Karaelmas Science and Engineering Journal
Journal home page: https://dergipark.org.tr/tr/pub/karaelmasfen
DOI: 10.7212/karaelmasfen.1778192

Research Article
Received / Geliş tarihi : 05.09.2025
Accepted / Kabul tarihi : 31.12.2025

Automatic Detection of Eye Diseases Using Enhanced Fundus Images and Pre-trained CNN Models

Geliştirilmiş Fundus Görüntüleri ve Önceden Eğitilmiş CNN Modelleri Kullanılarak Göz Hastalıklarının Otomatik Tespiti

Sajad Abdlkadhim[1], Sait Demir[2*], Ashwan A. Abdulmunem[3]

[1]Karabük University, Institute of Graduate Programs, Department of Computer Engineering, Karabük, Türkiye
[2]Karabük University, Faculty of Computing and Informatics Science, Department of Software Engineering, Karabük, Türkiye
[3]University of Kerbala, College of Computer Science and Information Technology, Kerbala, Iraq

Abstract

This study addresses the widespread blindness issue due to late diagnosis and treatment of ophthalmic diseases by proposing an automated diagnostic framework based on retinal fundus images. The system utilizes transfer learning with pre-trained deep convolutional neural networks (CNNs) to differentiate between healthy and pathological retinal fundus images (cataract, diabetic retinopathy, glaucoma) efficiently. We propose a clear four-class baseline performance for cataracts, diabetic retinopathy (DR), glaucoma, and normal instances on top of the ODIR dataset. CLAHE preprocessing with a contrast constraint, along with morphological filtering using a light approach, was employed for preprocessing. We employed a strict 70/10/20 split for training/validation/testing purposes with fixed seed values for reproducibility Model selection was conducted exclusively by means of validated metrics, with no utilization of test data. Various state-of-the-art CNN models—ResNet50, InceptionV3, GoogleNet, and MobileNet—were fine-tuned and compared regarding feature reduction and ensemble techniques. Experimental results validate that customized models achieve remarkable accuracy rates of 99.25%, 99.1%, 95.5%, and 94.7%, respectively. Among the models investigated, ResNet50 yielded the highest performance on the test data. Performance on the training/validation dataset can be examined through multiple runs, providing mean ± standard deviation summarizations: Accuracy 99.94 ± 0.57%, Balanced Accuracy 99.69 ± 0.59%, Macro-F1 99.71 ± 0.59%, Macro ROC-AUC 0.989 ± 0.001, and Macro PR-AP 0.965 ± 0.002. Class-wise performance is captured through per-class ROC and PR curve plots with 95% bootstrap confidence bands on held-out test data, accompanied by Grad-CAM attention highlighting relevant regions, which are indicative of clinical relevance. The findings demonstrate the clinical applicability of superior fundus image-based automated screening as a dependable method for the early identification and control of numerous ophthalmic conditions

Keywords: CNN, GoogLeNet, Grad-CAM, InceptionV3, MobileNet, Ocular, ResNet50, Retinal fundus, Transfer learning.

Öz

Bu çalışma, retina fundus görüntülerine dayalı otomatik bir tanı çerçevesi önererek, oftalmik hastalıkların geç teşhisi ve tedavisi nedeniyle oluşan yaygın körlük sorununu ele almaktadır. Sistem, sağlıklı ve patolojik retina fundus görüntüleri (katarakt, diyabetik retinopati, glokom) arasında etkili bir şekilde ayrım yapmak için önceden eğitilmiş derin evrişimli sinir ağları (CNN'ler) ile transfer öğrenmesini kullanır. ODIR veri kümesinin üstünde katarakt, diyabetik retinopati (DR), glokom ve normal durumlar için net bir dört sınıf temel performansı öneriyoruz. Ön işleme için, ışık yaklaşımı kullanılarak morfolojik filtreleme ile birlikte kontrast kısıtlamalı CLAHE ön işleme kullanıldı. Tekrarlanabilirlik için sabit tohum değerleri ile eğitim/doğrulama/test amaçları için katı bir 70/10/20 bölünmesi kullandık. Model seçimi, test verileri kullanılmadan, yalnızca doğrulanmış metrikler vasıtasıyla gerçekleştirildi. ResNet50, InceptionV3, GoogleNet ve MobileNet gibi çeşitli son teknoloji CNN modelleri, özellik azaltma ve topluluk teknikleri açısından ince

*Corresponding author: saitdemir@karabuk.edu.tr

Sajad Abdlkadhim orcid.org/0000-0001-5950-9838
Sait Demir orcid.org/0000-0001-8891-4082
Ashwan A. Abdulmunem orcid.org/0000-0002-1903-9269

ayar yapılarak karşılaştırıldı. Deneysel sonuçlar, özelleştirilmiş modellerin sırasıyla %99.25, %99.1, %95,. ve %94.7 gibi dikkate değer doğruluk oranlarına ulaştığını doğrulamaktadır. İncelenen modeller arasında test verileri üzerinde ResNet50 modeli en yüksek performansı göstermiştir. Eğitim/doğrulama veri kümesindeki performans, ortalama ± standart sapma özetleri sağlayan birden fazla çalışma ile incelenebilir: Doğruluk %99.94 ± 0.57, Dengeli Doğruluk %99.69 ± 0.59, Makro-F1 %99.71 ± 0.59, Makro ROC-AUC 0.989 ± 0.001 ve Makro PR-AP 0.965 ± 0.002. Sınıf bazında performans, tutulan test verileri üzerinde %95 önyükleme güven aralıklarına sahip sınıf bazında ROC ve PR eğrisi çizimleri ile elde edilir ve bunlara klinik önemi gösteren ilgili bölgeleri vurgulayan Grad-CAM dikkati eşlik eder. Bulgular, üstün fundus görüntü tabanlı otomatik taramanın çok sayıda oftalmik rahatsızlığın erken teşhisi ve kontrolü için güvenilir bir yöntem olarak klinik uygulanabilirliğini göstermektedir.

