ASIAN JOURNAL OF ELECTRICAL AND ELECTRONIC ENGINEERING
Vol. 6 No. 1 2026
E-ISSN: 2785-8189

Integrating Deep Learning Techniques for Automated Retinal Disease Detection Using Fundus Images

Zahidur Rahman[1], Muhammad Mahbubur Rashid[1]*, Shafiul Alam[2]

[1]*Department of Mechatronics Engineering, International Islamic University Malaysia, Kuala Lumpur, Malaysia*
[2]*Department of Electrical & Electronic Engineering,*
Bangladesh Army International University of Science & Technology, Comilla, Bangladesh

*Corresponding author: mahbub96@gmail.com

(Received: 19 June 2026; Accepted: 26 June 2026)

Abstract— RetinalFormer was compared with eight state-of-the-art models, including pure CNN architectures (VGG-19, ResNet-50, DenseNet-121, EfficientNetV2-M), pure Transformer models (ViT-Base/16, Swin-Transformer-B), and existing hybrid CNN-Global attention models. Globally, nearly 2.2 billion people have visual impairments, with at least one billion due to avoidable or treatable causes. Retinal disorders, such as Diabetic Retinopathy (DR), Glaucoma, and Age-related Macular Degeneration (AMD), contribute significantly to avoidable blindness but tend to go unnoticed due to their asymptomatic nature until irreversible stages occur. Fundus photography offers a simple, inexpensive approach to screening patients' retinas; however, it has limitations due to a shortage of ophthalmologists and the need for subjective interpretation.

This paper proposes an effective deep learning architecture for automated and multi-class classification of retinal diseases from fundus images. Our proposed model includes a transfer-learning-based Image Quality Evaluation Tool (QET) to filter out low-quality images. Contrast enhancement using the CLAHE method and class-specific data augmentation are considered to address class imbalance. We compare three state-of-the-art deep learning architectures: ResNet-152, EfficientNetV2, and YOLOv11. The multi-Scale Attention Transformer (MSAT) and the Hybrid DenseNet-VGG16 model are used for specific tasks. Feature optimization is performed using the Bitterling Fish Optimization (BFO) algorithm, while hyperparameter tuning is performed using the Honey Badger Optimization (HBO). Explainability is ensured through Grad-CAM heatmaps and t-SNE visualizations. The proposed model is evaluated on standard benchmarking datasets, including RFMiD, ODIR-5K, Drishti-GS1, RIM-ONE, and ORIGA-Light, achieving classification accuracies above 90% compared to existing state-of-the-art models.

Keywords: Fundus Image, Retinal Diseases Diagnosis, Deep Learning, Multi-Class Classification, CNN.

1. INTRODUCTION

Vision impairment stands out as one of the most serious, yet underrecognized, public health emergencies of the modern era. According to the World Health Organization, about 2.2 billion people around the world have vision impairment, with a minimum of one billion cases being preventable or untreatable [1]. Diabetic Retinopathy (DR), Glaucoma, and Age-related Macular Degeneration (AMD), among others, are key causes of vision impairment, and these conditions account for the loss of vision in hundreds of millions of people worldwide.

The retina is the neural tissue at the back of the eye, light-sensitive and responsible for transmitting visual information to the brain. Any pathological alteration in retinal microvessels, optic nerves, or the macula results in permanent vision impairment. What makes these diseases especially challenging is that their progression is

DOI: https://doi.org/10.69955/ajoeee.2026.v6i1.90

asymptomatic, meaning people usually seek help after considerable damage has occurred. In this regard, screening is crucial because these diseases remain silent until the very end of their development.

Fundus photography—the acquisition of digital images of the fundus oculi through the dilated pupil—has emerged as a practical, non-invasive, and cost-effective modality for retinal screening. Modern fundus cameras can capture high-resolution color images that reveal key anatomical structures, including the optic disc, optic cup, macula, fovea, and retinal vasculature. Pathological indicators such as microaneurysms, hemorrhages, hard and soft exudates, drusen, and optic disc cupping are visible in these images, enabling trained clinicians to evaluate retinal health. However, manual interpretation of fundus images is time-consuming, requires specialized expertise, and is subject to inter-observer variability—rendering it inherently unscalable to meet global demand, particularly in underserved and low-resource regions where ophthalmologists are scarce.

