JSUCIT

Journal of Shaqra University for Computing and Information Technology
ISSN 2961-4279 (Print) || ISSN 2961-4287 (Online)

An Ensemble Deep Learning Framework with Age-Guided Feature Fusion for Multi-Label Ocular Disease Classification and Explainable Predictions

Samira Sheneamer[1], Samia Dardouri[1]*

[1]Department of Computer Science/College of Computing and Information Technology, Shaqra University, Saudi Arabia
Email: s.dardouri@su.edu.sa

Abstract : Automated multi-label ocular disease detection remains a significant challenge due to the visual similarity between pathologies and the absence of clinical context in image-only models. This study proposes an ensemble deep learning framework that integrates retinal fundus images with patient age metadata for multi-label classification across eight ocular disease categories on the ODIR-5K dataset [1]. The framework combines three pretrained convolutional neural networks—EfficientNet-B1, ResNet-50, and DenseNet-121—with a custom age-gated HybridModel that applies attention-based feature modulation conditioned on patient age [3,4]. Training employed the AdamW optimizer with cosine annealing scheduling, weighted binary cross-entropy loss to address class imbalance, and early stopping. Per-class threshold optimization was applied following ensemble aggregation. The ensemble achieved a Hamming accuracy of 86.24%, a macro AUC of 0.8382, an F1-micro of 0.5758, and an F1-macro of 0.6164 on the patient-level validation set. Grad-CAM explainability [6] was integrated as a qualitative visualization of prediction-related image regions, while Monte Carlo Dropout was used to estimate predictive variability. Because the same validation set was used for threshold optimization and final metric calculation, these results should be interpreted as validation-set estimates rather than independent test performance. Independent patient-level testing, formal ablation, and quantitative uncertainty/calibration analysis remain necessary to establish generalizability and clinical utility.

***Index Terms*— deep learning; ocular disease; ensemble learning; multi-label classification; explainable AI; Grad-CAM; ODIR-5K; age-gated fusion; EfficientNet; ResNet-50; DenseNet**

I. INTRODUCTION

The rapid global increase in chronic ocular conditions, such as Diabetic Retinopathy, Glaucoma, and Age-related Macular Degeneration (AMD), has placed an unprecedented burden on healthcare systems [7,8]. Early and precise screening is important for preventing irreversible vision loss; however, traditional diagnostic workflows remain heavily dependent on the availability of expert ophthalmologists and manual interpretation of fundus photographs. In recent years, Deep Learning (DL), specifically Convolutional Neural Networks (CNNs), has emerged as a useful approach in medical imaging. Despite these advances, three gaps remain relevant to automated ocular disease detection. First, many models rely primarily on image-based features and may not exploit available clinical metadata such as age. Second, real-world clinical scenarios may involve patients presenting with multiple coexisting conditions, whereas many systems are formulated as single-label classification tasks. Third, the 'black-box' nature of deep neural networks motivates the use of explainability methods when predictions are intended for clinical review. To address these issues, this study proposes an ensemble framework that integrates EfficientNet-B1,

ResNet-50, and DenseNet-121 with a custom HybridModel featuring an age-gated attention mechanism inspired by multimodal learning approaches [4,12]. The ensemble aggregates model outputs using predefined weights, followed by per-class threshold optimization on the validation set. Explainable AI (XAI) is incorporated through Grad-CAM [6] to generate qualitative heatmaps, and Monte Carlo Dropout is used to estimate predictive variability. The study uses a patient-level train/validation split; however, the validation set is also used for threshold selection and final reporting, so an independent test-set assessment remains a necessary next step.

II. Related Works

The automated classification of ocular diseases has significantly advanced with the development of deep learning techniques [9]. Early approaches primarily focused on single-disease or single-label classification tasks, limiting their applicability in clinical environments where multiple ocular conditions may coexist simultaneously. Single-Model CNN Approaches. Several studies have utilized convolutional neural networks for ocular disease classification using retinal fundus images. Deepak and Bhat (2024) [14] applied transfer learning techniques for automated ocular disease classification, achieving high diagnostic performance. Similarly, Jeny et al. (2023) [15] employed deep neural network ensemble architectures for retinal disease detection, demonstrating improved classification robustness. Wang et al. (2024) [26] further explored transformer-based multi-label retinal disease classification, highlighting the effectiveness of attention mechanisms for capturing complex pathological relationships. Despite these advances, many existing approaches rely on standalone architectures that do not leverage the complementary representational strengths of multiple networks and do not integrate demographic metadata into the diagnostic process.

Addressing Class Imbalance. Class imbalance remains a persistent challenge in ocular disease datasets. Alqassab et al. (2026) [10] proposed a hybrid quantum-classical deep learning framework that improved classification robustness under imbalanced distributions. Vadduri and Kuppusamy (2023) [16] introduced a hybrid deep learning framework combining multiple feature representations to improve ocular disease classification performance. Although these studies address imbalance and feature representation challenges, they do not incorporate demographic metadata or provide comprehensive explainability mechanisms for clinical interpretation.

Metadata Integration and Multimodal Learning. The integration of clinical metadata and multimodal learning has gained increasing attention in recent years. Khan et al. (2025) [4] proposed a framework combining retinal image features with auxiliary clinical information, demonstrating improved classification robustness and diagnostic performance. Similarly, Binu (2025) [7] emphasized the effectiveness of multimodal data fusion for enhancing retinal disease analysis. These findings suggest that demographic and clinical metadata can provide complementary contextual information beyond image-only representations. However, many previously reported accuracy values are based on single-label or differently partitioned evaluation protocols and are therefore not directly comparable to results obtained using strict patient-level multi-label splits as adopted in the present study.

Explainable AI in Medical Imaging. Explainability has become a critical requirement for trustworthy clinical AI systems. Kher et al. (2025) [5] proposed an explainable ensemble framework integrating Grad-CAM visualization for ocular disease classification. Altayeb et al. (2025) [6] further developed an explainable AI framework for multi-disease retinal analysis and diabetic retinopathy grading. In addition, Sureja et al. (2025) [23] and Alqutayfi et al. (2025) [24] highlighted the importance of explainability techniques for improving transparency and clinical trust in retinal disease prediction systems. Despite these advances, most existing methods either lack uncertainty estimation mechanisms or fail to combine explainability with clinically reliable confidence assessment.