Anahtar Kelimeler: CNN, GoogLeNet, Grad-CAM, InceptionV3, MobileNet, Ocular, ResNet50, Retinal fundus, Transfer öğrenme.

1. Introduction

Vision impairment and blindness are serious global healthcare problems, with a vast patient base due to delay in diagnosis and subsequent failure to treat ocular disorders on time (Qi et al. 2025). It is essential to screen early to prevent late-stage permanent blindness, especially for disorders such as diabetic retinopathy, cataracts, and glaucoma, which have subtle symptoms in early stages and are often missed during conventional screenings (Abazaga & Fechtner, 2023). Proven imaging technologies such as color fundus photography (CFP), OCT (optical coherence tomography), and others provide exquisite visualization of retinal morphology but cannot be interrogated by all hospitals due to expertise required for these technologies (Ramakrishnan et al., 2024).

Examples include fundus photographs, which have been adopted due to their non-invasive and cost-effective nature for retinal imaging. Nevertheless, its general clinical applications remain affected by limited access to skilled ophthalmologists, particularly in underprivileged locations (Ahn & Kim, 2024). Therefore, automated diagnostic tools have become increasingly critical to complement screening initiatives, minimize clinical workload, and improve diagnostic precision (Jeong et al., 2025).

Recent advances in deep learning, especially convolutional neural networks (CNNs), have achieved successful results in ophthalmic disease detection (Dash et al., 2024). Transfer learning with pre-trained CNN architecture has enabled significant performance gains, especially when large, annotated datasets are unavailable (Gholizade et al., 2025). Numerous studies have validated the utility of CNN-based systems in classifying specific retinal pathologies, yet many remain restricted to single-disease classification or rely on limited preprocessing techniques (Ernest et al., 2024).

Previous research on this topic has shown that deep learning is applicable for automatic disease classification of eyes. According to Ejaz et al. (2025) proposed the use of a convolutional neural network (CNN)-based method for classifying fundus images with the help of two different architectures' feature concatenation. Their ensemble-based approach exploiting the complementary power of separate CNNs provided better diagnostic performance relative to individual models. Jiang et al. (2021) proposed multi-branch convolutional networks that obtained 98.78% for spatial input features. Shamsan et al. (2023) achieved 99.23% AUC and 98.5% accuracy based on mobile and dense feature aggregation. These initial works, based primarily on data sets such as Messidor (Decencière et al., 2014) and ODIR (Kaggle, 2023), constitute a base for today's highly accurate models.

Most studies of fundus images address only a single disease. In contrast, we provide a transparent four-class screening baseline on ODIR with a 70/10/20 stratified split, single-pass evaluation on the held-out test, and per-class ROC/PR curves with 95% bootstrap CIs. We also quantify robustness via mean ± SD across repeated train/validation runs (not on the test set) and implement a pre-specified Grad-CAM protocol with uniform scaling to support clinical interpretability. These collectively position the model as a reproducible, screening-oriented baseline suitable for extension to multi-center/device validation.

2. Methodology

The introduced methodology uses a deep learning approach based on transfer learning to automate ocular disease classification from retinal fundus images. The framework, as shown in Figure 1, is structured into a workflow of five main phases:

1. *Input Acquisition:* Color Fundus Images (CFPs) are collected from the ODIR dataset (Kaggle 2023), encompassing four target classes: cataract, diabetic retinopathy, glaucoma, and normal (healthy).
2. *Preprocessing:* The images are improved using adaptive histogram equalization and morphological transformations to enhance structure visibility and contrast.

3. *Data Augmentation:* To diversify datasets as much as possible and minimize overfitting, several augmentation methodologies are invoked involving rotation, flipping, zooming, and injection of noise.
4. *Model Deployment:* Four fine-tuned pre-trained CNN models—ResNet50, InceptionV3, GoogLeNet, and MobileNet—are utilized to extract discriminative information.
5. *Classification:* The last predictions consist of SoftMax-activated dense layers that generate multi-class probabilities for a given input image.

2.1. Dataset, Preprocessing and Augmentation

The 4217 CFP images in the ODIR collection were sourced from various locations, including Ocular Recognition, the Indian Diabetic Retinopathy Image Collection (IDRiD), and High-Resolution Fundus (HRF). The collection contains over a thousand additional images of cataracts compared to diabetic retinopathy. The healthy eyes category contains 1074 images, whereas the glaucoma category contains 1007 images. Figure 1 exhibits many sample CFP images obtained from the ophthalmic dataset (Kaggle, 2023).

We chose to study the four-class fundus problem of cataract, DR, Glaucoma, and Normal scans. The data was split using a strict stratified split, with train, validation, and test set proportions of 70%, 10%, and 20%, respectively. All random functions, including shuffling, sampling, and initialization, used fixed seeds to ensure the study's repeatability. The selection of models was performed only on the validation split, with the test split remaining unseen until the final evaluation.

Effective preprocessing is essential for fundus images due to the effects caused by noises, low contrast, and uneven illumination. This work proposes a model that makes use of two prominent pre-processing approaches to deal with these issues: histogram equalization with adaptive control and morphological processing.

Each image is then processed with adaptive histogram equalization (AHE) to enhance local contrast in such a way that small micro-pathologic features such as microaneurysms, exudates, and borders of the optic disc stand out (Härtinger & Steger, 2024). Unlike global histogram equalization, this process varies the contrast on a small part of the image to enhance features that would otherwise be hard to see in images related to diabetic retinopathy and glaucoma.

Next, morphological operations such as dilation, erosion, opening, and closing are used for improving vascular and lesion structures. All these operations aim to keep the spatial integrity and magnify the contours of the retina. An example illustrating the process of improvement is shown in Figure 2 below, which depicts the improvement in visualization of contrast and clarity.

In parallel, data augmentation techniques are applied to increase the variability and volume of training data, thus reducing overfitting and improving model generalization. The augmentation pipeline includes random horizontal and vertical flipping, rotation (±15 degrees), scaling, Gaussian noise injection, and zoom transformations (Wang et al., 2024). These operations are applied probabilistically during training to ensure a diverse and robust input distribution. All preprocessed and augmented images are resized to a standard input dimension (e.g., 224×224) to ensure compatibility with the CNN architectures.

