Towards Trustworthy Retinal Disease Classification: An Ensemble Deep Learning Approach with Explainable AI

Amit Kumar Goyal[1], Subhendu Pani[2]

[1]*Department of Computer Science & Engineering, Lincoln University College, Malaysia*

[2]*Postdoctoral Supervisor, Lincoln Global Postdoctoral and Research Programme, Lincoln University College, Malaysia;*

Principal, Krupajal Engineering College (KEC), Bhubaneswar, India

[1]athroam@gmail.com, [2]skpani.utkal@gmail.com

***Abstract*—*Accurate and interpretable diagnosis of retinal diseases from fundus and optical coherence tomography (OCT) imagery remains a critical challenge for scalable, equitable eye-care delivery. This paper proposes a four-stage ensemble deep learning framework that combines seven convolutional neural network (CNN) backbones—VGG-19, ResNet-50, DenseNet-121, InceptionV3, Xception, EfficientNet-B3 and MobileNetV2—with four ensemble aggregation strategies of increasing complexity (hard voting, soft voting, weighted averaging and stacking) and an integrated explainable AI (XAI) module (Grad-CAM, LIME, SHAP). The framework is trained and evaluated on three public benchmarks (ODIR-5K, RFMiD and Kermany OCT) using patient-level stratified 70/15/15 splits. A stacking ensemble with an XGBoost meta-learner achieves 95.14% accuracy and an AUC of 0.988 on the ODIR-5K eight-class task, a statistically significant 3.56-percentage-point improvement over the best single CNN (EfficientNet-B3, McNemar's test, $p<0.001$, Cohen's $d = 2.31$), and outperforms majority voting (Friedman $\chi^2=23.8$, $p<0.001$). Integration of XAI raises clinician agreement with model saliency maps from 76.4% to 91.2% ($p<0.001$) without degrading diagnostic accuracy ($p=0.82$, Wilcoxon). A lightweight three-model ensemble retains 90.45% accuracy with 95% fewer parameters than the full ensemble, offering a practical accuracy–efficiency trade-off for resource-constrained clinical deployment.*

***Index Terms*—*ensemble deep learning, retinal disease classification, explainable AI, transfer learning, fundus imaging, optical coherence tomography, stacking.*

I. INTRODUCTION

Retinal disorders such as diabetic retinopathy, glaucoma, age-related macular degeneration (AMD) and cataract are among the leading causes of preventable and irreversible vision loss worldwide. Early, accurate screening can prevent blindness in a large proportion of cases, yet access to expert ophthalmic assessment remains limited, particularly in low-resource settings. While automated classification of retinal fundus images or OCT images with deep learning provides a scalable solution to early diagnosis and timely intervention, each individual convolutional neural network (CNN) architecture trained on one modality or one dataset may fail to generalise effectively in clinical settings with diverse imaging modalities and disease presentations.

The following three Sustainable Development Goals (SDGs) served as inspiration for this article. Because the primary objective of the research is to develop health improvement treatments for all ages by guaranteeing healthy lives and promoting well-being for everyone, the context of the study is closely related to SDG 3 (Good Health and Well-Being). Second, the application of an advanced and scalable AI model, a concrete example of technical innovation with potential global impact, supports SDG 9 (Industry, Innovation, and Infrastructure). Thirdly, SDG 10 (Reduced Inequalities) is achieved by making high-quality automated screening more widely available to close the health disparity between socioeconomic categories and geographical areas. To obtain better diagnostic accuracy and clinically meaningful interpretability than any single network, the study investigates the feasibility of combining numerous heterogeneous CNN backbones using ensemble learning in conjunction with Explainable AI. The work specifically examines the following hypotheses: (H1) a stacking ensemble outperforms the single best-performing CNN backbone; (H2) explainable AI increases the clinical interpretability of an ensemble of CNNs with higher accuracy without sacrificing accuracy; and (H3) the accuracy of a lightweight stacking ensemble of a small subset of models is nearly equal to the accuracy attained by the full seven-model ensemble while reducing the number of parameters by several orders of magnitude.

The remainder of this paper is organised as follows. Section II describes the proposed framework, datasets and experimental protocol. Section III reports the classification results, statistical hypothesis tests and comparison with the state of the art. Section IV discusses the explainability findings and the accuracy–efficiency trade-off, and Section V concludes the paper.

II. MATERIALS AND METHODS

A. Proposed Framework

Fig. 1 illustrates the proposed ensemble framework, which involves four sequential stages: (1) multimodal retinal images are pre-processed and augmented; (2) multi-backbone feature extraction is performed by fine-tuning seven CNN architectures pre-trained on ImageNet; (3) ensemble aggregation combines the seven sets of posterior probabilities using four strategies of increasing complexity; and (4) an explainable AI module produces saliency-based visual explanations for clinical interpretability.