Research Gaps and Proposed Contributions. As summarized in Table 1, existing approaches present several important limitations, including reliance on single-model architectures, absence of demographic metadata fusion, limited explainability, and lack of patient-level dataset splitting. The proposed framework addresses these limitations by combining ensemble deep learning, age-guided metadata fusion, Grad-

CAM interpretability, and Monte Carlo Dropout uncertainty estimation within a unified multi-label ocular disease classification framework evaluated using a rigorous patient-level split on the ODIR-5K dataset [1].

Table 1. Comparison of existing methods.

Study	Method	Metadata	XAI	Accuracy / AU
Deepak and Bhat (2024)[14]	Transfer Learning CNN	No	No	Acc: 94.2%
Wang et al. (2025) [27]	Transformer-based Multi-label CNN	No	No	Acc: 95%
Alqassab et al. (2026) [10]	Hybrid Quantum CNN	No	No	Acc: 97.8%
Kher et al. (2025) [5]	Explainable Ensemble Framework	No	Yes	Enhanced interpretability
Jeny et al. (2023) [15]	Ensemble Deep Learning	No	No	Acc: 95.6%
Khan et al. (2025) [4]	CNN + Metadata Fusion	Yes	No	Improved classification robustness
Vadduri and Kuppusamy (2023)[16]	Hybrid Deep Learning	No	No	Acc: 96.4%
Altayeb et al. (2026) [6]	XAI + Retinal Disease Analysis	No	Yes	Improved clinical transparency
Proposed	**Ensemble + Age-Gate + XAI**	**Yes**	**Yes**	**Acc: 0.8624**

III. Materials and Methods

A. Dataset description

Label	Disease	Description	Train Coun
N	Normal	No pathological findings	2101
D	Diabetes	Diabetic Retinopathy	2123
G	Glaucoma	Optic disc abnormalities	397
C	Cataract	Lens opacity	402
A	AMD	Age-related Maci Degeneration	319
H	Hypertensio	Hypertensive Retinopathy	203
M	Myopia	Pathological Myopia	306
O	Others	Other retinal abnormalities	1588

The empirical foundation of this research is the Ocular Disease Intelligent Recognition dataset, known as ODIR-5K [1]. This structured ophthalmic database comprises fundus photographs from 5,000 patients, representing a real-world clinical environment with data collected from multiple medical institutions using diverse fundus camera systems. The dataset was constructed as a benchmark for intelligent ocular disease recognition and has been widely adopted in the research community [11]. Each clinical case in the ODIR-5K dataset is annotated with bilateral fundus images (left and right eye) and patient-level metadata including patient age and sex. The dataset is categorized into eight distinct patho-logical classes: Normal (N), Diabetes (D), Glaucoma (G), Cataract (C), Age-related Macular Degeneration (A), Hypertension (H), Pathological Myopia (M), and Other Abnormalities (O). This multi-label structure allows a single patient to be associated with more than one condition simultaneously, reflecting real clinical complexity [14].

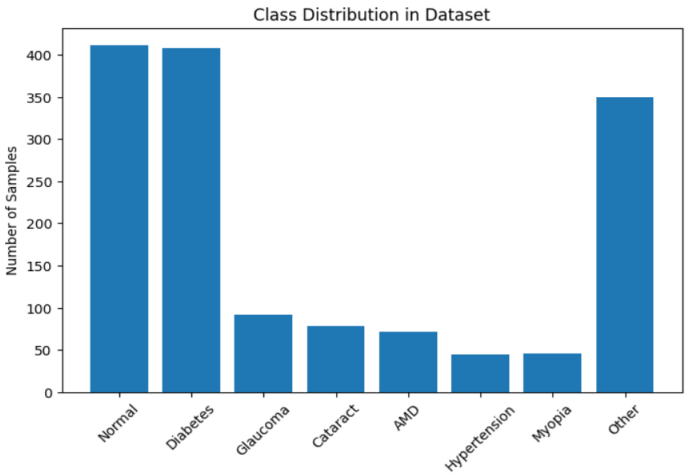


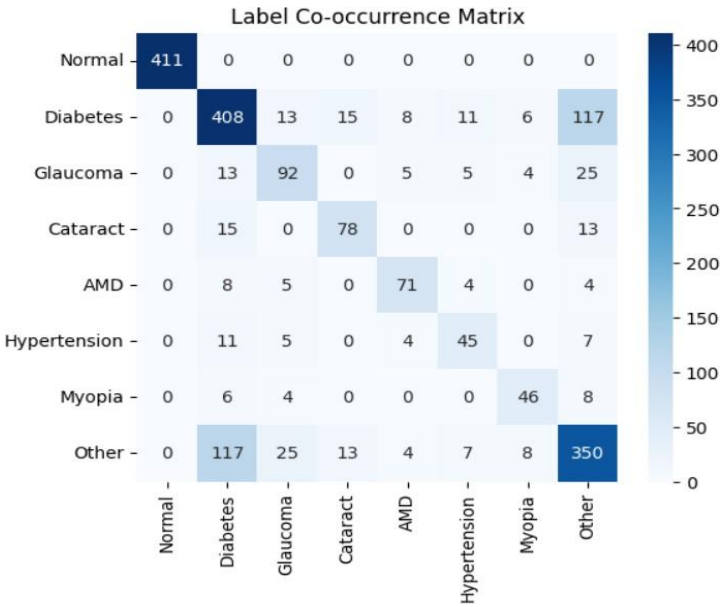
Figure 1. Class distribution of the ODIR-5K dataset showing significant imbalance across disease categories. Normal and Diabetes classes dominate the dataset, while Hypertension and AMD have relatively fewer samples, posing challenges for model generalization.

The class distribution shown in Figure 1 highlights a significant imbalance in the dataset. The Normal and Diabetes categories contain the majority of samples, whereas classes such as Hypertension and AMD are underrepresented. This imbalance introduces challenges for model training and was addressed using weighted loss functions during training.