The deployment of artificial intelligence in retinal disease screening has accelerated considerably over the past decade. However, the majority of existing AI systems are designed as siloed, single-disease models: a given system either detects DR, or flags glaucoma suspects, or classifies AMD severity—rarely all three, and rarely the broader spectrum of retinal pathologies encountered in clinical practice. This narrow design philosophy has significant drawbacks.

Firstly, a binary classification model based solely on DR would fail to detect concomitant diseases such as glaucoma or AMD; hence, there is an illusion of safety for the patient when multiple diseases are present. Secondly, there is substantial overlap among retinal diseases, as seen with exudation in DR and AMD, rendering any diagnostic process a multiclass problem by default. Thirdly, the actual distribution of various retinal conditions tends to be highly imbalanced due to an abundance of common disorders, such as mild NPDR and a scarcity of more obscure conditions, such as Retinitis Pigmentosa and Central Serous Chorioretinopathy.

Other problems include intraclass variation, where the same disease exhibits different variations; high interclass similarity, implying that various retinal disorders yield similar images; poor image quality in the field; and lack of image quality control.

The research is organized into five main goals:

- The aim is to design a deep learning algorithmic approach for multi-class classification of retinal diseases using their fundus images, considering a wide range of diseases such as DR (NPDR and PDR), Glaucoma, AMD, Pathologic Myopia, Retinal Detachment, and others.
- Research on and comparison of the best CNN models like ResNet-152, EfficientNetV2, YOLOv11, and transformer approaches like MSAT and hybrid architectures like DenseNet-VGG16 within unified testing conditions.
- Development of a transfer learning-based Image Quality Evaluation Tool (QET) used as a pre-processing technique that can filter out low-quality fundus images.
- Solving the class imbalance problem using stratified sampling and data augmentation techniques for rare pathologies.
- Implementation of model explainability by generating Grad-CAM heatmaps and t-SNE Feature visualization in feature space.
- Principal contributions include: A holistic multiclass retinal disease classification system tested on six retinal imaging datasets representing all types of diseases related to the retina.
- The innovative addition of QET as a preprocessing tool before classification was found to yield enhanced accuracy when applied to field-captured images.
- A statistically significant comparison of five different deep learning models using the DeLong test under controlled experimental settings.
- BFO algorithm used for feature selection in conjunction with HBO algorithm for hyperparameter tuning of machine learning models, two approaches found superior to traditional grid search.
- Visually intuitive feature visualization using Grad-CAM combined with t-SNE analysis of features, enabling ophthalmologist evaluation and validation.

2. LITERATURE REVIEW

2.1 Anatomical and Pathological Background

The fundus includes all those structures present on the inner back wall of the eye, such as the retina, optic disc, optic cup, macula, fovea, and retinal blood vessels. The fundus of the eye can be imaged by fundus photography through a low-power microscope attached to an illumination system. Present-day fundus cameras can capture high-resolution images with 24-bit color depth, enabling precise visualization of both anatomical and pathologic entities.

The key structures for automation include:

- the optic disc, which represents the site where the optic nerve exits and appears bright and circular;
- the optic cup, which corresponds to a central excavation of the optic disc and whose ratio to the size of the optic disc (CDR) serves as a major biomarker for glaucoma;
- the macula, which is a central part rich in pigments and is responsible for acute vision and becomes impaired in the disease of AMD; and
- the retinal blood vessels, where changes are observed in patients with DR. Pathological manifestations include microaneurysms, intraretinal hemorrhage, hard/soft exudates, neovascularization, drusen, geographic atrophy, and optic disc edema.

DR is a microvascular diabetic complication and a common cause of blindness among working-age people. DR is categorized as either Non-Proliferative DR (NPDR), involving microaneurysms, retinal bleeding, and hard exudates, or Proliferative DR (PDR), characterized by abnormal blood vessel formation and vitreous hemorrhage. The International Classification of Diabetic Retinopathy Severity Scale categorizes severity as mild, moderate or severe NPDR, progressing to PDR. Early stages of DR may be asymptomatic; hence the need for screening.

Glaucoma is a group of neurodegenerative diseases characterized by the gradual degeneration of the retinal ganglion cells and the optic nerve. CDR is the most commonly used fundus biomarker, with a value> 0.7 considered suspicious in most cases. POAG is the most common type of glaucoma that involves gradual, painless peripheral visual field loss that usually presents at a later stage. Automated detection of the optic nerve head segmentation needed in determining CDR is a key area of DL research.