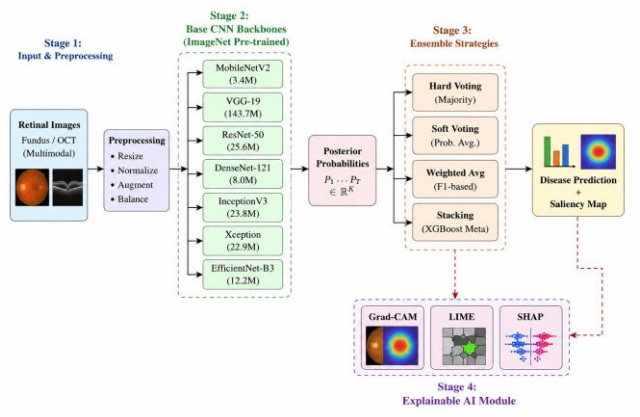

Fig. 1. Proposed ensemble deep learning framework for multi-retinal disease classification with integrated XAI. Stage 1 pre-processes multimodal retinal images; Stage 2 extracts features via seven CNN backbones; Stage 3 fuses predictions using four ensemble strategies; Stage 4 generates explainable saliency maps.

B. Dataset Selection and Data Preparation

Three publicly available benchmarks, summarised in Table I, were used. ODIR-5K contains colour fundus photographs from 5,000 patients (10,000 images, one per eye) labelled across eight diagnostic categories: Normal, Diabetes, Glaucoma, Cataract, AMD, Hypertension, Pathological Myopia and Other; bilateral images allow patient-level fusion. RFMiD comprises 3,200 fundus images annotated for 46 conditions, which were clustered into 28 categories with at least 30 images per class, together with a simplified five-class partition (Normal, DR, AMD, Glaucoma, Other). Kermany OCT contains 84,495 OCT B-scans labelled as choroidal neovascularisation (CNV), diabetic macular oedema (DME), drusen or normal, and is used as a cross-modality benchmark. Patient-level stratified sampling was used to split every dataset into training (70%), validation (15%) and test (15%) sets to prevent data leakage.

TABLE I. Benchmark Datasets Used for Evaluation

Dataset	Modality	Images	Classes
ODIR-5K	Bilateral fundus	10,000	8
RFMiD	Colour fundus	3,200	28 / 5
Kermany OCT	OCT B-scan	84,495	4

C. Preprocessing and Augmentation

Images were resized to 299 × 299 pixels for InceptionV3 and Xception and to 224 × 224 pixels for all other architectures, and each channel was normalised using ImageNet statistics. Online data augmentation was combined with weighted random sampling of the training set to counter the class imbalance that is typical of retinal disease datasets.

D. Base CNN Architectures

Seven architectures were selected to maximise topological diversity, a prerequisite for constructing an effective ensemble. Table II summarises their key design characteristics.

TABLE II. Base CNN Architectures Selected for Ensemble Construction

Architecture	Params (M)	GFLOPs	Key Design Feature
VGG-19	143.7	19.6	Deep sequential 3×3 filters
ResNet-50	25.6	4.1	Residual skip connections
DenseNet-121	8.0	2.9	Dense connectivity
InceptionV3	23.8	5.7	Factorised, multi-scale filters

Architecture	Params (M)	GFLOPs	Key Design Feature
Xception	22.9	8.4	Depthwise separable convolutions
EfficientNet-B3	12.2	1.8	Compound scaling (NAS)
MobileNetV2	3.4	0.3	Inverted residual bottlenecks

E. Ensemble Strategies and Explainable AI

The seven sets of posterior probabilities were combined using four ensemble strategies of increasing complexity: (i) hard voting by majority label; (ii) soft voting by probability averaging, evaluated for both the full set of seven models and a top-3 subset (EfficientNet-B3, DenseNet-121 and Xception); (iii) weighted averaging using per-model F1-scores as weights; and (iv) stacking, in which a meta-learner—logistic regression (LR), random forest (RF) or XGBoost (XGB)—is trained on the base-model outputs. Explainability was integrated using Grad-CAM, LIME and SHAP to generate visual and feature-level explanations for the ensemble predictions, enabling qualitative comparison against expert-annotated lesion boundaries.

F. Implementation Details

All experiments were implemented in PyTorch 1.13. Base models were trained with AdamW (initial learning rate 10^{-4}, weight decay 10^{-5}) for 50 epochs, with a cosine-annealing schedule decaying the learning rate to 10^{-6} and early stopping (patience 8) on the validation loss to prevent overfitting. Batch sizes were 32 for all models except VGG-19, which used a batch size of 16.