Table 2. Distribution of ocular disease labels in the ODIR-5K dataset.

Figure 2. Co-occurrence matrix illustrating the frequency of multiple disease labels appearing together in the same patient. The matrix reflects the multi-label nature of the dataset and the presence of overlapping pathological conditions.

Figure 2 demonstrates the co-occurrence patterns between different disease categories. The presence of multiple labels per sample confirms the multi-label



nature of the dataset, where patients may exhibit more than one ocular condition simultaneously. This complexity increases the difficulty of the classification task compared to traditional single-label settings. The processed dataset used in this study comprises 6,392 image-level records after pairing left and right eye images per patient. The distribution of diagnostic categories is presented in Table 2. The dataset exhibits significant class imbalance, with Normal and

Diabetes dominating the sample distribution, while Hypertension contains the fewest samples. This imbalance directly informed the class weighting strategy described in Section 3.2.

B. Data Processing

A structured multi-stage preprocessing pipeline was implemented to ensure numer- ical stability and optimal convergence during training [18]. The pipeline covers image standardization, normalization, augmentation, metadata encoding, and data splitting.

Image Resizing and Normalization: All raw fundus images were resized to a uniformresolution of 224 × 224 pixels, aligning with the input dimensions of the pretrained backbone architectures. Images were converted into standard deep learning tensors [19] and nor-malized using standard ImageNet parameters (= [0.485, 0.456, 0.406], = [0.229, 0.224, 0.225]) to ensure consistent data distribution across all models.

Data Augmentation: To improve generalization and reduce overfitting, the training set was subjected to an augmentation pipeline [18] consisting of: random horizontal flipping, random vertical flipping, random rotation of up to ±15 degrees, and color jitter with brightness=0.3, contrast=0.3, and saturation=0.2. No augmentation was applied to the validation set to ensure unbiased evaluation.

Metadata Encoding: Patient age was used as the sole clinical metadata feature. Age 138 values were normalized to the range [0, 1] by dividing by 100 before being passed 139 to the model, ensuring numerical compatibility with image-derived features and 140 preventing the age scalar from dominating the feature space [3].

Class Imbalance Handling: To address the significant class imbalance in the ODIR-5K dataset [10], class weights were computed as inversely proportional to class frequency from the training set. The computed weights were normalized by their mean and clamped to a maximum value of 10.0 to prevent extreme gradients. These weights were applied as positive class weights in the Binary Cross-Entropy loss function, en-couraging the model to pay greater attention to minority classes such as Hypertension (H) and Glaucoma (G).

Data Splitting: To prevent patient-level data leakage — a methodological flaw present

in several existing ODIR-5K studies — the dataset was partitioned using a Group-

ShuffleSplit strategy [20] that groups image samples by patient ID. This ensures that images from the same patient appear exclusively in either the training or validation set. The dataset was split into 80% training (5,122 samples) and 20% validation (1,270 samples). The overlap between training and validation patient IDs was verified to be zero, confirming the integrity of the split.

IV. METHODOLOGY

The task is formulated as a multi-label classification problem where each fundus image may simultaneously be associated with multiple disease labels. The proposed framework consists of four independently trained models combined through a weighted ensemble strategy with per-class threshold optimization. The overall architecture of the proposed system is illustrated in Figure 3.

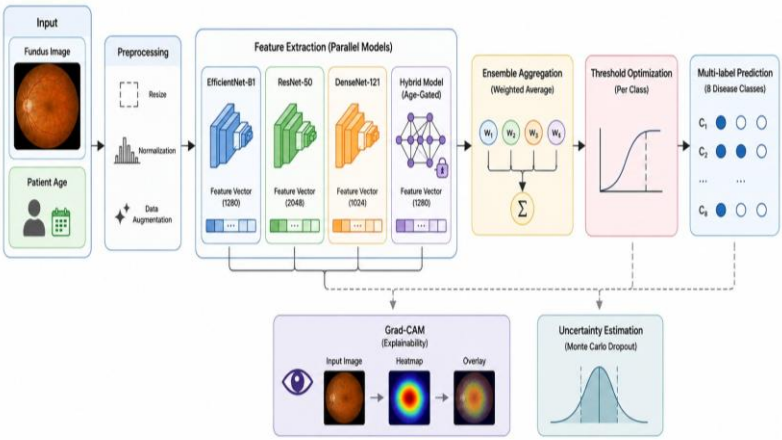


Figure 3. Proposed ensemble architecture integrating multiple CNN backbones with age-guided feature fusion and weighted aggregation.

A. Base CNN Backbones

Three pretrained CNN architectures were fine-tuned on the ODIR-5K dataset as base classifiers:

EfficientNet-B1 (pretrained on ImageNet1K-V2): The compound scaling method of EfficientNet provides an efficient balance between network depth, width, and input resolution [21]. The final classifier layer was replaced with a linear layer mapping the 1,280-dimensional feature vector to 8 output classes.

ResNet-50 (pretrained on ImageNet1K-V2): ResNet-50 utilizes residual connections to improve gradient flow in deep neural networks.

DenseNet-121 (pretrained on ImageNet1K-V1): DenseNet connects each layer to every other layer in a feed-forward fashion, promoting feature reuse and gradient flow. The classifier was replaced with a linear layer (1024 → 8). These three architectures were selected for their architectural diversity EfficientNet provides efficient compound scaling, ResNet provides residual connectivity, and DenseNet provides dense feature reuse — ensuring that the ensemble captures complementary spatial representations of retinal pathology.

