AUIQ Technical Engineering Science

Manuscript 1064

A Hybrid Deep Learning Framework for Multi-Class Retinal Disease Classification

Ali M. Duhaim

Abbas M. Al-Bakry

AUIQ TECHNICAL ENGINEERING SCIENCE 2026;3:46–56

Scan the QR to view the full-text article on the journal website

ORIGINAL STUDY

A Hybrid Deep Learning Framework for Multi-Class Retinal Disease Classification

Ali M. Duhaim [a,*], **Abbas M. Al-Bakry** [b]

[a] Informatics Institute for Postgraduate Studies, University of Information Technology and Communication (UoITC), Baghdad, Iraq
[b] University of Information Technology and Communication (UoITC), Baghdad, Iraq

ABSTRACT

Ocular diseases such as diabetic retinopathy, glaucoma, and cataract remain major causes of vision loss worldwide, creating an increasing demand for accurate and reliable automated diagnostic systems. Although deep learning techniques have demonstrated strong performance in retinal image analysis, their clinical adoption is still limited by challenges related to predictive reliability, robustness, and interpretability. This study presents HEFR-Net (Hybrid EfficientNet-B3 and ResNet-50 Feature-Fusion Network), a hybrid deep learning framework designed for multi-class classification of retinal fundus images. The proposed architecture integrates EfficientNet-B3 and ResNet-50 through feature-level fusion, enabling the model to capture both fine-grained texture patterns and high-level structural characteristics associated with retinal diseases. To enhance predictive reliability, label smoothing is incorporated as a calibration-oriented regularization strategy, reducing overconfidence in model predictions. In addition, Grad-CAM is employed to provide visual explanations by highlighting clinically relevant regions that contribute to the model's decisions. The framework is evaluated on a publicly available dataset consisting of 4,217 retinal fundus images across four diagnostic categories: Normal, Cataract, Glaucoma, and Diabetic Retinopathy. Experimental results show that HEFR-Net achieves an accuracy of 95.50%, indicating improved performance compared to conventional single-model and standard ensemble approaches. Further analysis demonstrates consistent performance across classes, along with interpretable activation patterns aligned with clinically meaningful features. Overall, the proposed framework provides an accurate, reliable, and interpretable approach for automated retinal disease classification, supporting its potential application in computer-aided diagnosis and clinical decision support systems.

Keywords: Trustworthy AI, Medical image classification, Hybrid deep learning, Feature fusion, EfficientNet-B3, ResNet-50

1. Introduction

Visual impairment remains a major global health challenge, with large-scale epidemiological studies indicating that hundreds of millions of individuals suffer from moderate to severe vision loss or blindness, the majority of which could be prevented through early diagnosis and timely intervention [1]. Among the leading causes of vision impairment are cataract, glaucoma, and diabetic retinopathy (DR), all of which significantly contribute to irreversible blindness, particularly among aging populations and individuals with chronic metabolic disorders such as diabetes mellitus [2]. Glaucoma, often referred to as the "silent thief of sight," is especially critical due to its progressive and asymptomatic nature, leading to irreversible optic nerve damage if not detected early [3]. Traditionally, retinal disease diagnosis relies on manual examination of fundus images by trained ophthalmologists; however, this process is time-consuming, subjective, and prone to inter-observer variability, particularly in resource-limited settings where expert availability is constrained [4, 5].

Recent advances in deep learning (DL), particularly convolutional neural networks (CNNs), have

Received 14 February 2026; revised 6 May 2026; accepted 6 May 2026.
Available online 25 May 2026

* Corresponding author.
E-mail address: alialloshy@gmail.com (A. M. Duhaim).

https://doi.org/10.70645/3078-3437.1064

significantly transformed medical image analysis by enabling automatic feature extraction directly from raw pixel data [6]. Early architectures such as LeNet and AlexNet laid the foundation for visual recognition tasks, while more advanced models, including VGG, Inception, and ResNet, have demonstrated superior performance in ophthalmic image analysis [7]. Building upon these developments, automated retinal disease classification has rapidly evolved, primarily leveraging transfer learning with pretrained architectures. Recent studies have reported multi-class classification accuracies ranging from 87% to over 94% using models such as ResNet [8–11], EfficientNet [12, 13], and custom-designed CNN architectures [14]. Furthermore, hybrid approaches combining deep learning representations with classical machine learning classifiers [15], as well as explainability techniques such as LIME, have been explored to enhance interpretability in clinical contexts [16]. In addition to these advancements, recent studies have further explored diverse CNN-based and hybrid strategies for retinal disease classification. Comparative analyses of multiple CNN architectures have demonstrated that models such as ResNet-152, EfficientNet, MobileNet, and DenseNet can achieve high diagnostic performance, with ResNet-152 reaching an accuracy of 94.82%, highlighting the importance of architecture design and preprocessing techniques such as CLAHE [24]. Moreover, semi-supervised learning approaches have been introduced to address the challenge of limited labeled medical data, where a dual-branch framework (DB-SSL) achieved an accuracy of 93.14% by effectively leveraging both labeled and unlabeled data, improving robustness and generalization in multi-class classification tasks [25]. In parallel, multi-model CNN approaches have shown that combining different architectures enhances feature representation by capturing diverse visual patterns, leading to improved performance compared to single-model methods [23]. Despite these advancements, single-architecture models inherently suffer from several limitations. Deep residual networks excel in capturing high-level semantic structures but may overlook fine-grained textural abnormalities critical for early-stage DR detection, whereas lightweight models often sacrifice representational depth for computational efficiency. To address these trade-offs, recent research has explored hybrid and ensemble strategies; however, prediction-level fusion and simple model averaging often fail to fully exploit complementary feature representations, resulting in suboptimal robustness and generalization [17]. Consequently, feature-level fusion approaches that integrate complementary architectures, such as EfficientNet for fine-grained texture representation and ResNet for structural feature extraction, have emerged as a promising direction for improving diagnostic performance [18].