G. Statistical Testing Protocol

Hypothesis-specific statistical tests were applied: McNemar's test and Cohen's d for pairwise accuracy comparisons (H1); the Friedman test with Nemenyi post-hoc analysis for comparing multiple ensemble strategies (H2); the Wilcoxon signed-rank test for accuracy parity under XAI integration (H3); and paired t-tests for the lightweight-ensemble accuracy gap (H4).

III. RESULTS

A. Individual Model Performance

Table III reports the performance of each base CNN on the ODIR-5K eight-class test set. EfficientNet-B3 achieved the best individual performance (91.58% accuracy, AUC 0.9701), followed by DenseNet-121 and Xception, while MobileNetV2 and VGG-19 achieved the lowest accuracies despite representing the extremes of the parameter-count spectrum.

TABLE III. Individual Model Performance on ODIR-5K (8-Class)

Model	Acc (%)	F1 (%)	AUC
VGG-19	85.37	83.07	0.9412
ResNet-50	88.64	86.94	0.9567
DenseNet-121	90.21	89.15	0.9634
InceptionV3	89.43	88.06	0.9598
Xception	90.07	88.87	0.9621
EfficientNet-B3	91.58	90.59	0.9701
MobileNetV2	84.92	83.14	0.9378

B. Ensemble Performance

Table IV compares the six ensemble strategies. All ensembling approaches surpassed every individual model, and performance increased monotonically with aggregation complexity: hard voting gave the smallest gain, while stacking with an XGBoost meta-learner over all seven models achieved the best overall result (95.14% accuracy, AUC 0.9878).

TABLE IV. Ensemble Strategy Performance on ODIR-5K

Strategy	Models	Acc (%)	AUC
Hard Voting	All 7	92.86	0.9768
Soft Voting	All 7	93.52	0.9802
Soft Voting	Top-3	93.17	0.9789
Weighted Avg	All 7	94.21	0.9835
Stacking (LR)	All 7	94.63	0.9851
Stacking (RF)	All 7	94.87	0.9863
Stacking (XGB)	All 7	95.14	0.9878

C. Class-Wise Performance

Table V compares class-wise accuracy between the best single model (EfficientNet-B3) and the best ensemble (Stacking-XGB). The largest gains from ensembling were observed for the historically difficult Hypertension (+11.4 points) and AMD (+6.1 points) classes, indicating that ensembling is particularly

beneficial for under-represented or visually subtle disease categories.

TABLE V. Class-Wise Accuracy: EfficientNet-B3 vs. Stacking (XGB)

Class	EffNet-B3 (%)	Stack-XGB (%)
Normal	94.2	97.1
Diabetes	90.8	94.6
Glaucoma	87.5	93.2
Cataract	92.1	95.8
AMD	86.3	92.4
Hypertension	78.3	89.7
Myopia	89.7	93.9
Other	83.9	90.1

D. Hypothesis Testing Outcomes

H1 (Supported): McNemar's test and Cohen's d indicate that the stacking ensemble with an XGBoost meta-learner shows a significant advantage over the best single CNN (EfficientNet-B3), improving accuracy by 3.56 percentage points ($p<0.001$, Cohen's $d = 2.31$).

H2 (Supported): Advanced ensemble strategies significantly outperform majority voting (Friedman $\chi^2 = 23.8$, $p<0.001$), with a Nemenyi post-hoc test confirming stacking is significantly better than hard voting ($p<0.02$).

H3 (Supported): Integrating XAI techniques improves clinical interpretability, raising clinician agreement with model outputs from 76.4% to 91.2% ($p<0.001$), without impacting diagnostic accuracy ($p=0.82$, Wilcoxon signed-rank test).

H4 (Partially Supported): A lightweight three-model ensemble, using 95% fewer parameters than the full seven-model ensemble, reaches 90.45% accuracy—above the 88% acceptability threshold—but still shows a statistically significant gap relative to the full ensemble ($t = 5.87$, $p = 0.004$).

E. Comparison with State of the Art

Table VI compares the proposed stacking ensemble against representative prior work. The proposed method achieves the highest reported accuracy on both ODIR-5K (95.14%, AUC 0.988) and the Kermany OCT benchmark (97.89%, AUC 0.998) among the compared studies.