B. Age-Gated HybridModel

To explicitly incorporate patient age as a clinical diagnostic context, a custom HybridModel was designed based on the EfficientNet-B1 backbone, inspired by multimodal fusion approaches [3,4]. Let x denote the fundus image and a the patient age. Age is normalized as a_norm = Age/100. The EfficientNet-B1 feature extractor produces a 1,280-dimensional feature vector f = E(x) ∈ R^1280. The normalized age is passed through a linear projection followed by a Sigmoid activation to generate an age-dependent gate g(a) = sigmoid(W_a a_norm + b_a) ∈ R^1280. The gated representation is then obtained by element-wise multiplication, h = f ⊙ g(a). This 1,280-dimensional representation is passed through a classification head consisting of Linear(1280→512), ReLU, and Dropout(0.4), followed by Linear(512→256), ReLU, and Dropout(0.3), and a final Linear(256→8) layer producing eight class logits. At inference, sigmoid probabilities are obtained from the output logits. Thus, age does not enter as a simple concatenated scalar; instead, it modulates the visual feature vector through a learned 1,280-dimensional gate. The HybridModel achieved lower standalone F1 performance than the strongest base CNN when evaluated at a fixed threshold of 0.5 (Table 4). The longer training duration observed in the current run is descriptive of the optimization behavior, but it is not sufficient to establish that architectural complexity caused the lower performance. A controlled ablation is therefore required before attributing any ensemble gain specifically to age gating.

C. Training Configuration

All four models were trained independently using the AdamW optimizer with a learning rate of 1e-4 and weight decay of 1e-4 [32]. A CosineAnnealingLR scheduler with T_max equal to the maximum training duration (20 epochs) was applied. Gradient clipping at a maximum norm of 1.0 was used to stabilize training. The batch size was 32 for both training and validation data loaders. Each model was trained for up to 20 epochs with early stopping (patience=5) based on minimum validation loss, with the best model state saved for inference. A threshold of 0.25 was used for the training/validation accuracy calculation during optimization. The patient-level GroupShuffleSplit used random_state=42 for the 80/20 split. The original implementation did not explicitly set a global PyTorch/NumPy/Python training seed; therefore, exact run-to-run determinism is not claimed for the reported training results.

Table 3. Per-class optimized classification thresholds.

Disease	Threshold	Validation F1
Normal (N)	0.10	0.573
Diabetes (D)	0.10	0.564
Glaucoma (G)	0.31	0.568
Cataract (C)	0.49	0.792
AMD (A)	0.41	0.558
Hypertension (H)	0.27	0.419
Myopia (M)	0.41	0.828
Others (O)	0.10	0.482

The optimized thresholds are reported in Table 3. The per-class optimization strategy is applied only to the ensemble predictions; individual model baselines in Table 4 are evaluated at a fixed threshold of 0.5 to describe their standalone behavior before threshold tuning. This distinction is important because the difference between the individual-model results and the final ensemble reflects both model aggregation and threshold optimization, rather than ensembling alone.

D. Explainability Module (Grad-CAM)

To enhance clinical transparency, the framework incorporates Grad-CAM (Gradient-weighted Class Activation Mapping) [6]. This module computes the gradients of the target class score relative to the feature maps of the final convolutional layer. The gradients are globally averaged to produce channel-wise importance weights, which are used to generate a weighted combination of the activation maps. A ReLU operation is applied to retain only the regions that positively influence the prediction. Grad-CAM visualizations were generated for both ResNet-50 (targeting the last residual block: layer4[-1]) and DenseNet- 121 (targeting the last dense block: features[-1]). By generating these weighted activation heatmaps, the system provides a visual justification for its predictions, highlighting specific retinal regions — such as the optic disc or lesion areas — that most influenced the diagnostic outcome [6,17].

E. Uncertainty Estimation

To estimate predictive variability, Monte Carlo Dropout (MC Dropout) [31] was applied during inference by keeping the dropout layers active and performing 30 stochastic forward passes for each input image. For each pass, the model produced a probability vector for the eight disease classes. The mean probability across the 30 stochastic predictions was used as the final predictive probability, while the standard deviation across these predictions was used as an empirical measure of predictive variability. Samples with a mean uncertainty value greater than 0.15 were flagged for potential specialist review. This threshold was selected empirically from the validation-set uncertainty distribution. Since no independent uncertainty calibration or uncertainty-discrimination analysis was performed, this procedure should be interpreted as an empirical variability-based screening mechanism rather than evidence of calibrated clinical uncertainty or improved clinical reliability.

V. EXPERIMENTAL RESULTS

A. Training Behavior

All four models were successfully trained on the ODIR-5K training set. EfficientNet-B1 triggered early stopping at epoch 10 with a best validation loss of 0.1752. ResNet-50 stopped at epoch 8 with a best validation loss of 0.1799. DenseNet-121 stopped at epoch 9 with a best validation loss of 0.1785. The HybridModel required more epochs to converge, stopping at epoch 12 with a best validation loss of 0.1791. All models demonstrated stable training curves with consistent loss reduction in early epochs followed by convergence, with no signs of divergence or instability.

B. Individual Model Performance

Table 4 presents the performance of each individual model on the validation set evaluated at a fixed threshold of 0.5. This fixed threshold is intentionally used for individual model evaluation to reflect standalone behavior prior to ensemble aggregation and threshold optimization. As expected in heavily imbalanced multi-label settings, the fixed threshold of 0.5 produces conservative recall values across all models. The ensemble with per-class threshold optimization is presented separately in Section 5.3 to avoid conflating the gains from ensembling with those from threshold tuning.

Table 4. Performance comparison of individual models and the proposed ensemble.

Model	F1 Macro	F1 Micro	AUC (Macro)	Precision	Recall
EfficientNet-B1	0.3582	0.1942	0.8289	0.7626	0.1113
ResNet-50	0.3393	0.1884	0.8141	0.7118	0.1086
DenseNet-121	0.3758	0.2513	0.8225	0.7759	0.1499
HybridModel (Age-Gated)	0.3270	0.1776	0.8125	0.7204	0.1013
Ensemble (Proposed)	**0.5978**	**0.5591**	**0.8355**	**0.4919**	**0.6476**

DenseNet-121 achieved the highest standalone F1-macro (0.3758) and AUC (0.8225) among the individual models, while the HybridModel achieved a lower standalone F1-macro (0.3270) and AUC (0.8125) at the fixed threshold of 0.5. The lower standalone performance of the HybridModel does not by itself establish that age gating caused the difference. A controlled ablation study is required to isolate the effects of age information and the gating mechanism and to determine whether the HybridModel provides a measurable contribution to the ensemble.