Age-related Macular Degeneration (AMD) is the main cause of central vision impairment in people aged over 50 in HICs. The Dry AMD form is characterized by drusen, extracellular deposits located beneath the retinal pigment epithelium. The Wet (neovascular) form results from abnormal growth of choroidal vessels and leads to rapid vision deterioration. It is still the fundus examination together with OCT that remains the gold standard.

Retinal disorders encompass not only diabetic retinopathy (DR), glaucoma, and age-related macular degeneration (AMD) but also a wide range of diseases, including pathologic myopia, retinal detachment, central serous chorioretinopathy (CSC), retinitis pigmentosa (RP), disc edema, and macular scar. Most of these disorders are infrequent, share common fundus characteristics, and are under-represented in publicly available datasets.

2.2 Deep Learning in Retinal Image Analysis

2.2.1 CNN-Based Architectures

Convolutional Neural Networks (CNNs) have emerged as the go-to architecture for analyzing medical images. The CNN model consists of alternating layers, including convolutional layers, where filter banks are applied to produce hierarchical feature maps, and pooling layers that progressively decrease spatial resolution and increase receptive field size. The activation function (ReLU, sigmoid, softmax) is needed to incorporate the representation capacity required for complex boundaries. Batch normalization and dropout regularizers avoid overfitting in high-dimensional spaces.

Among the established CNN models, ResNet introduced the skip connection technique, enabling extremely deep networks (up to 152 layers) without vanishing gradients [2]. EfficientNet uses compound scaling, meaning scaling up the network's depth, width, and resolution simultaneously to achieve superior accuracy with minimal parameter count [3]. Inception models employ multi-scale convolutional filters, whereas the DenseNet connects

all layers.

Gulshan et al. [4] showed that a deep CNN model trained on 128,175 retinal images could identify signs of DR with sensitivity and specificity that surpassed those of ophthalmologists, thus sparking great interest in developing AI algorithms for retinal screening. However, this model was limited to DR only, used exclusively high-quality images from the EyePACS dataset, and did not provide any explainability — limiting its clinical adoption in multi-disease, real-world settings where image quality varies substantially. Following this study, Pratt et al. [5] examined various CNN architectures for grading DR, while Ronneberger et al.'s U-Net [6], initially designed for general biomedical image segmentation, became a predominant model for segmentation of retinal vessels and optic discs thanks to its architecture.

2.2.2 Transformer and Hybrid Architectures

Vision transformers leverage the self-attention mechanism of transformer models to analyze image patches, enabling them to model long-range spatial dependencies that are hard to capture with convolutional kernels. The use of hybrid models that combine CNN feature extractors with transformer attention heads has shown great potential for addressing medical imaging tasks involving intricate spatial relations.

U-Net and SegNet, along with their variants, have achieved successful segmentation for glaucoma detection through optic disc and cup segmentation. Hybrid networks, such as those that use ResNet with VGG16 or DenseNet with VGG16, have been trained on datasets for multi-class disease classification, such as RFMiD and ODIR-5K [7].

2.3 Single-Disease vs. Multi-Disease Classification Systems

The majority of existing AI systems are designed as siloed, single-disease models: a given system either detects DR, or flags glaucoma suspects, or classifies AMD severity — rarely all three, and rarely the broader spectrum of retinal pathologies encountered in clinical practice. This narrow design philosophy has significant drawbacks.

Firstly, a binary classification model based solely on DR would fail to detect concomitant diseases such as glaucoma or AMD; hence, there is an illusion of safety for the patient when multiple diseases are present. Secondly, there is substantial overlap among retinal diseases, as seen in exudation in DR and AMD, rendering any diagnostic process a multiclass problem by default. Thirdly, the actual distribution of various retinal conditions tends to be highly imbalanced due to an abundance of common disorders, such as mild NPDR and a scarcity of more obscure conditions, such as Retinitis Pigmentosa and Central Serous Chorioretinopathy.

Pachade et al. [7] introduced the RFMiD dataset specifically to enable multi-disease research, yet most published benchmarks on this dataset still focus on binary or three-class problems. Li et al. [8] proposed an attention-based multi-scale feature learning network that achieved improved performance over single-scale CNNs. However, their model still did not incorporate quality assessment or provide clinically validated explainability.