In parallel, the clinical deployment of deep learning systems remains constrained by challenges related to reliability, robustness, and interpretability—key principles of Trustworthy AI [19]. In particular, predictive calibration has received limited attention in existing studies, despite its critical role in ensuring that model confidence aligns with actual predictive accuracy. Models trained using conventional cross-entropy loss with hard labels are prone to overconfidence, often producing misleadingly high probability estimates in uncertain or ambiguous cases [20]. Such behavior can undermine clinical trust and pose risks in decision-support systems. Label smoothing has been proposed as an effective regularization technique to mitigate overconfidence by softening target distributions, thereby improving generalization and calibration performance [21]. However, the integration of calibration mechanisms with hybrid feature-learning frameworks remains largely unexplored.

To address these gaps, this study proposes a unified Trustworthy AI framework for retinal disease classification. This study introduces HEFR-Net (Hybrid EfficientNet-B3 and ResNet-50 Feature-Fusion Network), which integrates feature-level fusion, calibration-aware training, and explainability within a single end-to-end architecture. Unlike existing approaches that treat these components independently, the proposed framework performs bottleneck-level feature fusion between EfficientNet-B3 and ResNet-50 to jointly capture fine-grained textural features and high-level structural representations. In addition, label smoothing is incorporated to enhance predictive calibration, while Grad-CAM is employed to provide clinically meaningful visual explanations. This integrated design aims to improve robustness, reliability, and interpretability, particularly in challenging diagnostic scenarios such as glaucoma detection. The main contributions of this work can be summarized as follows:

1. **Hybrid Feature-Fusion Architecture:** HEFR-Net is proposed as a dual-stream CNN framework that integrates EfficientNet-B3 and ResNet-50 through feature-level fusion to capture complementary textural and structural representations of retinal images.
2. **Calibration-Oriented Training Strategy:** The training process incorporates label smoothing and class weighting to reduce overconfidence, improve predictive calibration, and address class imbalance.

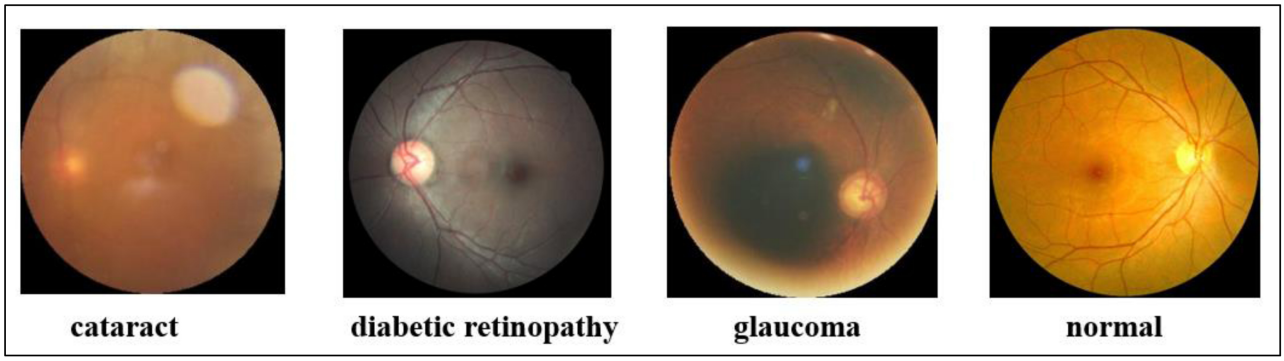

Fig. 1. Sample retinal fundus images for the four diagnostic classes.

3. **Enhanced Classification Performance:** The proposed model achieves a testing accuracy of 95.50% on a publicly available dataset of 4,217 retinal fundus images, outperforming conventional single-model and standard ensemble approaches.
4. **Explainable and Reproducible Framework:** Grad-CAM-based visual explanations and detailed methodological design are provided to ensure interpretability, transparency, and reproducibility.

The remainder of this paper is organized as follows. Section 2 describes the proposed methodology, including the dataset, preprocessing pipeline, hybrid architecture, and training strategy. Section 3 presents and discusses the experimental results, including quantitative evaluation, confusion-matrix-based error interpretation, explainability, ablation analysis, comparative evaluation, and reproducibility considerations. Section 4 concludes the paper and outlines future research directions.