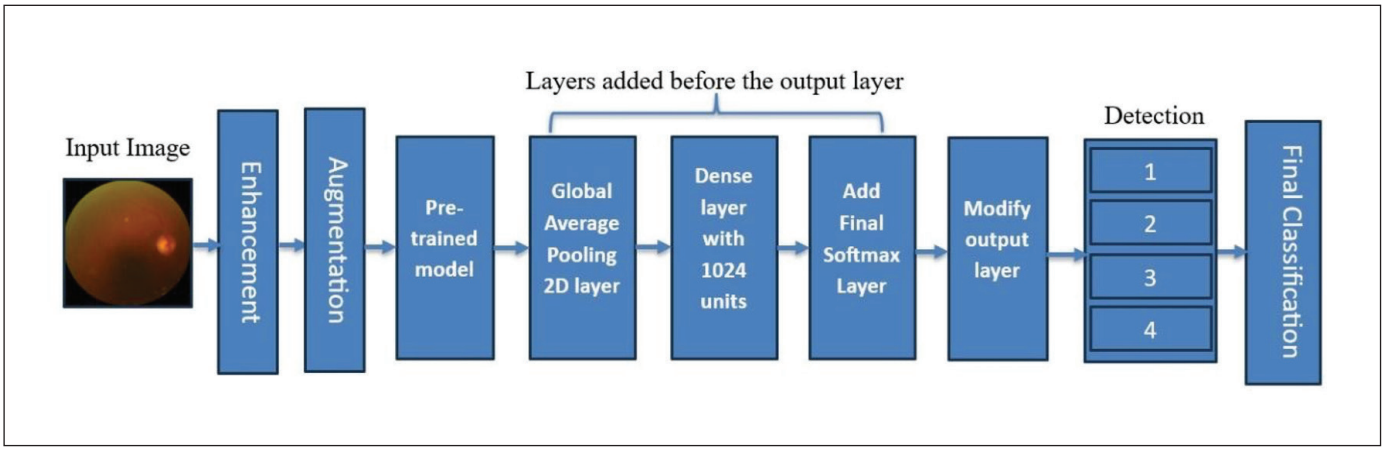

Figure 1. Schematic overview of the proposed deep learning pipeline for multi-disease classification in retinal fundus images.

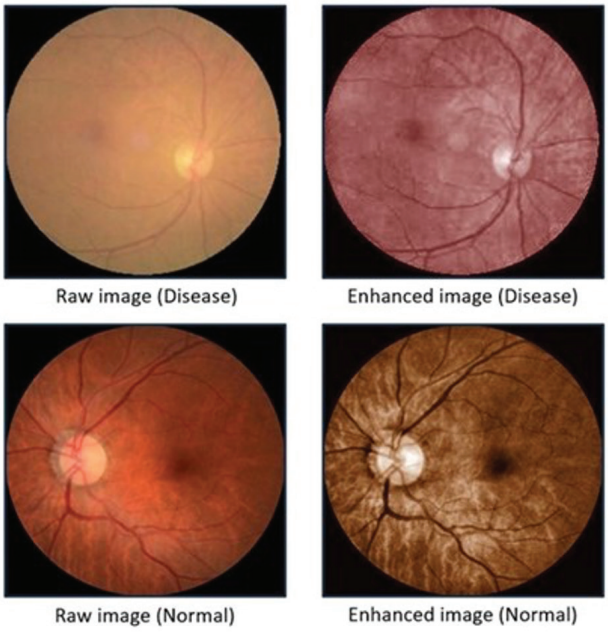

Figure 2. Visual comparison of original versus enhanced fundus images after applying adaptive histogram equalization and morphological preprocessing.

2.2. Model Architecture and Transfer Learning

This work uses a transfer learning strategy with four state-of-the-art CNN architectures: ResNet50, InceptionV3, GoogLeNet, and MobileNet. The CNNs are initialized with ImageNet-pre-trained weights and fine-tuned for multi-class ocular disease classification. The choice of fine-tuning was driven by the domain gap between natural images and retinal fundus images. The modified models are trained using the stochastic gradient descent with momentum optimizer, with the best single performance obtained using the fine-tuned ResNet50 architecture.

2.3. Training Strategy and Hyperparameters

Supervised learning was adopted for training purposes, in which the fundus images that were preprocessed and augmented were combined with their respective diagnostic labels for four classes. All of the models were created with *MATLAB's Deep Learning Toolbox* and trained on a computer that equipped with a High-Performance GPU for fast convergence. To ensure reproducibility, we fixed all random seeds for operations like data shuffling, augmentation sampling, and weight initialization, and logged the complete configuration along with model checkpoints.

2.3.1. Protocol and Model Selection

The dataset was split using a strict split strategy, with 70% allocated for training, 10% for validation, and 20% for testing. The hyperparameters for tuning and the stopping criteria for the model's training process relied solely on the metrics of the validation set. The test dataset was left unchanged until it was used for evaluation.

We selected the best model checkpoint based on the highest score on the macro F1 measure on the validation data, ensuring a balanced sensitivity for all classes. We repeated the experiments on 5 stratified folds/runs, using only the training and validation datasets. The summary statistics for all experiments are reported in mean and standard deviation. The test data is evaluated only once.

2.3.2. Optimizer

We used Stochastic Gradient Descent with Momentum (SGDM) to optimize model weights because it is known for robust fine-tuning performance on mid-sized datasets for medical imaging. The loss function used is categorical cross-entropy, which is defined over four categories. We employed a weighted categorical cross-entropy loss function with slight oversampling on the minority categories. Resampling is not done on test datasets because it is not part of best practices.