C. Final Ensemble Results

After applying per-class threshold optimization, the ensemble achieved the results shown in Table 5. The improvement from fixed-threshold individual models (F1-macro: 0.3270–0.3758) to the optimized ensemble (F1-macro: 0.6164) reflects the combined benefit of architectural diversity and per-class threshold tuning [20,23].

*Specificity is computed per-label and averaged; as with Hamming accuracy, this value may be inflated in multi-label settings where negative labels dominate [20].

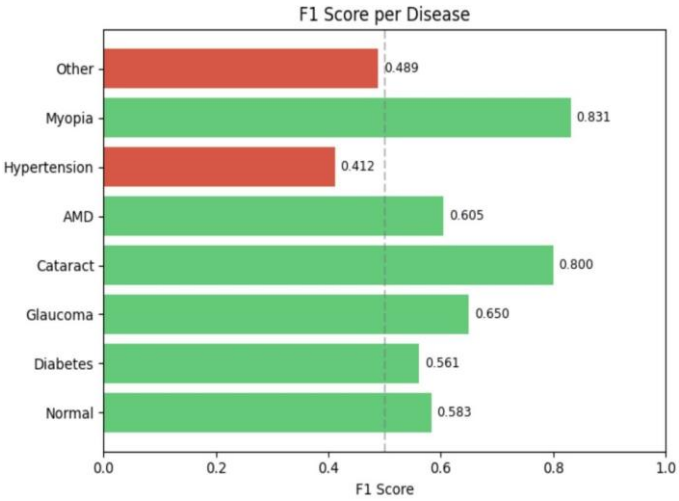
Table 5. Per-class F1 scores of the ensemble.

Metric	Value
Accuracy	0.8624
F1-micro	0.5758
F1-macro	0.6164
F1-sample	0.5715
Precision (micro)	0.5287
Recall (micro)	0.6322
Specificity*	0.8817
AUC (macro)	0.8382

D. Per-Class F1 Scores

Table 6 presents the per-class F1 scores of the final ensemble. Myopia (M) and Cataract (C) achieved the highest F1 scores (0.8315 and 0.8000 respectively), consistent with their visually distinct retinal features. Hypertension (H) and Others (O) achieved the lowest scores (0.4118 and 0.4886), attributable to severe class imbalance and the *heterogeneous nature of the Others category respectively. These results indicate that while the proposed class weighting strategy partially mitigates imbalance effects, minority classes remain challenging and represent an avenue for future improvement.*

Figure 5. F1-score per disease class for the proposed ensemble model.



E. ROC-AUC Analysis

The per-class AUC scores from the ensemble are: Normal (0.74), Diabetes (0.75),Glaucoma (0.92), Cataract (0.95), AMD (0.90), Hypertension (0.83), Myopia (0.94), Others (0.69). Cataract and Myopia achieve the highest AUC values (0.95 and 0.94 respectively), confirming the model's strong discriminative ability for visually distinct conditions. The lower AUC for Others (0.69) reflects the heterogeneous nature of this category, which encompasses a wide range of unspecified retinal abnormalities. The macro-averaged AUC of 0.8382 indicates good overall discriminative performance across all eight classes [9].

Class	F1 Score	Interpretation
Normal (N)	0.5832	Moderate — high overlap
Diabetes (D)	0.5614	Moderate — subtle features
Glaucoma (G)	0.6497	Moderate — small class
Cataract (C)	0.8000	High — distinct
AMD (A)	0.6050	Moderate — similar patterns
Hypertension (H)	0.4118	Low — minority class
Myopia (M)	0.8315	Highest — strong features
Others (O)	0.4886	Low — heterogeneous

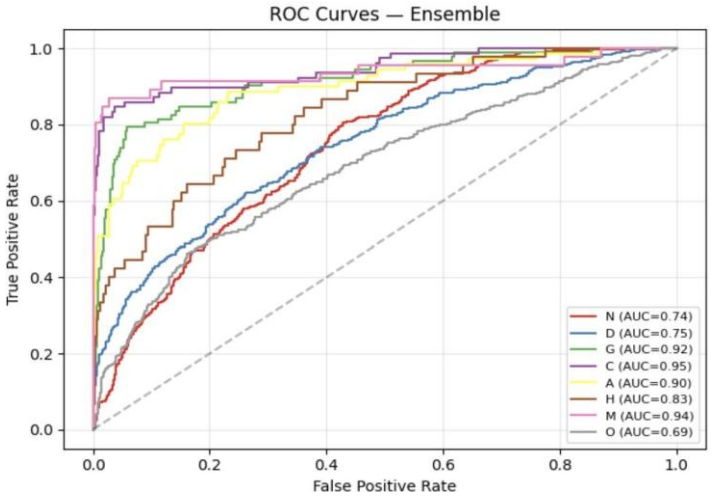


Figure 6. ROC curves for each disease class using the ensemble model.

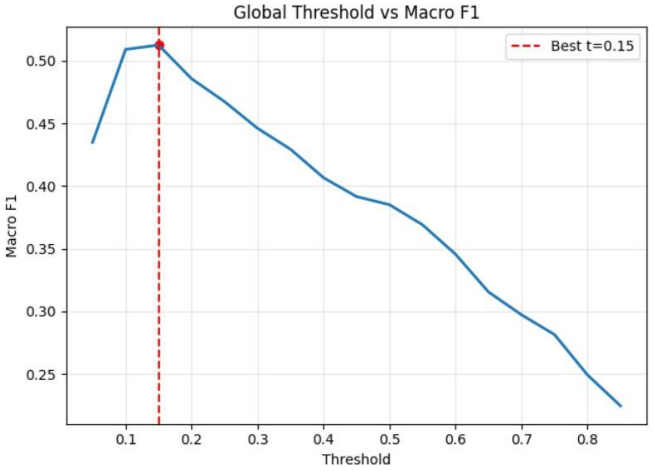


Figure 7. Effect of threshold selection on macro F1-score.