2. Methodology

In this study, retinal diseases are not modeled using a physiological or anatomical simulation of the human eye. Instead, a phenotypic image-based approach is adopted, where each class is defined by its visual manifestation in color fundus images. Cataract is associated with diffuse lens opacity and reduced image clarity, glaucoma with structural changes such as optic disc cupping, and diabetic retinopathy (DR) with vascular abnormalities including microaneurysms, hemorrhages, and exudates. The dataset used in this work is the publicly available Kaggle Eye Disease Classification Dataset [22], which contains 4,217 images distributed across four classes: Normal, Cataract, Glaucoma, and DR. To illustrate the visual variability and diagnostic characteristics of these categories, representative samples are presented in Fig. 1.

Table 1. Dataset distribution and stratified splits.

Class	Total Images	Training	Validation	Testing
Normal	1,074	866	100	108
Cataract	1,038	843	100	95
Glaucoma	1,007	816	100	91
Diabetic Retinopathy	1,098	892	100	106
Total	**4,217**	**3,417**	**400**	**400**

The dataset was divided using a stratified split into training (3,417 images), validation (400 images), and testing (400 images) subsets to ensure balanced class distribution and prevent data leakage. A fixed train–validation–test split was adopted to provide a consistent and unbiased evaluation protocol, where the test set remained completely unseen during model training and hyperparameter tuning. To further ensure the absence of data leakage, the dataset splitting was performed prior to any preprocessing or augmentation steps, and no overlap exists between the training, validation, and testing subsets. All preprocessing and augmentation operations were applied only to the training data, while validation and test sets were used strictly for evaluation purposes.

To provide a clear overview of the dataset structure and class distribution used in this study, the detailed breakdown of images across all subsets is summarized in Table 1.

The proposed HEFR-Net adopts a dual-stream hybrid convolutional architecture to capture complementary retinal features, as illustrated in Fig. 2. EfficientNet-B3 and ResNet-50 are employed as parallel feature extractors due to their complementary representational strengths. EfficientNet-B3 effectively captures fine-grained textural patterns associated with subtle lesions, while ResNet-50 learns high-level structural representations relevant to global retinal anatomy, particularly for glaucoma detection. The choice of these architectures, instead

Table 2. HEFR-Net architectural blocks.

Stage	Component	Description	Output Dimension
Input	Fundus Image	RGB retinal image	325 × 325 × 3
Feature Extraction (Branch 1)	EfficientNet-B3	Extracts fine-grained textural features + GAP	1,536
Feature Extraction (Branch 2)	ResNet-50	Extracts structural features + GAP	2,048
Feature Fusion	Concatenation	Combines EfficientNet and ResNet features	3,584
Regularization	Batch Normalization + Dropout	Reduces overfitting	3,584
Feature Refinement	Dense + ReLU	Learns compact representation	512
Output	Softmax Layer	Multi-class classification (4 classes)	4

of more recent models such as transformer-based approaches, is motivated by their strong performance on moderate-sized medical datasets and their computational efficiency. Transformer-based models typically require larger datasets and higher computational resources to achieve optimal performance, whereas CNN-based models remain more stable and data-efficient in practical clinical settings. The extracted features from both branches are reduced using Global Average Pooling (GAP) and combined through feature-level concatenation to form a unified representation, which is subsequently passed to a regularized classification head for final prediction.

To clearly describe the internal structure of the model, including feature dimensions and processing stages, the architecture of HEFR-Net is summarized in Table 2.

The preprocessing and training strategy are designed to ensure robustness, reliability, and reproducibility. All images are normalized to the range [0, 1] and resized to 325 × 325 pixels to preserve fine pathological details. During training, online data augmentation is applied, including rotation, shear, zoom, and horizontal flipping, to improve generalization. The model is trained using categorical cross-entropy with label smoothing ($\varepsilon = 0.05$) to reduce overconfidence and improve predictive reliability, while class weighting is applied to address class imbalance. Optimization is performed using the Adam optimizer with a learning rate of 3×10^{-6} and a batch size of 16. Training stability is ensured using ModelCheckpoint, EarlyStopping, and ReduceLROnPlateau callbacks.

To provide a clear summary of all training parameters and ensure reproducibility of the proposed approach, the full training configuration is presented in Table 3.

Table 3. Training configuration and hyperparameters.

Category	Parameter	Value
Input	Resolution	325 × 325 × 3
Training	Batch Size	16
Optimization	Optimizer	Adam
Optimization	Learning Rate	3×10^{-6}
Loss Function	Type	Cross-entropy + Label Smoothing ($\varepsilon = 0.05$)
Regularization	Class Weighting	Inverse-frequency
Data Augmentation	Techniques	Rotation (±15°), Shear (±5%), Zoom (±10%), Horizontal Flip
Training Control	Callbacks	EarlyStopping, ModelCheckpoint, ReduceLROnPlateau
Hardware	GPU	NVIDIA Tesla T4 (16GB)
Implementation	Framework	TensorFlow/Keras
Reproducibility	Random Seeds	Fixed

3. Results and discussion

The effectiveness of the proposed HEFR-Net framework was evaluated using an independent test set of 400 retinal fundus images that were not used during training or validation, ensuring an unbiased assessment of the model's generalization capability. The results demonstrate that the hybrid architecture, combined with a calibration-aware training strategy, provides accurate and clinically relevant multi-class retinal disease classification. In this section, quantitative results, class-wise behavior, error analysis, interpretability, and comparative insights are presented and discussed in a unified manner to improve clarity and readability.