2.3.3. Hyperparameters

Unless specified, all these parameters for the best model (ResNet50) will be used on all other backbone models:

- Initial learning rate: 0.001
- Learning rate schedule: Step decay by a factor of 0.1 if validation loss plateaus, with a minimum learning rate of 1e-6
- Momentum / weight decay: 0.9 / 1e-4
- Mini batch size: 64 (reduced to 32 if memory is a concern on GPUs).
- Epochs: At most 100, with early stopping if there is no improvement on validation loss for 10 epochs.
- Model selection: Due to highest validation Macro-F1.
- Data shuffling: Done every epoch to prevent ordered-input bias.
- Mixed precision: Allowed (if available on the GPU) to speed up training without sacrificing validation accuracy.

2.3.4. Regularization and Augmentation Policy

To improve generalization, we applied a two-pronged strategy:

(i) The class-weighted loss and oversampling approach as noted above, and

(ii) A conservative augmentation pipeline applied only to training images, including random rotations, horizontal flips, and mild variations in zoom, brightness, and contrast. No augmentations or test-time tricks were applied to validation or test images to maintain a clean hold-out protocol.

We applied class-weighted cross-entropy and mild oversampling of minority classes during training only; no resampling or TTA was applied on validation or test.

2.3.5. Monitoring and Convergence

Training and validation loss and accuracy were tracked at each epoch. Across runs, the training curves demonstrated stable convergence with no significant divergence between training and validation—validating the use of early stopping and the effectiveness of our regularization strategies. The final checkpoints selected from each run were later used to generate per-class ROC and PR curves, along with confidence intervals on the held-out test set, as required by reviewers.

2.4. Evaluation Metrics

The main metrics are Accuracy, Macro-F1, and Macro-AUC. We also provide per-class ROC and PR curves on the test set, along with 95% bootstrap confidence intervals based on predicted probabilities (Johnson & Khoshgoftaar, 2020). Unless otherwise specified, all summary statistics are aggregated over 5 runs on the training/validation splits only. The test set is evaluated just once, using the best-performing checkpoint on the validation set, to maintain the integrity of the hold-out protocol. While binary class-specific scores can be reported—defined explicitly as One-vs-Rest (OvR) tasks per class—we encourage the use of per-class Precision (Cabot & Ross, 2023), Recall (Tatbul et al., 2018), F1 (Diallo et al., 2025), Specificity, and Balanced Accuracy instead (Miller, 1975). These provide a more comprehensive assessment than relying on binary accuracy alone, and all are computed using test set predictions.

2.5. Explainability

We created Grad-CAM maps from the final convolutional block of the top-performing model (ResNet50). For each disease category, we selected representative examples of both correct and incorrect predictions—specifically, the top-k most confident correct cases and the highest-confidence errors among the misclassified ones. Selection criteria were pre-defined and applied consistently to prevent any post-hoc cherry-picking.

The heatmaps were overlaid on the original images using uniform color scaling across all samples. Two clinicians independently reviewed a subset of these maps to confirm that the highlighted regions correspond to disease-relevant anatomical features, such as the optic disc/cup for glaucoma and microaneurysms or exudates for diabetic retinopathy.

3. Results and Discussion

3.1. Overall Test Performance

We deal with a four-class classification problem in this competition involving the ODIR dataset (cataract, diabetic retinopathies, glaucoma, normal). All images have been pre-processed and augmented as mentioned in section 2. We have conducted model training/validation on a stratified split with fixed seeds, without any access to test data.

ResNet50 showed the best performance on all metrics on the test data: Accuracy = 99.25%, AUC = 0.9991, F1 Score = 99.89%, Precision = 99.80%, Recall = 100% (Table 1). The performances of other pre-trained models are shown for completeness. They are all unchanged.

Out of all investigated CNN architectures, it was observed that ResNet50 outperformed all other networks with an average accuracy of 99.25%, an AUC value of 0.9991, and a recall of 100%. InceptionV3 also demonstrated satisfactory performance with an accuracy of 99.1% along with 100% precision and recall rates mostly in cataract and diabetic retinopathy.

We provide a qualitative comparison of all four models on all performance measures in Figure 3. Although GoogLeNet and MobileNet produced comparable results, there was a slight reduction in F1-score and recall for some classes notably glaucoma and normal.

Confusion matrices for all models were created to get a clearer insight into class-structured behavior. Figures 3 and 4 show classification results for each structure. ResNet50 achieved flawless cataract and diabetic retinopathy classification with minor misclassifications for glaucoma (98.7%) and normal (99.3%) classes. On the other hand, GoogLeNet and MobileNet demonstrated more variability with significant drops in normal class identification.

Table 1. Performance evaluation of the proposed CNN models on the test set using AUC, accuracy, F1 score, precision, and recall.

Tools	Accuracy	AUC	F1 Score	Precision	Recall
ResNet50	**99.25**	**0.9991**	**99.89**	**99.80**	**100**
Inceptionv3	99.1	0.996	100	100	100
GoogLeNet	95.3	0.992	92.2	95.2	89.4
MobileNet	94.7	0.994	92.4	93.4	91.5