F. Grad-CAM Visualization Results

The Grad-CAM visualizations (Figures 8 and 9) suggest that the model focuses on retinal regions associated with the selected predictions [6]. For Cataract cases (label C), the ResNet-50 Grad-CAM highlights lens opacity regions in the central field of view. For Myopia (label M), the model's attention is directed toward peripheral retinal areas associated with myopic degeneration. For Other conditions (label O), activation maps show focus on the optic disc region, consistent with various disc-related abnormalities. These qualitative results illustrate the retinal regions associated with selected model predictions. However, the visualizations should not be interpreted as evidence of clinical validity because the current study does not quantitatively compare the generated heatmaps with expert-annotated lesion or anatomical masks. Such quantitative evaluation is left for future work.

Future work should include quantitative localization evaluation against expert-annotated lesion masks to further substantiate these finding [17].

Figure 8. Grad-CAM visualizations using ResNet50 highlighting important retinal regions.

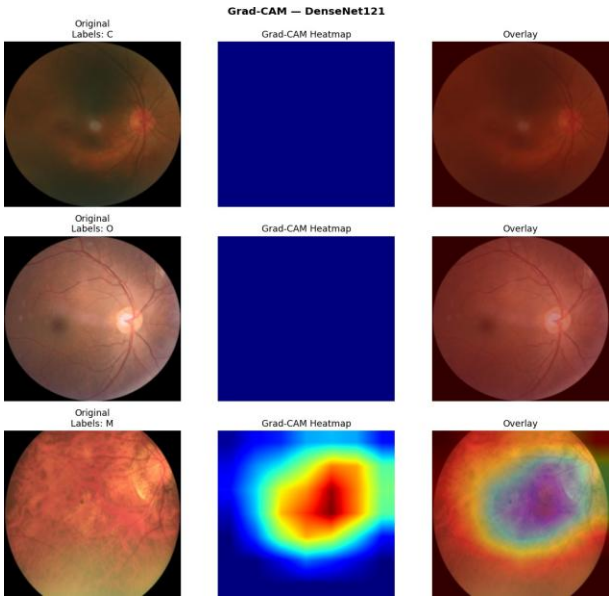
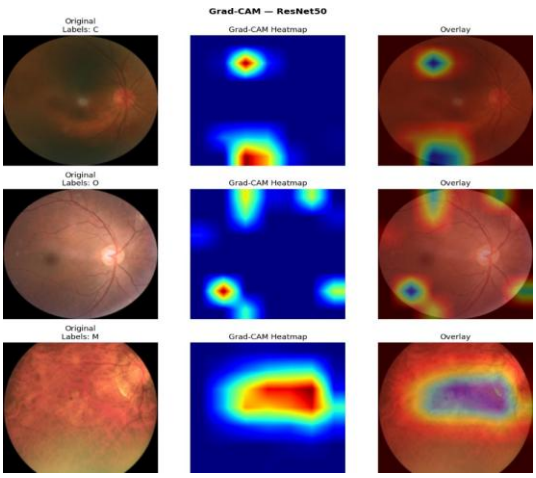


Figure 9. Grad-CAM visualizations using DenseNet121 highlighting important retinal regions.



VI. EVALUATION METRICS

Given the multi-label nature of the task, a comprehensive set of evaluation metricswas computed following standard procedures for multi-label classification [20,23].

A. Hamming Accuracy

Hamming Accuracy measures the proportion of correctly predicted label-instance pairs across all samples and all classes. A value of 0.8624 indicates that 86.24% of all label assignments are correct. It is important to note that this metric can be inflated in multi-label settings when negative classes dominate the label space — a known characteristic of the ODIR-5K dataset — and should therefore be interpreted alongside F1 scores rather than in isolation [23]. Direct comparison of this figure with single-label accuracy values reported in prior work is not appropriate.

B. F1 Scores

Three F1 averaging strategies were applied [20]:

F1-micro (0.5758): Aggregates contributions of all classes, giving equal weight to each label instance. This metric is influenced by majority classes such as Normal and Diabetes.

F1-macro (0.6164): Computes F1 independently per class and averages without weight-ing, providing a more balanced view of performance across all eight disease categories including minority classes.

F1-sample (0.5715): Computes F1 per sample and averages, reflecting per-patient prediction quality.

C. Precision and Recall

Micro-averaged Precision of 0.5287 and Recall of 0.6322 indicate that the model prioritizes recall over precision. This trade-off may be relevant to screening-oriented applications, where reducing false-negative predictions can be an important consideration; however, the present study does not evaluate clinical outcomes or screening utility.

D. Specificity

Specificity of 0.8817 measures the proportion of true negative labels correctly identified across all classes. As with Hamming Accuracy, this value may be inflated in multi-label settings where negative labels dominate, and should be interpreted with this limitation in mind [20].

E. AUC-ROC

The macro-averaged AUC of 0.8382 reflects the model's discriminative ability across all eight classes. The per-class AUC scores are reported in Section 5.6 and visualized in Figure 6, which shows the ROC curves.

F. Discussion

The experimental results demonstrate that the proposed ensemble approach achieves higher validation-set performance than the individual models. The improvement in F1-macro from the individual models (0.3270–0.3758 at a fixed threshold of 0.5) to the optimized ensemble (0.6164) reflects two combined effects: architectural diversity across the four models and per-class threshold optimization tailored to the imbalanced label distribution. Because the ensemble thresholds were optimized on the same validation set used for final evaluation, this improvement should be interpreted as a validation-set result rather than as an independent estimate of generalization.

Effect of Per-Class Threshold Optimization. A fixed threshold of 0.5 proved suboptimal for minority classes such as Hypertension and Others, producing near-zero recall for these categories. By independently selecting thresholds between 0.10 and 0.49, the ensemble achieves substantially better sensitivity for underrepresented conditions [10,23]. It is acknowledged that thresholds were optimized on the same validation set used for final evaluation, which introduces a degree of optimistic bias. Future work should adopt a held-out test set for final threshold evaluation to eliminate this leakage.