3.1. Overall quantitative performance

The proposed HEFR-Net achieved an overall test accuracy of 95.50% with a test loss of 0.3702, indicating both high predictive performance and well-calibrated probability estimates. To provide a comprehensive evaluation, multiple performance metrics were analyzed, as summarized in Table 4.

These results indicate that the model maintains strong performance across all categories without bias toward any dominant class. This improvement is attributed to the feature-level fusion strategy and calibration-aware training.

3.2. Class-wise performance analysis

To better understand model behavior across different disease categories, class-wise evaluation metrics are presented in Table 5.

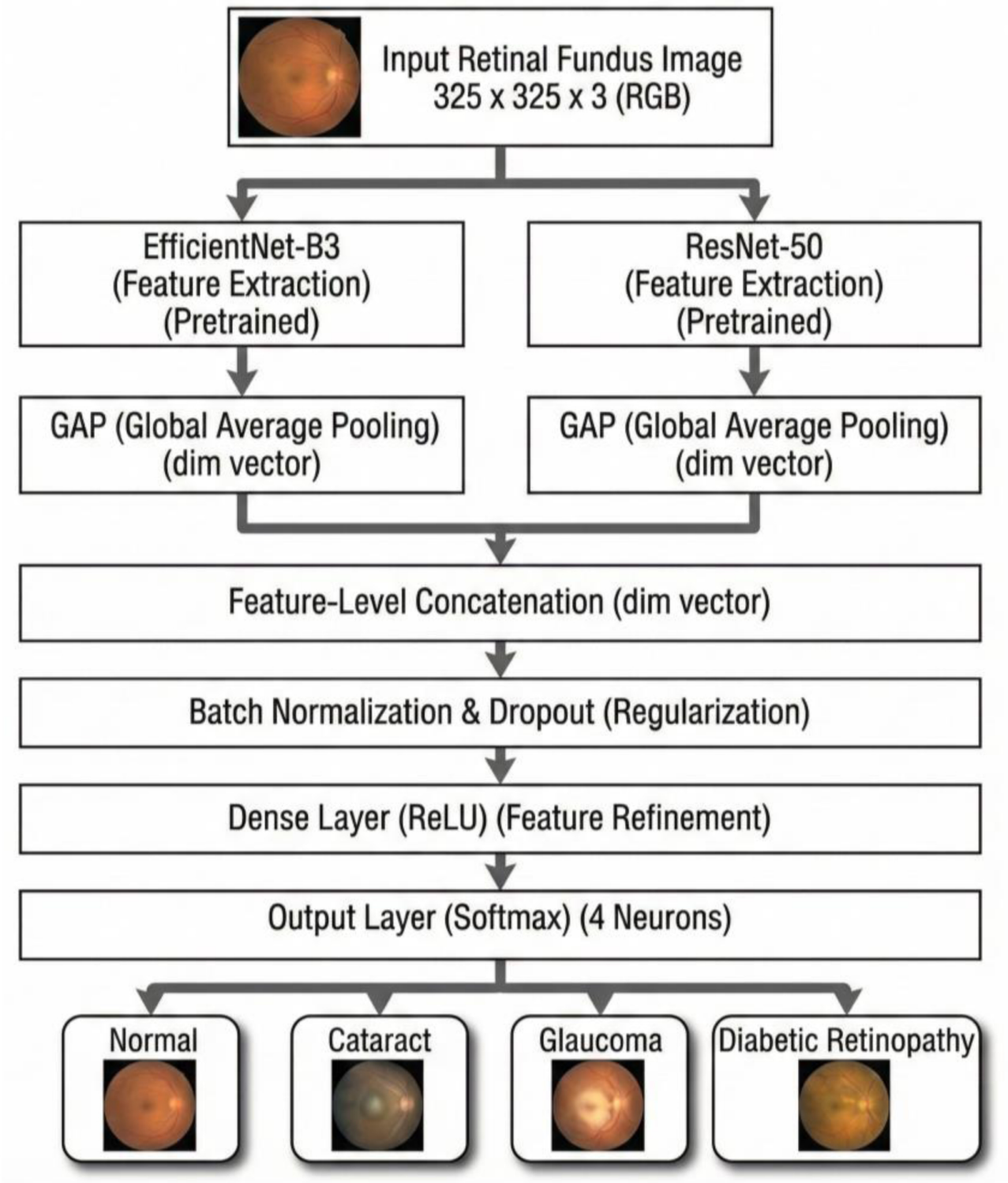

Fig. 2. Overview of the proposed HEFR-Net hybrid architecture.

The results show consistently high performance across all classes. Cataract detection achieved near-perfect results due to its distinct global visual characteristics, while diabetic retinopathy detection benefited from the model's ability to capture fine vascular abnormalities. Glaucoma classification, although slightly lower, reflects the inherent difficulty of detecting subtle structural changes. The reduced

Table 4. Performance metrics of HEFR-Net on the test set.

Metric	Value	Interpretation
Accuracy	95.50%	High overall classification correctness
Loss	0.3702	Indicates reliable and calibrated predictions
Macro Precision	96%	Consistent prediction reliability across classes
Macro Recall	95%	Strong ability to detect all disease categories
Weighted F1-Score	95%	Balanced performance across all classes

Table 5. Class-wise classification performance.