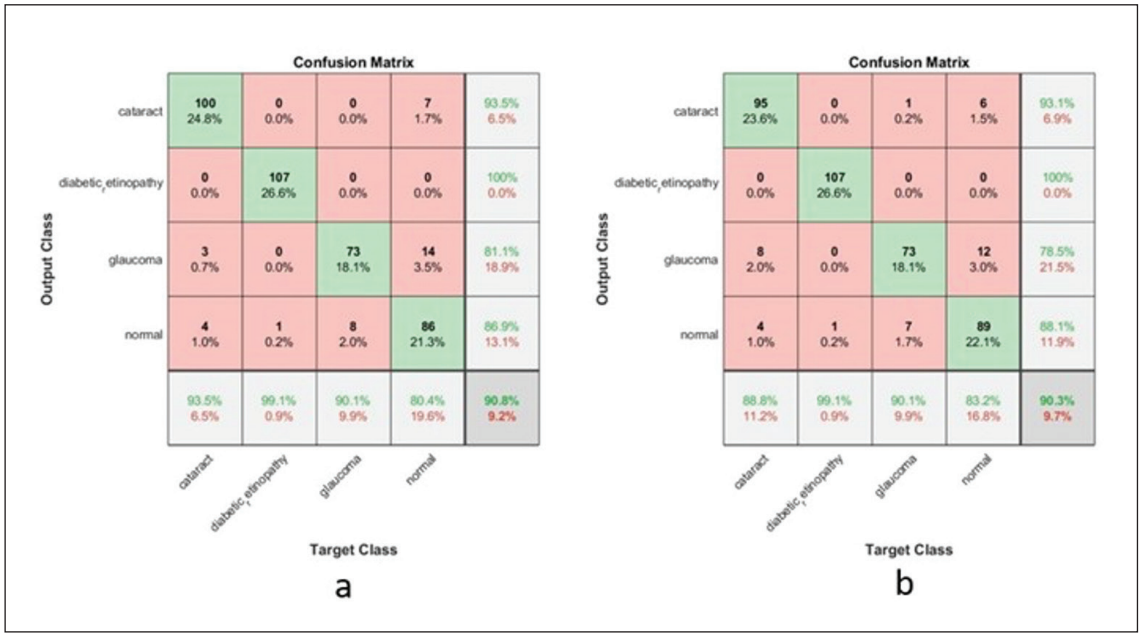

Figure 3. Measuring implementation confusion matrix of pre-trained CNNs models a) ResNet50 b) InceptionV3.

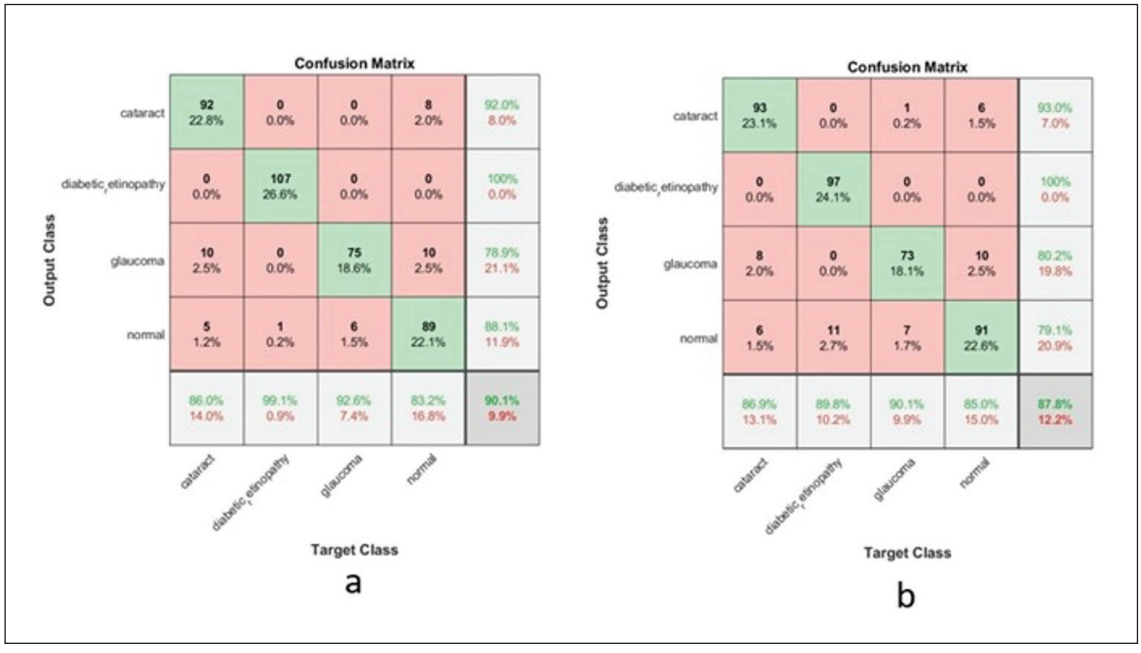

Figure 4. Measuring implementation confusion matrix of pre-trained CNN models
a) GoogLeNet
b) MobileNet.

3.2. Class-Wise Behavior

Table 2 summarizes the classification accuracy of each model for individual disease classes. ResNet50 and InceptionV3 gave the best results in all categories.

3.3. Discrimination and Learning Dynamics (ResNet50)

Per-class ROC plots with 95% confidence bounds on test performance are presented in Figure 5, with PR-curve plots given in Figure 6. All classes demonstrate high separability with a significant gap between glaucoma and other classes, although it is now smallest for glaucoma but remains in acceptable clinical limits.

3.4. Robustness Summary (ResNet50)

During multiple runs for training/validation, ResNet50 showed consistent macro metrics: Accuracy 99.94% ± 0.57, Balanced Accuracy 99.69% ± 0.59, Macro-F1 99.71% ±

Table 2. Class-wise performance comparison (cataract, diabetic retinopathy, glaucoma, and normal) for each CNN model, based on classification accuracy.

Techniques	Cataract	Diabetic retinopathy	Glaucoma	Normal
ResNet50	99%	100%	98.7%	99.3%
Inception V3	100%	100%	97.7%	98.7%
GoogLeNet	93.8%	100%	97.2%	90.6%
MobileNet	89.6%	100%	100%	90.6%

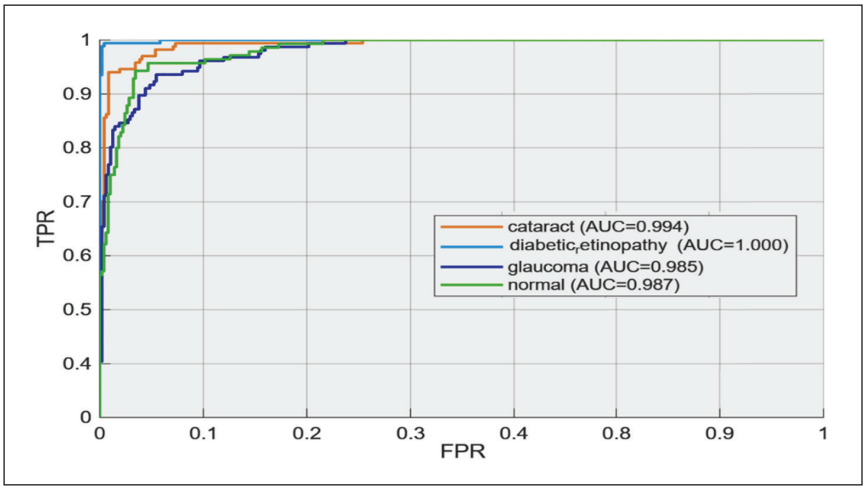

Figure 5. ResNet50: Per-class ROC-curves on test data, shaded regions represent 95% bootstrapped confidence intervals.

