoch Time (s)
InceptionV3	94.2	92.7	91.5	95.3	**93.4**	**58.4**	140.2
DenseNet121	95.1	94.0	92.6	96.1	**94.5**	**43.6**	104.6
Ensemble (Dense+Incep)	96.0	95.3	93.8	96.8	**95.5**	**62.1**	—

5.3.1 Results of best model: Bagging (Random Forest)

Table 7 reports class-wise performance for bagging with Random Forest. Cataract, diabetic retinopathy, and normal achieved perfect precision, recall, and F1-score. Glaucoma achieved near-perfect performance with an F1-score of 0.99. Overall test accuracy reached 0.99, and macro-average metrics remained consistently high across classes.

5.3.2 Improve boost results

Table 8 shows performance for the optimized boosting model. All classes achieved precision, recall, and F1-score of 1.00, resulting in 1.00 overall accuracy on the test set. Although this outcome is strong, it should be interpreted cautiously given the single-dataset evaluation and the potential for overfitting.

Table 7. Classification of bagging (Random Forest)

Class	Precision	Recall	F1-score
Cataract	1.00	1.00	1.00
Diabetic Retinopathy	1.00	1.00	1.00
Glaucoma	0.99	0.99	0.99
Normal	1.00	1.00	1.00
Accuracy		0.99	
Macro Average	0.99	0.99	0.99

Table 8. Classification of optimized boosting

Class	Precision	Recall	F1-score
Cataract	1.00	1.00	1.00
Diabetic Retinopathy	1.00	1.00	1.00
Glaucoma	1.00	1.00	1.00
Normal	1.00	1.00	1.00
Accuracy	-	-	1.00
Macro Average	1.00	1.00	1.00

Table 9. Comparison results

Model	Accuracy	Precision	Recall	F1-score
Bagging (Random Forest)	0.994076	0.994069	0.994076	0.994067
Stacking	0.928910	0.928800	0.928910	0.928442
Soft Voting	0.915877	0.915457	0.915877	0.915097
DenseNet121	0.912322	0.912409	0.912322	0.911648
Hard Voting	0.911137	0.910418	0.911137	0.910225
InceptionV3	0.851896	0.850582	0.851896	0.850469

5.4 Final comparison

Table 9 summarizes the performance of all evaluated models. Bagging with Random Forest produced the highest accuracy among the evaluated strategies. DenseNet121 outperformed InceptionV3, suggesting that it captured discriminative retinal patterns more effectively in this setting. Overall, ensemble models improved performance over individual CNNs, which supports the effectiveness of combining deep features with classical ensemble learners for retinal disease classification.

5.5 Grad-CAM results

Figure 8 shows representative Grad-CAM visualizations. The heatmaps highlight image regions that contributed most to the predicted class. High-importance regions are indicated by warm colors. In several examples, activations focused on clinically relevant areas, such as the optic disc and vascular structures, supporting interpretability of the model decisions.

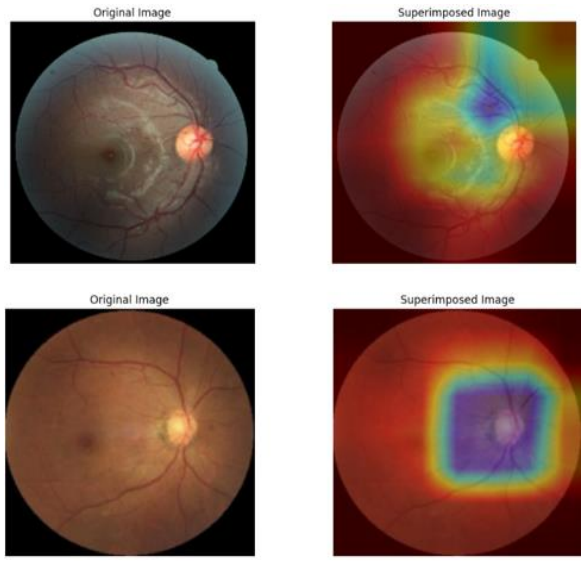

Figure 8. Grad-CAM results

Strong activation around the optic disc and surrounding vascular structures is shown in the first row of the heat map, indicating that the model uses these features for classification. With a distinct focus in the central retinal region, the activation in the second row is more localized and may suggest pathology. These findings demonstrate that the model successfully detects important diagnostic characteristics, improving the explanatory ability and reliability of AI-assisted medical diagnostics.