Class	Precision	Recall	F1-Score	Support
Cataract	0.99	0.99	0.99	95
Diabetic Retinopathy	0.97	0.98	0.98	106
Glaucoma	0.94	0.92	0.93	91
Normal	0.92	0.93	0.92	108

recall indicates that some glaucoma cases visually resemble normal retinal structures. The relatively lower performance can be attributed to the nature of glaucoma, which is primarily characterized by subtle structural variations such as optic disc cupping and changes in the cup-to-disc ratio. These features are often less visually prominent and may overlap with normal anatomical variations, making accurate classification more challenging. Furthermore, glaucoma diagnosis typically requires additional clinical context and three-dimensional structural assessment, which are not fully captured in two-dimensional fundus images. Although class weighting was considered during model development, the dataset is relatively balanced across all classes. Therefore, its impact on performance was minimal. The consistent precision, recall, and F1-score values confirm that the model does not exhibit significant bias and can learn all classes effectively without relying on imbalance-handling techniques.

3.3. Computational efficiency and practical accessibility

Although HEFR-Net adopts a dual-backbone architecture, it demonstrates efficient computational performance and practical feasibility for deployment. The model was successfully trained and evaluated on a standard cloud-based GPU (NVIDIA Tesla T4), indicating that high-performance ophthalmic AI systems can be developed without requiring specialized hardware. In terms of inference, the model achieves fast prediction speed, with inference time per image typically within a fraction of a second on GPU hardware. This enables real-time or near real-time screening, which is essential for clinical applications.

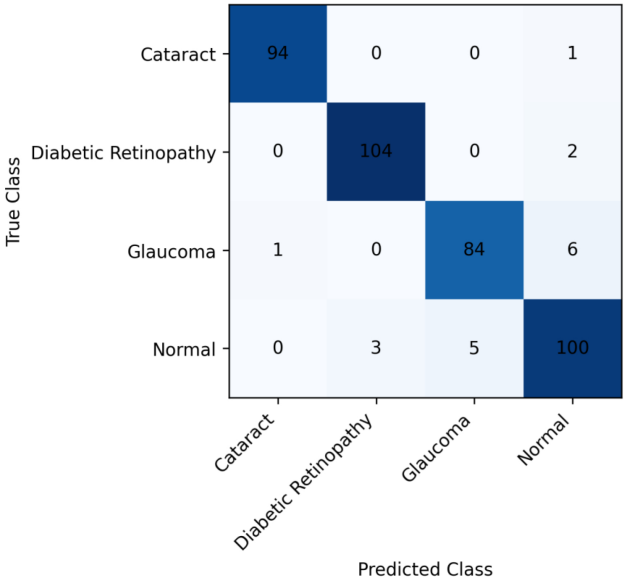

Fig. 3. Confusion matrix of HEFR-Net for multi-class retinal disease classification.

From a deployment perspective, the model is compatible with existing medical AI infrastructures and can be integrated into hospital systems and clinical decision-support platforms. Its moderate computational requirements, combined with high accuracy and interpretability, make it suitable for large-scale automated retinal screening in real-world healthcare environments.

3.4. Confusion matrix and error interpretation

To analyze prediction behavior at the instance level, the confusion matrix is presented in Fig. 3.

The confusion matrix shows strong diagonal values, indicating high classification accuracy across all classes. Specifically, 94 cataract, 104 diabetic retinopathy, 84 glaucoma, and 100 normal cases were correctly classified.

However, 18 misclassifications provide important insights. The most critical errors occur in glaucoma cases misclassified as normal (6 cases), representing false negatives that are clinically significant since untreated glaucoma can lead to irreversible vision loss.

Other errors include minor confusion between glaucoma and cataract, and normal cases misclassified as diseased categories, which represent safer false positives. These findings indicate that the main limitation lies in detecting subtle glaucoma features.

3.5. Visual explainability and clinical interpretation

To further validate model transparency, Grad-CAM visualizations are presented in Fig. 4. The visual explanations demonstrate that the model consistently

focuses on clinically relevant regions. In diabetic retinopathy cases, activation maps highlight lesion areas such as microaneurysms and hemorrhages, while in glaucoma cases, attention is concentrated on the optic disc and surrounding regions, reflecting structural changes associated with optic nerve damage. Cataract predictions exhibit diffuse activation patterns consistent with global opacity, whereas normal images show more uniform attention without localized pathological focus. These observations confirm that HEFR-Net produces interpretable and clinically meaningful predictions aligned with established ophthalmic knowledge. Grad-CAM provides valuable visual insights into the regions influencing model predictions and enhances transparency by linking model decisions to anatomically meaningful features. This contributes to improved trust in the model's outputs, particularly in clinical decision-support scenarios, where understanding the reasoning behind predictions is essential.

3.6. Ablation study

To evaluate the contribution of individual components in the proposed HEFR-Net framework, an ablation study was conducted by comparing the performance of each backbone model independently with the hybrid feature-fusion architecture. This analysis aims to verify whether the integration of multiple feature extractors leads to measurable performance improvements. The quantitative results of this comparison are summarized in Table 6.