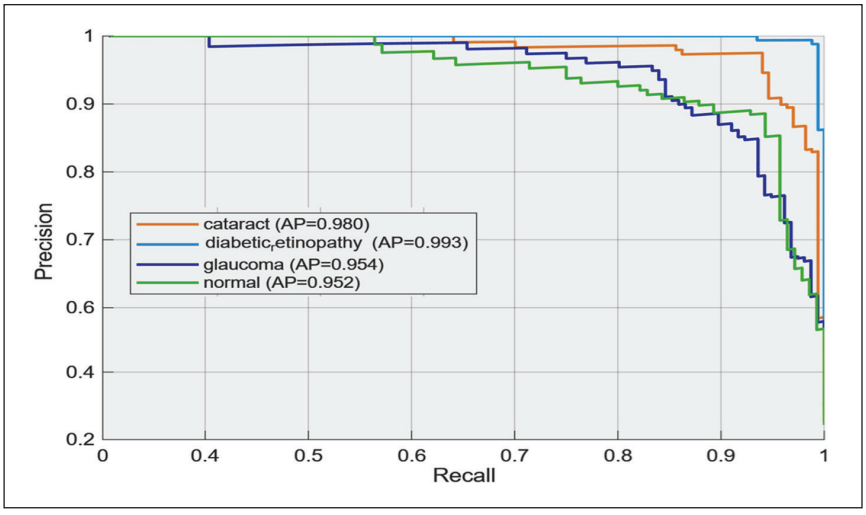

Figure 6. ResNet50: Per-class PR curves on held-out test set; shaded bands: 95% bootstrap CIs.

0.59, Macro Precision 99.75% ± 0.59, Macro Recall 99.69% ± 0.59, Macro ROC-AUC 0.989 ± 0.001, Macro PR-AP 0.965 ± 0.002. These metrics are a measure of variation across multiple runs for training/validation, while the held-out test set is evaluated once using the validation-selected checkpoint to preserve a clean hold-out protocol. Unless noted otherwise, mean ± SD summaries are computed across repeated runs on the training/validation portion only; the held-out test set is evaluated once using the validation-selected checkpoint.

3.5. Qualitative Explainability (ResNet50)

The Grad-CAM visualization highlights important structures for each disease, such as retinal disc/cup for Glaucoma and microaneurysms and exudates for DR. Grad-CAM visualization for correct and incorrect test samples for each class is shown in Figure 7, using a fixed color scale with a pre-defined selection policy.

3.6. Positioning in Current Literature

Table 3 places current work in the framework of recent multi-disease fundus analyses involving collections ranging from ODIR, RFMiD, Kaggle fundus images, through to ultra-widefield datasets. Notably, current proposed benchmarking permeates the highest reported accuracy levels on datasets such as ODIR but is mindful of clear communication (strict split, class-by-class evaluation on ROC/PR with CI, and model explanations).

4. Limitations

Although the proposed baseline demonstrated excellent discrimination on ODIR, several areas remain for improvement. Firstly, this assessment is limited to one public data source, gathered in a variety of environments, and thus does not cover actual out-of-distribution robustness. Secondly, only one model, ResNet50, has been examined with multiple assessments. The rest have only been succinctly presented due to space reasons. They could have otherwise been used for comparison with other submissions. Thirdly, our pre-processing is generic and non-lesion-oriented. Future work could be oriented towards pre-processing strategies such as lesion-guided enhancements. Finally, decision thresholds have been chosen for class balance. However, cost-sensitive decision-making strategies relevant to screening operations (such as maximizing sensitivity for disease categories) have many untapped areas for improvement.

5. Clinical Relevance and Deployment Considerations

The operating points employed in this work put greater emphasis on sensitivity towards pathological images with acceptable levels of precision on par with referral practices. Thus, this deep model can serve as a pre-screening triage system for directing images indicative of cataract/DR/glaucoma towards expert evaluation. Validation related to improved reading times and inter-grader agreements with and without AI aid is required.