5.6 Discussion

Although the proposed ensembles achieved very high test accuracies (up to 99–100%), these results should be interpreted cautiously. First, evaluation was conducted on a single public dataset, which may not reflect the variability observed in real clinical environments, including differences in acquisition devices, illumination, and patient demographics.

Second, even with data augmentation, early stopping, and a held-out test set, overfitting cannot be fully excluded, particularly for tree-based models that can fit complex decision boundaries when feature representations are highly separable.

Future work will therefore focus on external validation using additional datasets, cross-dataset testing, and prospective evaluation in clinical workflows. Further analysis should also investigate calibration, error patterns by disease severity, and model robustness under domain shifts.

6. CONCLUSION

This study proposed a hybrid framework for multi-class classification of retinal fundus images into normal, cataract, glaucoma, and diabetic retinopathy. InceptionV3 and DenseNet121 were used as deep feature extractors, and several ensemble strategies were evaluated on the fused feature representations. DenseNet121 offered a favorable balance between predictive performance and training time. Ensemble learning further improved classification results, with bagging (Random Forest) and optimized boosting achieving the best performance on the held-out test set.

Grad-CAM visualizations suggested that the models relied on clinically plausible retinal regions, which supports interpretability. However, the study was limited to a single dataset and did not include external validation or expert-based quantitative assessment of explanations.

Future work will evaluate generalization across datasets and imaging devices, extend the framework to additional disease categories and multi-label settings, and assess clinical impact through prospective studies. Integrating multimodal inputs (e.g., OCT and clinical variables) is also a promising direction to further improve reliability.

REFERENCES

[1] Mahadiuzzaman, A.S.M., Hoque, M.E., Alam, S., Chawdhury, Z.T., Hasan, M., Rashid, A.B. (2024). Visual neuroprostheses for impaired human nervous system: State-of-the-art and future outlook. International Journal of Cell Biology, 2024(1): 2651763. https://doi.org/10.1155/ijcb/2651763

[2] Dutta, S., Wilson, M. (2021). Spatial mapping of distributed sensors biomimicking the human vision system. Electronics, 10(12): 1443. https://doi.org/10.3390/electronics10121443

[3] Kovács-Valasek, A., Rák, T., Pöstyéni, E., Csutak, A., Gábriel, R. (2023). Three major causes of metabolic retinal degenerations and three ways to avoid them. International Journal of Molecular Sciences, 24(10): 8728. https://doi.org/10.3390/ijms24108728

[4] Assi, L., Chamseddine, F., Ibrahim, P., Sabbagh, H., et al. (2021). A global assessment of eye health and quality of life: A systematic review of systematic reviews. JAMA Ophthalmology, 139(5): 526-541. https://doi.org/10.1001/jamaophthalmol.2021.0146

[5] Steinmetz, J.D., Bourne, R.R., Briant, P.S., Flaxman, S.R., et al. (2021). Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: The Right to Sight: An analysis for the Global Burden of Disease Study. The Lancet Global Health, 9(2): e144-e160. https://www.thelancet.com/JOURNALS/LANGLO/ARTICLE/PIIS2214-109X(20)30489-7/FULLTEXT.

[6] Daien, V., Eldem, B.M., Talks, J.S., Jean-Francois Korobelnik, J.F., et al. (2019). Real-world data in retinal diseases treated with anti-vascular endothelial growth factor (anti-VEGF) therapy – A systematic approach to identify and characterize data sources. BMC Ophthalmology, 19(1): 206. https://doi.org/10.1186/s12886-019-1208-9

[7] Ahn, S.J. (2024). Real-world research on retinal diseases using health claims database: A narrative review. Diagnostics, 14(14): 1568. https://doi.org/10.3390/diagnostics14141568

[8] Nguyen, D.M., Alam, H.M.T., Nguyen, T., Srivastav, D., Profitlich, H.J., Le, N., Sonntag, D. (2025). Deep learning for ophthalmology: The state-of-the-art and future trends. arXiv preprint arXiv:2501.04073. https://doi.org/10.48550/arXiv.2501.04073

[9] Ting, D.S.W., Pasquale, L.R., Peng, L., Campbell, J.P., et al. (2018). Artificial intelligence and deep learning in ophthalmology. British Journal of Ophthalmology, 103(2): 167-175. https://doi.org/10.1136/bjophthalmol-2018-313173