Contribution of the Age-Gated HybridModel. Despite The HybridModel has lower standalone F1-macro than the strongest base CNN at a fixed threshold of 0.5, while the full ensemble has higher validation-set performance than the individual models. The present analysis does not isolate the contribution of age information, age gating, or the HybridModel because no controlled ablation experiment was included in the original run. Therefore, the current results should not be interpreted as evidence that age gating specifically improves AMD or Cataract classification. A formal ablation should compare at minimum: (i) EfficientNet-B1 without age information, (ii) EfficientNet-B1 with age concatenation but without gating, (iii) the age-gated HybridModel, and (iv) the ensemble with and without the HybridModel, using the same patient-level evaluation protocol.

Grad-CAM Clinical Alignment. The Grad-CAM visualizations provide qualitative indications of the image regions associated with selected predictions. They should not be described as confirmation of clinical alignment or validity because the current study does not compare the heatmaps with expert-annotated lesion or anatomical masks. Quantitative localization evaluation is therefore an important direction for future work [17].

Uncertainty Estimation. To estimate predictive variability, Monte Carlo Dropout (MC Dropout) was applied during inference by keeping the dropout layers active and performing 30 stochastic forward passes for each input image. For each pass, the model produced a probability vector for the eight disease classes. The mean probability across the 30 stochastic predictions was used as the final predictive probability, while the standard deviation across these predictions was used as an empirical measure of predictive variability. Samples with

a mean uncertainty value greater than 0.15 were flagged for potential specialist review. This threshold was selected empirically from the validation-set uncertainty distribution. Since no independent uncertainty calibration or uncertainty-discrimination analysis was performed, this procedure should be interpreted as an empirical variability-based screening mechanism rather than evidence of calibrated clinical uncertainty or improved clinical reliability.

Limitations. Several limitations should be acknowledged. First, the remaining class imbalance for Hypertension and Others indicates that the class weighting strategy does not fully resolve the challenge of minority-class detection. Second, the study uses a single 80/20 patient-level train/validation split rather than an independent test set; moreover, the validation set is used for both threshold optimization and final metric calculation, introducing potential optimistic bias. Third, no confidence intervals, repeated runs, or k-fold cross-validation were performed, so statistical uncertainty around the reported metrics is not quantified. Fourth, no formal ablation study was included to isolate the effects of age information, age gating, and ensemble fusion. Fifth, the uncertainty analysis does not include calibration or uncertainty-discrimination metrics, and the Grad-CAM analysis is qualitative. Finally, no permanent public code/model-weight repository was available in the current materials; a versioned repository and model archive should be provided if the authors make these resources publicly available.

Conclusion and Future Work

G. Conclusion

This study presented an ensemble deep learning framework for automated multi-label ocular disease detection using fundus images and patient age metadata. By combining EfficientNet-B1, ResNet-50, DenseNet-121, and an age-gated HybridModel, the framework achieved an F1-macro of 0.6164 and a macro AUC of 0.8382 on the ODIR-5K patient-level validation set. The age-gated mechanism provides an explicit way to condition visual features on patient age, while the ensemble combines predictions from architecturally diverse models. Per-class threshold optimization improved the reported validation metrics, but the same validation set was used for threshold selection and final evaluation; therefore, the reported performance should not be interpreted as an independent-test estimate. Grad-CAM provides qualitative visualization of prediction-related image regions, and Monte Carlo Dropout provides predictive-variability estimates, but neither component was quantitatively validated for clinical utility in the current study. A held-out patient-level test set, formal ablation, repeated or k-fold evaluation, uncertainty calibration, and quantitative explainability assessment are needed before claims about generalization or clinical deployment can be established.

H. Future Work

Several directions are identified for future investigation:

Ablation Study . A formal ablation study should compare the image-only EfficientNet-B1 baseline, age concatenation without gating, the age-gated HybridModel, and ensemble variants with and without the HybridModel. These experiments are required to quantify the individual contribution of age information, age gating, and ensemble fusion; numerical ablation results are not inserted here because they were not present in the original experimental run.

Expanded Metadata. Incorporating additional clinical features beyond patient age such as sex, refractive error, and intraocular pressure — may further improve diagnostic accuracy, particularly for conditions with strong demographic associations [12].

Advanced Architectures. Exploring Vision Transformer (ViT) backbones [26] as ensemble components may capture long-range retinal dependencies that CNN-basedarchitectures cannot model effectively.

Rigorous Validation. Adopting repeated patient-level cross-validation and/or an independent held-out test set would allow confidence intervals and more reliable estimates of generalization. In particular, thresholds should be selected using training/validation data and then fixed before evaluation on an untouched patient-level test set.

Quantitative Explainability. Quantitative validation of Grad-CAM localizations against expert-annotated lesion masks — for example using pointing-game accuracy would provide objective evidence of clinical alignment beyond qualitative heatmap visualization [17].

Larger and Multi-Center Datasets. Evaluating the framework on larger, multi-center datasets [8] would assess generalizability across diverse patient populations, imaging devices, and clinical settings.

3D Imaging Integration. Incorporating OCT scans alongside fundus images would provide volumetric retinal information, potentially improving detection of conditions with subtle 2D fundus manifestations [12].

Clinical Deployment. Developing a real-time clinical decision support system would facilitate evaluation of the framework's practical utility in ophthalmology screening workflows [9].

References

1. Larxel. Ocular Disease Recognition (ODIR-5K). Kaggle, 2019. Available online: https://www.kaggle.com/datasets/andrewmvd/ocular-disease-recognition-odir5k (accessed on 1 May 2026).
2. Dipu, N.M.; Shohan, S.A.; Salam, K.M.A. Ocular disease detection using advanced neural network