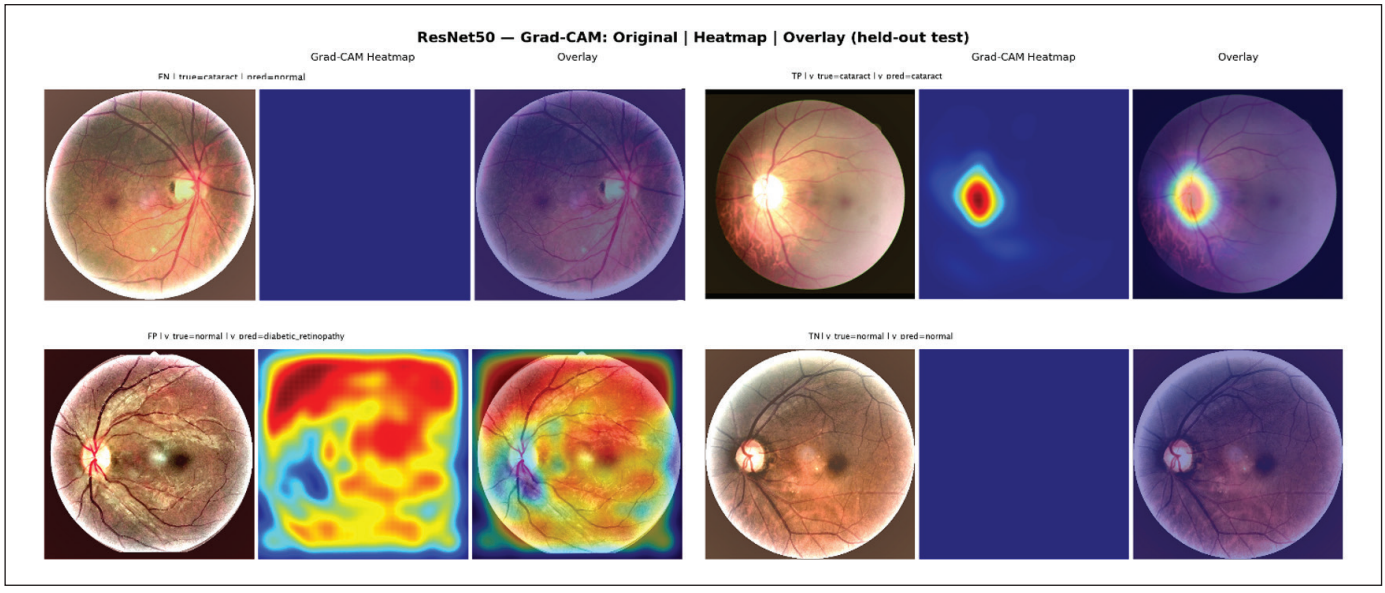

Figure 7. ResNet50-Grad-CAM attention weights for indicative correct (in columns 1–2) and incorrect (in columns 3–4) predictions for all four categories. Uniform color scaling. Relevant areas are pointed out using arrows.

Table 3. Comparative positioning against recent fundus-based studies.

Ref. (Year)	Dataset / #Classes	Method / Backbone	Split / External Validation	Reported Metrics	XAI
This study (2025)	ODIR (4 classes)	ResNet50 TL + preprocessing (AHE + morphology)	Train/Val/Test	Acc ≈ 99.25, AUC ≈ 0.999	Grad-CAM
Yu & Dong (2025)	MuReD + RFMiD; also reports on MESSIDOR2, EyePACS-1	Dual-branch (RGB fundus + vessel segmentation) + TL + soft-voting (SVM/MLP/XGB)	Train/Test; also reports performance on external datasets	DR task Acc 99.2%; multi-disease Acc 98.8%	Grad-CAM
Alsohemi & Dardouri (2025)	Kaggle fundus (4 classes: DR/Glaucoma/Cataract/Healthy)	EfficientNet-B3 (fine-tuned; Adam + cosine scheduler)	Internal split	Acc 95.12%; Prec 95.21%; Rec 94.88%; F1 95.00%; MCC 0.925	Not reported
Kansal et al. (2025)	ODIR (8 classes)	DenseNet201/EfficientNetB3/Inception-ResNet-V2 features → LDA → DNN/LSTM/BiLSTM	Train/Val/Test	Validation >98% (BiLSTM best)	Not reported
Al-Fahdawi et al. (2024)	ODIR (multi-label)	Fundus-DeepNet (feature/data fusion for multi-label)	Internal split	Reports multi-label F1/AUC on ODIR	Mentions interpretability
Chavan & Pete (2024)	RFMiD (multi-disease)	MGSCNN + Glowworm Swarm Optimization (hyper-param search)	Internal split	Acc 95.09% (with additional Sens/Spec)	Not reported
Ejaz et al. (2024)	RFMiD (multi-disease)	CNN (12- & 20-layer variants)	Train/Val/Test	Val Acc up to 89.81%; Test Acc 88.72–89.59%	Brief discussion
Duan & Tu (2025)	UWF fundus (16 diseases)	DenseNet121 features + XGBoost	External test set reported	Common diseases: AUC >0.975, Acc >0.98; rare diseases: high AUC/Acc reported	Grad-CAM

6. Conclusion and Suggestions

The research introduced a highly accurate deep-learning system for automated multi-disease screening from retinal fundus images with fine-tuned CNN models. The proposed model demonstrated state-of-the-art accuracy with strong robustness for cataract, diabetic retinopathy, glaucoma, and normal cases. These results identify the clinical potential of deep learning for early and precise ocular disease screening. Future research will aim to test this with larger, multi-centre datasets to expand its real-world utility.

Acknowledgements: This article is produced from the master's thesis of SAJAD Abdlkadhim Abdlhusein Alkhaykane titled *"A Modified ResNet-50 CNN Model for Classification of Eye Diseases"*, completed in 2023 at Karabük University under the primary supervision of Sait Demir and co-supervision of Ashwan A. Abdulmunem.

Author contributions: All authors contributed equally to the conception, literature review, writing, and revision of the manuscript. All authors have read and approved the final version.

Ethics approval and consent to participate: Not applicable.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of interest: The authors certify that they have no conflict of interest.

References

Abazaga, M., & Fechtner, R. (2023). Changes and diseases of the aging eye. In M. R. Wasserman, D. Bakerjian, S. Linnebur, S. Brangman, M. Cesari, & S. Rosen (Eds.), Geriatric medicine (pp. 663–689). Springer. https://doi.org/10.1007/978-3-030-01782-8_58-1

Ahn, S. J., & Kim, Y. H. (2024). Clinical applications and future directions of smartphone fundus imaging. Diagnostics, 14(13), 1395. https://doi.org/10.3390/diagnostics14131395

Al-Fahdawi, S., Al-Waisy, A. S., Zeebaree, D. Q., Qahwaji, R., Natiq, H., Mohammed, M. A., Nedoma, J., Martinek, R., & Deveci, M. (2024). Fundus-deepnet: Multi-label deep learning classification system for enhanced detection of multiple ocular diseases through data fusion of fundus images. Information Fusion, 102, 102059. https://doi.org/10.1016/j.inffus.2023.102059

Alsohemi, R., & Dardouri, S. (2025). Fundus image-based eye disease detection using EfficientNetB3 architecture. Journal of Imaging, 11(8), 279. https://doi.org/10.3390/jimaging11080279

Cabot, J. H., & Ross, E. G. (2023). Evaluating prediction model performance. Surgery, 174(3), 723–726. https://doi.org/10.1016/j.surg.2023.05.023