As shown in Table 6, both EfficientNet-B3 (93.0%) and ResNet-50 (92.7%) achieve competitive performance when used as standalone models. This comparison directly illustrates how each backbone performs individually relative to the proposed hybrid model, where neither backbone alone achieves the same level of performance as HEFR-Net. However, the proposed hybrid HEFR-Net demonstrates a clear improvement in classification accuracy. This improvement can be attributed to the complementary nature of the extracted features, where EfficientNet-B3 captures fine-grained textural patterns associated with subtle lesions, while ResNet-50 focuses on high-level structural representations related to global retinal anatomy. The fusion of these heterogeneous features at the feature level enables the model to construct a more comprehensive representation of retinal characteristics, leading to enhanced classification performance. These results confirm that the hybrid architecture plays a significant role in improving model effectiveness and supports the design choice of combining multiple backbone networks within a unified framework.

Table 6. Ablation study of HEFR-Net components.

Model Variant	Accuracy (%)
EfficientNet-B3	93.0%
ResNet-50	92.7%
HEFR-Net (Proposed)	95.50%

3.7. Comparative evaluation of existing eye-disease classification studies

This subsection presents a structured comparative analysis of representative studies in multi-class

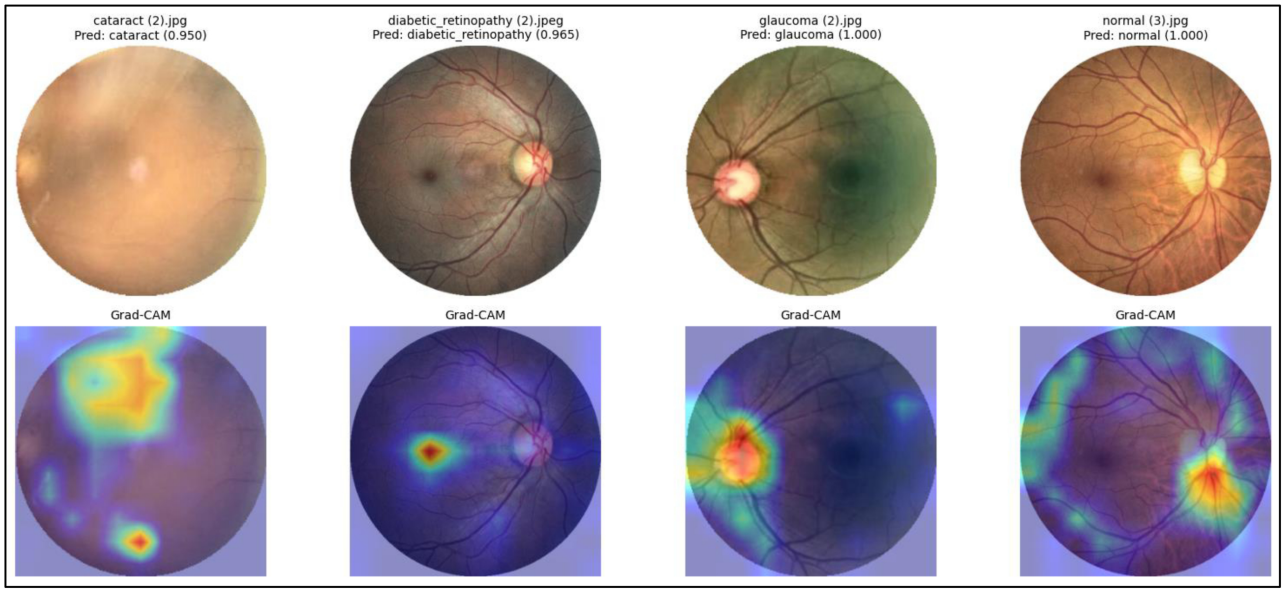

Fig. 4. Grad-CAM visualizations highlighting regions influencing HEFR-Net predictions.

retinal disease classification, as summarized in Table 7. The comparison considers key methodological aspects, including classification accuracy, model architecture, feature fusion strategy, calibration mechanisms, and the integration of explainability techniques. By jointly analyzing these factors, the table highlights the evolution of existing approaches and identifies the critical gaps addressed by the proposed HEFR-Net framework.

As shown in Table 7, most previous studies rely on single-stream convolutional neural networks that are typically fine-tuned using transfer learning, achieving accuracy ranging from approximately 85% to 94%. While these results indicate strong classification capability, the majority of these approaches are limited by the use of isolated backbone architectures without feature fusion, which restricts their ability to capture complementary representations of retinal abnormalities. Even hybrid approaches, such as CNN–ML pipelines, do not enable deep interaction between features within an end-to-end learning framework. Another important limitation observed across the literature is the absence of predictive calibration strategies. None of the compared studies explicitly incorporate calibration-aware training mechanisms, which may lead to overconfident predictions, particularly in clinically ambiguous cases. Furthermore, explainability remains largely underexplored, with only one study employing weak post-hoc interpretation using LIME, while no prior work integrates strong visual explainability with high classification accuracy. It is also worth noting that more recent architectures, such as DenseNet, Vision Transformers (ViT), and ConvNeXt, were not explicitly included in this comparative analysis. This is primarily because the comparison focuses on studies conducted under similar experimental settings and using comparable datasets to ensure fairness and consistency. In addition, these architectures typically require larger-scale datasets and higher computational resources to achieve optimal performance, which may not align with the constraints of the current study. In contrast, the proposed HEFR-Net introduces a unified hybrid CNN architecture that performs feature-level fusion between EfficientNet-B3 and ResNet-50, enabling early interaction between fine-grained texture features and high-level structural representations. The framework further incorporates calibration-aware training through label smoothing to improve predictive reliability and integrates Grad-CAM to provide clinically meaningful visual explanations. This combination of architectural design and trustworthy AI mechanisms results in the highest reported accuracy in Table 7 (95.50%), while simultaneously addressing key limitations related to feature complementarity, calibration, and interpretability. Overall, the comparative analysis demonstrates that HEFR-Net not only improves classification performance but also provides a more reliable, interpretable, and clinically applicable solution for automated retinal disease classification.