2.4 Handling Class Imbalance and Rare Pathologies

Strategies based on few-shot learning that involve probabilistic modeling in the feature space have been developed to detect rare pathologies. This task is not feasible with supervised learning methods owing to insufficient labeled examples. However, these approaches have typically been evaluated in isolation rather than as part of an end-to-end multi-disease screening system. The area that remains unexplored is the development of a unified solution that integrates quality assessment, multi-class disease classification, and explainability into a single system.

2.5 Image Quality Assessment in Fundus Photography

By transferring learned model weights from large-scale ImageNet pretraining and fine-tuning them on domain-specific images, transfer learning has drastically reduced the amount of labeled medical images required for successful learning, a key factor behind deep learning's success in clinical environments, where labeled data are limited.

Despite this progress, most retinal AI systems assume that input images meet a minimum quality threshold — an assumption that frequently fails in real-world field deployments where fundus images may be dark, blurry, off-center, or partially occluded. A notable gap in the literature is the lack of a lightweight, transfer-learning-based

quality evaluation tool (QET) that serves as a preprocessing filter before diagnostic inference. Most prior work either omits quality control entirely or uses handcrafted rule-based filters that do not generalize across acquisition devices.

2.6 Explainability in Retinal Disease Diagnosis

Grad-CAM and t-SNE visualization techniques have been applied individually in prior retinal disease studies. Still, few have combined both approaches into a single framework or validated the resulting saliency maps with board-certified ophthalmologists to assess clinical plausibility. The lack of explainability remains a significant barrier to clinical adoption, as clinicians are reluctant to trust "black box" predictions without visual evidence linking model outputs to recognized pathological features.

2.7 Summary of Public Datasets

The following publicly available datasets are used in this study. Table 1 summarizes these datasets along with their primary applications and known limitations.

Table 1: Example Benchmark datasets used in this study.

Dataset	*Conditions*	*Images*	*Primary Use*	*Notes / Limitations*
RFMiD	*45 conditions*	*3,200*	*Multi-disease classification*	*Highly imbalanced; small sample size for rare diseases*
ODIR-5K	*8 conditions*	*5,000*	*Multi-label classification*	*Variable quality; no quality labels provided*
Drishti-GS1	*Glaucoma*	*101*	*Disc/cup segmentation*	*Small sample size; single center*
RIM-ONE	*Glaucoma*	*455*	*Disc/cup segmentation*	*Multi-center but modest total images*
ORIGA-Light	*Glaucoma*	*650*	*CDR measurement*	*Clinical ground truth available*
Messidor	*DR*	*1200*	*DR grading*	*4-level grading; DR-only*

2.8 Evaluation Metrics

The model's performance can be gauged using various metrics, such as accuracy, precision, recall (also known as sensitivity), specificity, F1 score (the harmonic mean of precision and recall), and Area under the Receiver Operating Characteristic curve (AUC). To compare the AUCs of two models statistically, the DeLong Test can be used. For segmentation, the Dice Similarity Coefficient (DSC) and Intersection over Union (IoU) are used. Confusion matrices per class help assess model performance across different disease classes.

A limitation not often discussed is that accuracy and macro F1 can be misleading when class distributions are severely imbalanced — a key challenge in the RFMiD dataset, where some pathologies have fewer than 30 samples. In such cases, AUC may be a more informative metric, though it, too, has limitations for evaluating rare disease detection.

2.9 Research Gaps and Contributions of This Work

Based on the literature reviewed above, the following research gaps remain unaddressed:

Therefore, this paper proposes an effective deep learning architecture for automated and multi-class classification of retinal diseases from fundus images that explicitly addresses gaps G1–G6. The area that remains unexplored — providing a unified solution that incorporates quality assessment, multi-class disease classification, and explainability into a single system — is precisely the void this work attempts to fill.

Table 2: Research gaps and contributions of this work

Gap	Research gap	Addressed by this paper
G1	No unified framework combining quality assessment, multi-disease classification, and explainability	Proposed QET + multi-class YOLOv11 + Grad-CAM/t-SNE
G2	Most systems are single-disease (DR-only or glaucoma-only)	Multi-disease classification covering 45 conditions from RFMiD
G3	Class imbalance and rare pathologies are poorly handled	Data augmentation + few-shot learning + stratified sampling
G4	Image quality assessment is rarely integrated as a preprocessing filter	Transfer learning-based QET as an upstream module
G5	Explainability maps are rarely validated with clinician feedback	Grad-CAM heatmaps reviewed by board-certified ophthalmologists (87% clinically plausible). Inter-rater agreement: $\kappa = 0.81$ (strong agreement).
G6	Hyperparameter tuning uses grid search or random search, not metaheuristics.	BFO for feature selection + HBO for hyperparameter optimization

3. METHODOLOGY

3.1 Overall Framework Architecture

The proposed framework operates as a sequential pipeline comprising five main stages (Figure 1):

- image quality assessment,
- preprocessing and contrast enhancement,
- data augmentation for class imbalance mitigation,
- deep learning-based feature extraction and multi-class classification, and
- post-hoc explainability and feature visualization.