Chavan, R., & Pete, D. (2024). Automatic multi-disease classification on retinal images using multilevel glowworm swarm convolutional neural network. Journal of Engineering and Applied Science, 71,26. https://doi.org/10.1186/s44147-023-00335-0

Dash, S. K., Sethy, P. K., Das, A., Jena, S., & Nanthaamornphong, A. (2024). Advancements in deep learning for automated diagnosis of ophthalmic diseases: A comprehensive review. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3496565

Decencière, E., Zhang, X., Cazuguel, G., Lay, B., Cochener, B., Trone, C., Gain, P., Ordóñez-Varela, J. R., Massin, P., & Erginay, A. (2014). Feedback on a publicly distributed image database: The Messidor database. Image Analysis & Stereology, 33(3), 231–234. https://doi.org/10.5566/ias.1155

Diallo, R., Edalo, C., Awe, O.O. (2025). Machine learning evaluation of imbalanced health data: A comparative analysis of balanced accuracy, MCC, and F1 Score. In: Awe, O.O., A. Vance, E. (eds) Practical Statistical Learning and Data Science Methods. STEAM-H: Science, Technology, Engineering, Agriculture, Mathematics & Health. (pp. 283–312). Springer,. https://doi.org/10.1007/978-3-031-72215-8_12

Duan, M. M., & Tu, X. (2025). Deep learning-based classification of multiple fundus diseases using ultra-widefield images. Frontiers in Cell and Developmental Biology, 13, 1630667. https://doi: 10.3389/fcell.2025.1630667

Ejaz, S., Baig, R., Ashraf, Z., Alnfiai, M. M., Alnahari, M. M., & Alotaibi, R. M. (2024). A deep learning framework for the early detection of multi-retinal diseases. PLOS ONE, 19(7), e0307317. https://doi.org/10.1371/journal.pone.0307317

Ejaz, S., Zia, H. U., Majeed, F., Shafique, U., Altamiranda, S. C., Lipari, V., & Ashraf, I. (2025). Fundus image classification using feature concatenation for early diagnosis of retinal disease. Digital Health, 11. https://doi.org/10.1177/20552076251328120

Ernest, M., Godakanda, S., Chandrasiri, S., & Panduwawala, P. (2024). Diabetic retinal disease detection through transfer learning techniques. In Proceedings of the 2024 6th International Conference on Advancements in Computing (ICAC) (pp. 312–317). IEEE. https://doi.org/10.1109/ICAC64487.2024.10851057

Gholizade, M., Soltanizadeh, H., Rahmanimanesh, M., & Sana, S. S. (2025). A review of recent advances and strategies in transfer learning. International Journal of System Assurance Engineering and Management, 16, 1123–1162. https://doi.org/10.1007/s13198-024-02684-2

Härtinger, P., & Steger, C. (2024). Adaptive histogram equalization in constant time. Journal of Real-Time Image Processing, 21(3), 93. https://doi.org/10.1007/s11554-024-01465-1

Jeong, J., Kim, S., Pan, L., Hwang, D., Kim, D., Choi, J., Kwon, Y., Yi, P., Jeong, J., & Yoo, S. J. (2025). Reducing the workload of medical diagnosis through artificial intelligence: A narrative review. Medicine, 104(6), e41470. https://doi.org/10.1097/MD.0000000000041470

Jiang, Y., Liu, W., Wu, C., & Yao, H. (2021). Multi-scale and multi-branch convolutional neural network for retinal image segmentation. Symmetry, 13(3), 365. https://doi.org/10.3390/sym13030365

Johnson, J. M., & Khoshgoftaar, T. M. (2020). Thresholding strategies for deep learning with highly imbalanced big data. In M. A. Wani, T. M. Khoshgoftaar, & V. Palade (Eds.), Deep learning applications, Advances in Intelligent Systems and Computing (Vol. 2, pp. 199–227). Springer. https://doi.org/10.1007/978-981-15-6759-9_9

Kaggle. (2023). Eye diseases classification dataset. https://www.kaggle.com/datasets/gunavenkatdoddi/eye-diseases-classification

Kansal, I., Khullar, V., Sharma, P., Singh, S., Hamid, J. A., & Santhosh, A. J. (2025). Multiple model visual feature embedding and selection method for an efficient ocular disease classification. Scientific Reports, 15, 5157. https://doi.org/10.1038/s41598-024-84922-y

Miller, D. (1975). The accuracy of predictions. Synthese, 30(1–2), 159–191.

Qi, T., Liu, H., Frühn, L., Löw, K., Cursiefen, C., & Prokosch, V. (2025). Understanding glaucoma: Why it remains a leading cause of blindness worldwide. Klinische Monatsblätter für Augenheilkunde, 242(7), 712–717. https://doi.org/10.1055/a-2617-1575

Ramakrishnan, M. S., Kovach, J. L., Wykoff, C. C., Berrocal, A. M., & Modi, Y. S. (2024). American Society of Retina Specialists clinical practice guidelines on multimodal imaging for retinal disease. Journal of VitreoRetinal Diseases, 8(3), 234–246. https://doi.org/10.1177/24741264241237012

Shamsan, A., Senan, E. M., & Shatnawi, H. S. A. (2023). Automatic classification of colour fundus images for prediction eye disease types based on hybrid features. Diagnostics, 13(10), 1706. https://doi.org/10.3390/diagnostics13101706

Tatbul, N., Lee, T. J., Zdonik, S., Alam, M., & Gottschlich, J. (2018). Precision and recall for time series. In Advances in Neural Information Processing Systems (Vol. 31).

Wang, Z., Wang, P., Liu, K., Wang, P., Fu, Y., Lu, C. T., Aggarwal, C. C., Pei, J., & Zhou, Y. (2024). A comprehensive survey on data augmentation. (preprint) arXiv. https://doi.org/10.48550/arXiv.2405.09591

Yu, H., & Dong, X. (2025). Ensemble-based eye disease detection system utilizing fundus and vascular structures. Scientific Reports, 15, 19298. https://doi.org/10.1038/s41598-025-04503-5