3.8. Trustworthy AI perspective and practical implications

The performance gains of HEFR-Net are not only due to architectural complexity but also due to its alignment with Trustworthy AI principles. Label smoothing improves calibration by reducing overconfidence, ensuring that predicted probabilities better reflect true uncertainty. Feature-level fusion enhances robustness by combining texture and structural representations, allowing the model to handle diverse imaging conditions. This

Table 7. Comparative evaluation of retinal disease classification studies.

Ref.	Accuracy (%)	Model Type	Fusion Level	Calibration Strategy	XAI Support
[17]	84.7%	Ensemble CNN	Prediction-level	None	None
[16]	85.0%	Sequential CNN	None	None	LIME
[11]	88.0%	Single CNN	None	None	None
[15]	90.9%	Hybrid CNN–ML	Feature-level	None	None
[9]	92.18%	Single CNN	None	None	None
[10]	92.7%	Single CNN	None	None	None
[8]	92.91%	Single CNN	None	None	None
[13]	93.0%	Single CNN	None	None	None
[14]	94.0%	Single CNN	None	None	None
[12]	94.30%	Single CNN	None	None	None
[23]	87.0%	CNN (ManualNet, LeNet, DenseNet)	None	None	None
[24]	94.82%	CNN (ResNet-152, EfficientNetB0, MobileNetV1, DenseNet-121)	None	None	None
[25]	93.14%	Dual-Branch CNN + Semi-Supervised	Feature-level	Implicit (ArcMargin + Focal Loss)	None
HEFR-Net (Proposed)	**95.50%**	**Hybrid CNN**	**Feature-level**	**Label Smoothing**	**Grad-CAM**

complementary interaction significantly improves generalization. Despite its dual-backbone design, the model remains computationally efficient and was successfully trained using standard GPU resources. This demonstrates its practical feasibility for real-world deployment in clinical screening systems.

3.9. Experimental reproducibility

All experiments in this study were conducted under fixed and controlled conditions, including consistent data splits, model architecture, and training configurations. The reported results correspond to a stable and representative run, obtained after careful tuning and validation of the training process. To ensure reliability, multiple stabilization mechanisms such as early stopping, model checkpointing, and learning rate scheduling were employed. These strategies help reduce variability and ensure consistent convergence behavior, resulting in robust and reproducible performance.

4. Conclusion and future directions

This study introduces a carefully evaluated trustworthy deep learning framework, termed HEFR-Net, for multi-class retinal disease classification. By adopting a hybrid architecture that fuses complementary feature representations from EfficientNet-B3 and ResNet-50 at the feature level, and by incorporating calibration-aware training through label smoothing and class weighting, the proposed model achieves an accuracy of 95.50% on the Kaggle Eye Disease dataset, outperforming existing single-backbone approaches such as ResNet50 and MobileNet-based models. The results highlight the critical role of feature fusion in capturing diverse pathological characteristics associated with cataract, diabetic retinopathy, and glaucoma, while explicitly addressing predictive calibration—an often overlooked yet clinically vital aspect of medical AI systems. It is important to emphasize that the proposed framework does not rely on a physiological simulation of the human eye. Instead, it adopts a phenotypic image-based representation, where retinal diseases are identified through observable visual patterns in fundus images. This design choice enhances the practical applicability of the model and enables direct integration into real-world clinical screening systems. Detailed performance analysis, including class-wise metrics, confusion matrix evaluation, and Grad-CAM–based visual explanations, confirms the robustness, reliability, and interpretability of the framework, while also identifying glaucoma false negatives as the primary remaining challenge. Future work will focus on several promising directions. These include deeper integration of explainability mechanisms, such as advanced Grad-CAM and attention-based visualization techniques, incorporation of transformer-based components to capture long-range dependencies, and external validation on real-world clinical datasets as well as heterogeneous multi-source retinal datasets to further assess generalizability and clinical reliability. In summary, HEFR-Net represents an accurate, interpretable, calibrated, and practically deployable solution with strong potential to enhance automated ophthalmic screening and support clinical decision-making.

Acknowledgement

The authors would like to acknowledge the academic support and research facilities that contributed to the completion of this work.

Conflict of interest statement

The authors declare that they have no known financial, institutional, or personal conflicts of interest that could have influenced the work reported in this manuscript. All authors have reviewed and approved the final version of the manuscript and agree to its submission to the journal.