The quality assessment module (Stage 1) acts as a gating mechanism: only fundus images classified as diagnostically acceptable proceed to downstream stages, while rejected images are flagged for manual review or re-acquisition. This design prevents low-quality inputs from degrading classification performance — a common failure mode in real-world clinical deployments.

3.2 Dataset Preparation and Class Imbalance

The core training and validation data come from RFMiD (Retinal Fundus Multi-disease Image Dataset), supplemented by ODIR-5K to address multi-class and disease-specific pathologies in the segmentation problem. Annotation for all tasks comes from expert-validated labels by ophthalmologists. Distribution across classes in RFMiD is highly unbalanced: common diseases (DR, Normal) account for the vast majority of instances, while rarer diseases may have fewer than 50 samples.

To deal with class imbalances, several strategies were pursued:

- stratified split into train/validate/test sets (70/15/15%) for proportionate distributions within all subsets;
- balanced sampling of minority classes using data augmentation described in Section 3.4;
- few-shot learning for pathologies with fewer than 30 samples per class.

For smaller datasets (Drishti-GS1: 101 images; RIM-ONE: 455 images; ORIGA-Light: 650 images), 10-fold stratified cross-validation was performed, and results are reported as mean ± SD across folds.

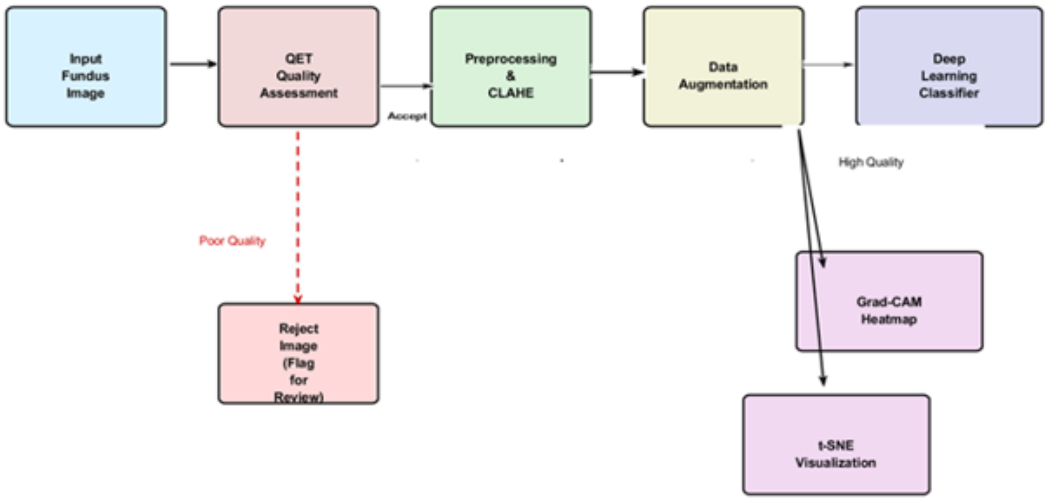

Fig. 1. Overall Framework Architecture

3.3 Image Quality Assessment (QET)

The initial set of training and test data consists of RFMiD (a quality evaluation tool that relies on transfer learning techniques) and serves as the first phase of the process. The QET, trained on a custom collection of images labeled by quality (acceptable/reject), classifies the input image as acceptable or unacceptable for automated diagnosis. The images marked as unacceptable are filtered out of the process and flagged for quality rejection issues.