[10] Masud, M. (2022). DL-CNN-based approach with image processing techniques for diagnosis of retinal diseases. Multimedia Systems, 28(4): 1417-1438. https://doi.org/10.1007/s00530-021-00769-7

[11] Haja, S.A., Mahadevappa, V. (2023). Advancing glaucoma detection with convolutional neural networks: A paradigm shift in ophthalmology. Romanian Journal of Ophthalmology, 67(3): 222-237. https://doi.org/10.22336/rjo.2023.39

[12] Anton, N., Doroftei, B., Curteanu, S., Catalin, L., Ilie, O.D., Târcoveanu, F., Bogdănici, C.M. (2022) Comprehensive review on the use of artificial intelligence in ophthalmology and future research directions. Diagnostics, 13(1): 100. https://doi.org/10.3390/diagnostics13010100

[13] Shree, R., Madagaonkar, S., Prateek, L.A., Tony, A., Rathnamma, M.V., Venkata Ramana, V., Chandrasekaran, K. (2022). Application of ensemble methods in medical diagnosis. In International Conference on Innovations in Data Analytics (ICIDA 2022), West Bengal, India, pp. 355-367. https://doi.org/10.1007/978-981-99-0550-8_29

[14] Mahajan, P., Uddin, S., Hajati, F., Moni, M.A., Gide, E. (2024). A comparative evaluation of machine learning ensemble approaches for disease prediction using multiple datasets. Health and Technology, 14(3): 597-613. https://doi.org/10.1007/s12553-024-00835-w

[15] Mahajan, P., Uddin, S., Hajati, F., Moni, M.A. (2023). Ensemble learning for disease prediction: A review. Healthcare, 11(12): 1808-1808. https://doi.org/10.3390/healthcare11121808

[16] Alavee, K.A., Hasan, M., Zillanee, A.H., Mostakim, M., et al. (2024). Enhancing early detection of diabetic retinopathy through the integration of deep learning models and explainable artificial intelligence. IEEE Access, 12: 73950-73969. https://doi.org/10.1109/ACCESS.2024.3405570

[17] Sarki, R., Ahmed, K., Wang, H., Zhang, Y., Wang, K. (2021). Convolutional neural network for multi-class

classification of diabetic eye disease. EAI Endorsed Transactions on Scalable Information Systems, 9(4). http:// doi.org/10.4108/eai.16-12-2021.172436

[18] Sarki, R., Ahmed, K., Wang, H., Zhang, Y. (2020). Automated detection of mild and multi-class diabetic eye diseases using deep learning. Health Information Science and Systems, 8(1): 32. https://doi.org/10.1007/s13755-020-00125-5

[19] Joshi, R.C., Sharma, A.K., Dutta, M.K. (2024). VisionDeep-AI: Deep learning-based retinal blood vessels segmentation and multi-class classification framework for eye diagnosis. Biomedical Signal Processing and Control, 94: 106273. https://doi.org/10.1016/j.bspc.2024.106273

[20] Vadduri, M., Kuppusamy, P. (2023). Enhancing ocular healthcare: Deep learning-based multi-class diabetic eye disease segmentation and classification. IEEE Access, 11: 137881-137898. https://doi.org/10.1109/access.2023.3339574

[21] Nawaz, A., Ali, T., Mustafa, G., Babar, M., Qureshi, B. (2023). Multi-class retinal diseases detection using deep CNN with minimal memory consumption. IEEE Access, 11: 56170-56180. https://doi.org/10.1109/access.2023.3281859

[22] Vaiyapuri, T., Srinivasan, S., Sikkandar, M.Y., Balaji, T.S., Kadry, S., Meqdad, M.N. (2022). Intelligent deep learning based multi-retinal disease diagnosis and classification framework. Computers, Materials & Continua, 73(3): 5543-5557. https://doi.org/10.32604/cmc.2022.023919

[23] Gualsaqui, M.G., Cuenca, S M., Rosero, I.L., Almeida, D.A., Cadena, C., Villalba, F., Cruz, J.D. (2023). Multi-class classification approach for retinal diseases. Journal of Advances in Information Technology, 14(3): 392-398. https://doi.org/10.12720/jait.14.3.392-398

[24] Ho, E., Wang, E., Youn, S., Sivajohan, A., Lane, K., Chun, J., Hutnik, C.M. (2022). Deep ensemble learning for retinal image classification. Translational Vision Science & Technology, 11(10): 39. https://doi.org/10.1167/tvst.11.10.39