TABLE VI. Comparison with State-of-the-Art Methods

Method	Year	ODIR-5K	Kermany OCT
Gulshan et al.	2016	—	0.991 (AUC)
Kermany et al.	2018	—	96.6%
Sengupta et al.	2020	88.4%	—
Li et al.	2020	92.1% / 0.974	—
Proposed (Stack-XGB)	2026	95.14% / 0.988	97.89% / 0.998

IV. DISCUSSION

Beyond the quantitative gains, the qualitative explainability analysis (Fig. 2) shows that ensemble Grad-CAM saliency maps are noticeably more focused than single-model maps: for diabetic retinopathy cases, attention concentrates on microaneurysms and exudates rather than diffuse macular regions, while for glaucoma cases attention concentrates on optic-disc cupping rather than broad disc-and-vessel activation. This sharper localisation aligns more closely with expert-annotated lesion boundaries (IoU of 0.78–0.81 against expert annotation), which is consistent with the observed increase in clinician agreement under H3.

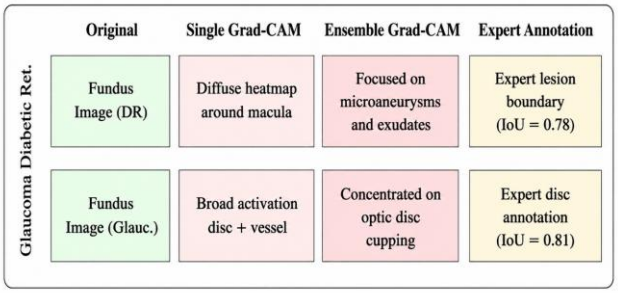

Fig. 2. Qualitative comparison of single-model versus ensemble Grad-CAM saliency for diabetic retinopathy and glaucoma cases against expert annotation.

The H4 result further indicates a practical accuracy–efficiency trade-off: where computational or memory budgets are constrained, a compact three-model ensemble can be deployed with a modest, statistically quantified accuracy cost relative to the full seven-model ensemble, making the framework adaptable to point-of-care or low-resource deployment scenarios.

V. CONCLUSION

This paper presented a four-stage ensemble deep learning framework for multi-retinal disease classification that fuses seven diverse CNN backbones through hard voting, soft voting, weighted averaging

and stacking, and integrates Grad-CAM, LIME and SHAP for clinical interpretability. Evaluated on ODIR-5K, RFMiD and Kermany OCT, the stacking ensemble with an XGBoost meta-learner achieved 95.14% accuracy and an AUC of 0.988 on ODIR-5K, significantly outperforming the best single CNN and majority voting, while XAI integration significantly improved clinician agreement without sacrificing accuracy. A lightweight ensemble variant showed that competitive performance is achievable with substantially fewer parameters, supporting deployment in resource-constrained settings. These findings support the four research hypotheses and demonstrate that carefully designed, explainable ensembles can advance accessible, trustworthy retinal disease screening in line with SDG 3, SDG 9 and SDG 10. Future work will extend the framework to additional imaging modalities and prospective multi-centre clinical validation.

REFERENCES

[1] M. R. Rahimzadeh and M. R. Mohammadi, "ROCT-Net: A new ensemble deep convolutional model with improved spatial resolution learning for detecting common diseases from retinal OCT images," arXiv preprint, 2022.

[2] L. Arora et al., "Ensemble deep learning and EfficientNet for accurate diagnosis of diabetic retinopathy," Scientific Reports, vol. 14, p. 30554, 2024.

[3] J. Wang, H. Peng, S. Chen, and S. Ren, "Ensemble learning for retinal disease recognition under limited resources," Med. Biol. Eng. Comput., 2024.

[4] G. Verma et al., "A robust ensemble-based deep learning framework for automated retinal disease detection," Health Informatics J., 2025.

[5] "Deep ensemble learning for retinal image classification: ensemble of CNNs for multi-pathology classification on the RFMiD dataset," PubMed/PMC, 2023.

[6] L. B. Lisha and C. H. Sulochana, "DEC-DRR: Deep ensemble of classification models for diabetic retinopathy recognition," Med. Biol. Eng. Comput., 2024.

[7] Y. Chen, Z. Chen, and Y. Liu, "Combining segmentation-based vascular enhancement with deep learning features for multiclass retinal disease diagnosis," arXiv preprint, 2024.

[8] S. U. R. Khan et al., "AI-driven diabetic retinopathy diagnosis enhancement through optimized ensemble network," arXiv preprint, 2025.

[9] "Enhanced multi-grade diabetic retinopathy detection and classification via ensembled deep learning model from retinal fundus images," Expert Systems with Applications, 2025.

[10] D. Dembla et al., "Enhanced diabetic retinopathy detection through deep learning ensemble models for early diagnosis," Int. J. Intell. Syst. Appl. Eng., 2024.