based classification algorithms. Asian J. Converg. Technol. 2021, 7, 91–99.
3. Kansal, I.; Khullar, V.; Sharma, P.; Singh, S.; Hamid, J.A.; Santhosh, A.J. Multiple model visual feature embedding and selection method for an efficient ocular disease classification. Sci. Rep. 2025, 15, 5157.
4. Khan, M.A.; Hussain, M.Z.; Khan, M.F.; Ahmad, M.; Abbas, S.; Mazhar, T.; Saeed, M.M. Blockchain enabled IoMT and transfer learning for ocular disease classification. Discov. Appl. Sci. 2025, 7, 458.
5. Kher, G.; Mehra, S.; Bala, R.; Singh, R.P. DeB5-XNet: An explainable ensemble model for ocular disease classification using feature extraction and Grad-CAM. Inform. Med. Unlocked 2025, 54, 101632.
6. Altayeb, K.; Ansari, S.; Bonny, T.; Jarndal, A. A novel explainable AI framework for multi-disease ocular classification and diabetic retinopathy severity grading. Neural Comput. Appl. 2026, 38, 251.
7. Binu, D. Comprehensive framework for ocular disease detection: utilizing Gegenbauer graph neural networks and fundus image data fusion techniques for enhanced classification of diverse ocular conditions. Biomed. Signal Process. Control 2025, 110, 108275.
8. Masum, A.K.M.; Khan, M.F.I.; Hassan, M.M.; Farid, D.M.; Bitto, A.K.; Rahman, M.A. Multi-Model Ensemble Approach for Accurate Classification of Ocular Disorders. In Proceedings of the 2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN), 2025; pp. 1–6.
9. Sah, U.K.; Chatterjee, J.M.; Sujatha, R. Multi-class eye disease classification using deep learning EfficientNetB0 fusion techniques. Sci. Rep. 2026, 16, 6368.
10. Alqassab, A.I.M.; Luque-Nieto, M.Á.; Mohammed, M.A. Identification of multiple ocular diseases using a hybrid quantum convolutional neural network with fundus images. Sci. Rep. 2026.
11. Wahab Sait, A.R. Artificial intelligence-driven eye disease classification model. Appl. Sci. 2023, 13, 11437.
12. Al-Fahdawi, S.; Al-Waisy, A.S.; Zeebaree, D.Q.; Qahwaji, R.; Natiq, H.; Mohammed, M.A.; Deveci, M. Fundus-deepnet: Multi-label deep learning classification system for enhanced detection of multiple ocular diseases through data fusion of fundus images. Inf. Fusion 2024, 102, 102059.
13. Aizaldeen Abdullah, A.; Aldhahab, A.; Al Abboodi, H.M. Review of eye diseases detection and classification using deep learning techniques. BIO Web Conf. 2024, 97, 00012.
14. Deepak, G.D.; Bhat, S.K. Deep learning-based CNN for multiclassification of ocular diseases using transfer learning. Comput. Methods Biomech. Biomed. Eng. Imaging Vis. 2024, 12, 2335959.
15. Jeny, A.A.; Junayed, M.S.; Islam, M.B. Deep neural network-based ensemble model for eye diseases detection and classification. Image Anal. Stereol. 2023, 42, 77–91.
16. Vadduri, M.; Kuppusamy, P. Enhancing ocular healthcare: deep learning-based multi-class diabetic eye disease segmentation and classification. IEEE Access 2023, 11, 137881–137898.
17. Sarki, R.; Ahmed, K.; Wang, H.; Zhang, Y.; Ma, J.; Wang, K. Image preprocessing in classification and identification of diabetic eye diseases. Data Sci. Eng. 2021, 6, 455–471.
18. Civit-Masot, J.; Luna-Perejon, F.; Vicente-Diaz, S.; Corral, J.M.R.; Civit, A. TPU cloud-based generalized U-Net for eye fundus image segmentation. IEEE Access 2019, 7, 142379–142387.
19. Civit-Masot, J.; Domínguez-Morales, M.J.; Vicente-Díaz, S.; Civit, A. Dual machine-learning system to aid glaucoma diagnosis using disc and cup feature extraction. IEEE Access 2020, 8, 127519–127529.
20. Bhati, A.; Gour, N.; Khanna, P.; Ojha, A. Discriminative kernel convolution network for multi-label ophthalmic disease detection on imbalanced fundus image dataset. Comput. Biol. Med. 2023, 153, 106519.
21. Chea, N.; Nam, Y. Classification of fundus images based on deep learning for detecting eye diseases. Comput. Mater. Contin. 2021, 67, 411–424.
22. Thanki, R. A deep neural network and machine learning approach for retinal fundus image classification. Healthc. Anal. 2023, 3, 100140.
23. Sureja, N.; Parikh, V.; Rathod, A.; Patel, P.; Patel, H.; Sureja, H. Explainable Artificial Intelligence based Deep Learning for Retinal Disease Detection. J. Electron. Electromed. Eng. Med. Inform. 2025, 7, 471–483.
24. Alqutayfi, A.; Almattar, W.; Al-Azani, S.; Khan, F.A.; Al Qahtani, A.; Alageel, S.; Alzahrani, M. Explainable Disease Classification: Exploring Grad-CAM Analysis of CNNs and ViTs. J. Adv. Inf. Technol. 2025, 16, 264–273.
25. Huo, G.; Lin, Z.; Wang, Z.; Dai, R.; Tang, H. DMS-Net: Dual-Modal Multi-Scale Siamese Network for Binocular Fundus Image Classification. arXiv 2025, arXiv:2504.18046.
26. Wang, D.; Lian, J.; Jiao, W. Multi-label classification of retinal disease via a novel vision transformer model. Front. Neurosci. 2024, 17, 1290803.
27. Wang, J.; Yang, L.; Huo, Z.; He, W.; Luo, J. Multi-label classification of fundus images with EfficientNet. IEEE Access 2020, 8, 212499–212508.
28. Sujah, A.A.; Sharfana, A.F.; Sazni, A.M. A multi-label classifier with transfer learning model for

detecting multiple ocular diseases in fundus images. Sri Lanka J. Technol. 2022, 1, 25–31.
29. Bali, A.; Mansotra, V. Transfer learning-based one versus rest classifier for multiclass multi-label ophthalmological disease prediction. Int. J. Adv. Comput. Sci. Appl. 2021, 12, 1–8.
30. Harage, S.A.; Sharma, S.; Chinchawade, A.J.; Gupta, M.K. Deep learning-based automated detection and classification of multiple ocular diseases using fundus images. SGVU Int. J. Converg. Technol. Manag. 2026, 12, 33–
31. Gal, Y.; Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In Proceedings of the 33rd International Conference on Machine Learning (ICML), 2016; pp. 1050–1059.
32. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the International Conference on Learning Representations (ICLR), 2019.