[25] Manikandaprabhu, P., Subaash, S.S. (2024). Harnessing deep learning methods for detecting different retinal diseases: A multi-categorical classification methodology. International Journal of Innovative Science and Research Technology (IJISRT), 9(3): 2381-2391. https://doi.org/10.38124/ijisrt/IJISRT24MAR1824

[26] Khan, O.S., Abbas, R., Gilani, S.O., Waris, A. (2022). Ensemble based multi-retinal disease classification and application with RFMiD dataset using deep learning. SSRN, Electronic Journal. https://doi.org/10.2139/ssrn.4247846

[27] Doddi, G.V. (2022). Eye_diseases_classification. https://www.kaggle.com/datasets/gunavenkatdoddi/eye-diseases-classification.

[28] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., et al. (2025). Going deeper with convolutions. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, pp. 1-9. https://doi.org/10.48550/arXiv.1409.4842

[29] Azis, F.A., Suhaimi, H., Abas, P.E. (2023). The development of an automated waste segregator. International Journal of Integrated Engineering, 15(4): 19-30. https://doi.org/10.30880/ijie.2023.15.04.002

[30] Zhou, T., Ye, X., Lu, H., Zheng, X., Qiu, S., Liu, Y. (2022). Dense convolutional network and its application in medical image analysis. BioMed Research International, 2022(1): 2384830. https://doi.org/10.1155/2022/2384830

[31] Li, G., Zhang, M., Li, J., Lv, F., Tong, G. (2021). Efficient densely connected convolutional neural networks. Pattern Recognition, 109: 107610. https://doi.org/10.1016/j.patcog.2020.107610

[32] Chetan, R., Ashoka, D.V., Ajay Prakash, B.V. (2023). HybridTransferNet: Advancing soil image classification through comprehensive evaluation of hybrid transfer learning. https://doi.org/10.21203/rs.3.rs-3032907/v1

[33] Abdulla, A.N.A., Nair, L.R. (2024). A comprehensive study of ensemble models to improve the performance of cluster algorithms. Revue d'Intelligence Artificielle, 38(4): 1183-1192. https://doi.org/10.18280/ria.380412

[34] Salur, M.U., Aydın, İ. (2022). A soft voting ensemble learning-based approach for multimodal sentiment analysis. Neural Computing and Applications, 34(21): 18391-18406. https://doi.org/10.1007/s00521-022-07451-7

[35] Wang, H., Yang, Y., Wang, H., Chen, D. (2013). Soft-voting clustering ensemble. In 11th International Workshop on Multiple Classifier Systems (MCS 2013), Nanjing, China, pp. 307-318. https://doi.org/10.1007/978-3-642-38067-9_27

[36] Dey, R., Mathur, R. (2023). Ensemble learning method using stacking with base learner, a comparison. In International Conference on Data Analytics and Insights, Kolkata, India, pp. 159-169. https://doi.org/10.1007/978-981-99-3878-0_14

[37] Odegua, R. (2019). An empirical study of ensemble techniques (bagging, boosting and stacking). In Conference on Deep Learning IndabaXAt. https://doi.org/10.13140/RG.2.2.35180.10882

[38] Ngo, G., Beard, R., Chandra, R. (2022). Evolutionary bagging for ensemble learning. Neurocomputing, 510: 1-14. https://doi.org/10.1016/j.neucom.2022.08.055

[39] Selvaraju, R.R., Cogswell, M., Abhishek, D., Ramakrishna, V., Devi, P., Dhruv, B. (2020). Grad-CAM: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision, 128(2): 336-359. https://doi.org/10.1007/s11263-019-01228-7

[40] Foody, G.M. (2023). Challenges in the real world use of classification accuracy metrics: From recall and precision to the Matthews correlation coefficient. PLoS ONE, 18(10): e0291908. https://doi.org/10.1371/journal.pone.0291908

[41] Owusu-Adjei, M., Hayfron-Acquah, J. B., Frimpong, T., Abdul-Salaam, G. (2023). A systematic review of prediction accuracy as an evaluation measure for determining machine learning model performance in healthcare systems. medRxiv, 2023-06. https://doi.org/10.1101/2023.06.01.23290837

[42] Powers, D.M. (2020). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies, 2(1): 37-63. https://doi.org/10.48550/arXiv.2010.16061

[43] Keldenich, T. (2021). Recall, Precision, F1 Score – Simple Metric Explanation in ML [in French]. https://inside-machinelearning.com/recall-precision-f1-score/.