Author contributions

Ali M. Duhaim: Conceptualization, methodology, model development, experiments, data analysis, writing the manuscript.

Abbas M. Al-Bakry: Supervision, validation, manuscript review and editing.

Ethics statement

This study used publicly available retinal fundus image datasets that are fully anonymized. No human subjects or personal data were involved; therefore, ethical approval and informed consent were not required.

Funding statement

This research received no external funding. The authors conducted this study independently as part of academic research activities at the University of

Information Technology and Communication (UoITC), Baghdad, Iraq.

References

1. R. R. A. Bourne *et al.*, "Causes of vision loss worldwide, 1990–2010: a systematic analysis," *lancet Glob. Heal.*, vol. 1, no. 6, pp. e339–e349, 2013.
2. Y. Zeng, Y. Liu, X. Chen, J. Kenny, R. Rong, and X. Xia, "Global, re gional, and national burden of blindness and vision loss attributable to smoking from 1990 to 2021, and forecasts to 2030: findings from the global burden of disease study 2021," *BMC Public Health*, vol. 25, no. 1, p. 440, 2025.
3. R. N. Weinreb, T. Aung, and F. A. Medeiros, "The pathophysiology and treatment of glaucoma," *JAMA*, vol. 311, p. 1901, 1981.
4. M. D. Abràmoff *et al.*, "Automated analysis of retinal images for detection of referable diabetic retinopathy," *JAMA Ophthalmol.*, vol. 131, no. 3, pp. 351–357, 2013.
5. V. Gulshan *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *jama*, vol. 316, no. 22, pp. 2402–2410, 2016.
6. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
7. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
8. Y. Cui, R. Sun, B. Liu, Z. Liu, and T. T. Toe, "Eye diseases classification using transfer learning of residual neural network," in *2023 3rd International Symposium on Computer Technology and Information Science (ISCTIS)*, 2023, pp. 244–248.
9. M. A. Mohamed *et al.*, "Multi-class eye disease classification using deep learning," in *2023 Eleventh International Conference on Intelligent Computing and Information Systems (ICICIS)*, 2023, pp. 489–494.
10. S. Benbakreti, S. Benbakreti, and U. Ozkaya, "The classification of eye diseases from fundus images based on CNN and pretrained models," 2024.
11. G. Verma, "Efficient retinal disease detection with ResNet-18: A deep learning approach," in *2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI)*, 2024, pp. 812–816.
12. A. Rafay, Z. Asghar, H. Manzoor, and W. Hussain, "Eye-CNN: exploring the potential of convolutional neural networks for identification of multiple eye diseases through retinal imagery," *Int. Ophthalmol.*, vol. 43, no. 10, pp. 3569–3586, 2023.
13. P. Kaushik and P. Sharma, "Optimizing automated eye disease detection: A deep learning approach using EfficientNetB3 for accurate multi-class classification," in *2024 IEEE International Conference on Intelligent Signal Processing and Effective Communication Technologies (INSPECT)*, 2024, pp. 1–6.
14. R. Singh, N. Sharma, and R. Gupta, "The proposed convolutional neural network architecture for the detection and classification of eye diseases," in *2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE)*, 2023, pp. 1–6.
15. K. İbrahim and İ. ÇINAR, "Evaluation of Machine Learning and Deep Learning Approaches for Automatic Detection of Eye Diseases," *Intell. Methods Eng. Sci.*, vol. 3, no. 1, pp. 37–45, 2024.
16. A. Aaron Samuel, N. Muhammad Fadil, and R. Beaulah Jeyavathana, "Sequential model using explainable AI method to detect eye diseases," in *International Conference on Deep Sciences for Computing and Communications*, 2023, pp. 268–280.
17. B. Naseeba, L. Burra, C. S. Kumari, B. Yohoshiva, and K. Shyam, "Eye disease classification using ensemble learning," in *2024 3rd Edition of IEEE Delhi Section Flagship Conference (DELCON)*, 2024, pp. 1–6.
18. M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, 2019, pp. 6105–6114.
19. M. Cannarsa, "Ethics guidelines for trustworthy AI," *Cambridge Handb. lawyering Digit. age*, pp. 283–297, 2021.
20. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International conference on machine learning*, 2017, pp. 1321–1330.
21. R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?," *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.
22. "eye-diseases-classification." https://www.kaggle.com/datasets/gunavenkatdoddi/eye-diseases-classification.
23. D. Jose, "Classification of EYE diseases using multi-model CNN," in *2023 IEEE 4th Annual Flagship India Council International Subsections Conference (INDISCON)*, 2023, pp. 1–6.

24. E. C. Yaurentius, T. R. D. Saputri, E. Tanuwijaya, and R. E. Sutanto, “Comparative study of CNN-based architectures on eye diseases classification using fundus images to aid ophthalmologist,” *J. Tek. Inform.*, vol. 6, no. 1, pp. 249–257, 2025.
25. M. H. Malik, Z. Wan, Y. Gao, and D.-W. Ding, “Efficient diagnosis of retinal disorders using dual-branch semi-supervised learning (DB-SSL): An enhanced multi-class classification approach,” *Comput. Med. Imaging Graph.*, vol. 121, p. 102494, 2025.