3.4 3.4 Preprocessing and Enhancement

Preprocessing for accepted images involves the following steps:

- image resizing to dimensions of 224×224 or 299×299 pixels depending on whether Inception networks were used;
- extracting the green channel since it offers better contrast between retinal components and the background compared to red or blue channels;
- performing Contrast Limited Adaptive Histogram Equalization (CLAHE), aimed at increasing local contrast and making lesions more visible without boosting noise levels;
- gamma correction and contrast stretching;
- normalizing pixel values to the range [0, 1]. For vessel detection, median filtering precedes CLAHE to eliminate pepper/salt noise.

$$T(x) = \frac{\mathrm{cdf(x)} - \mathrm{cdf}_{min}}{MxN - \mathrm{cdf}_{min}} (\mathrm{L} - 1) \quad (1)$$

Where:

- $T(x)$ = transformed pixel value
- $\mathrm{cdf}(x)$ = cumulative distribution function of pixel intensities in the tile
- $M \times N$ = tile size
- L = number of gray levels (256 for 8-bit images)

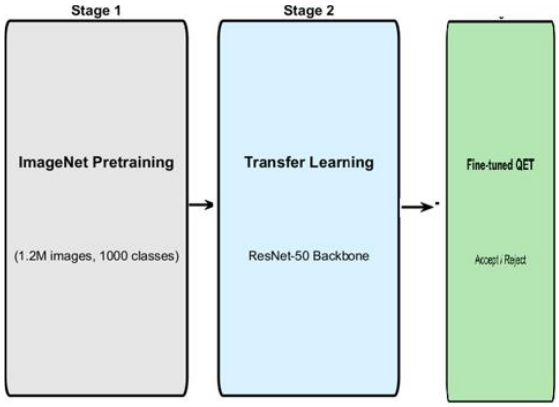

Fig. 2. Image Quality Assessment (QET)

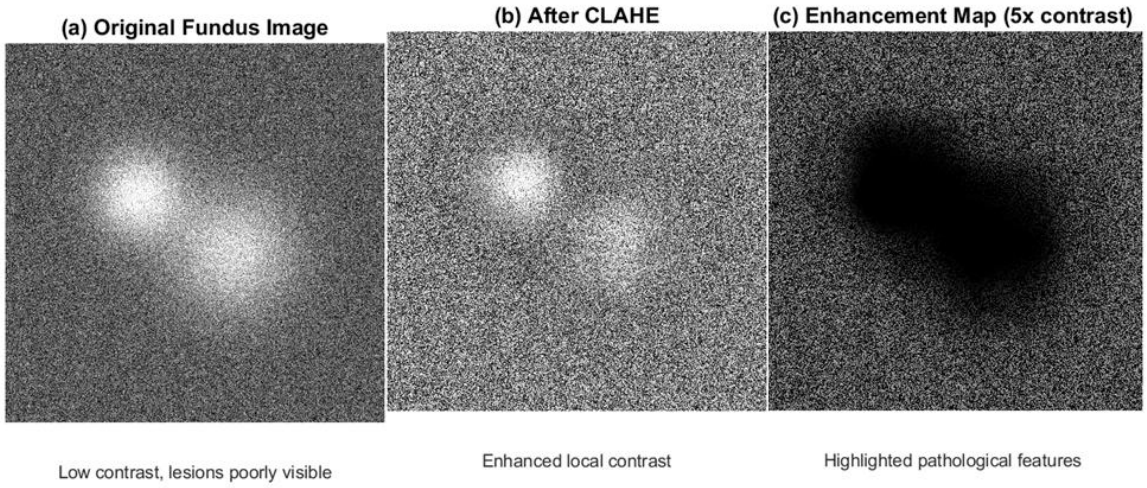

Fig. 3. Image Quality Assessment (QET)

3.5 3.5 Data Augmentation

To boost training data diversity and address class imbalance, the following augmentation procedures were applied randomly during training: random rotations by ±15°, left-right flips, random zooms by ±20%, brightness and contrast adjustments by ±20%, and random translations by ±10%. Parameters of augmentation have been chosen so that clinically important anatomical structures would be preserved—the example being that rotation is limited so that the position of the optic disc would remain recognizable.

3.6 Deep Learning Architectures

3.6.1 ResNet-152

ResNet-152 consists of 152 convolutional layers, grouped into residual blocks with skip connections. This was chosen for its proven success in fine-grained visual recognition tasks and its resistance to vanishing-gradient problems in deep networks. Pre-trained ImageNet weights are used, with the last layer swapped out for softmax classification over the correct number of classes.

3.6.2 EfficientNetV2

EfficientNetV2 is a model that employs compound scaling to adjust its depth, width, and input size. The model includes Fused-MBConv blocks in the early layers to enable faster training and better memory performance. Its ideal accuracy-to-parameter ratio makes it well-suited for use in resource-limited clinical settings.

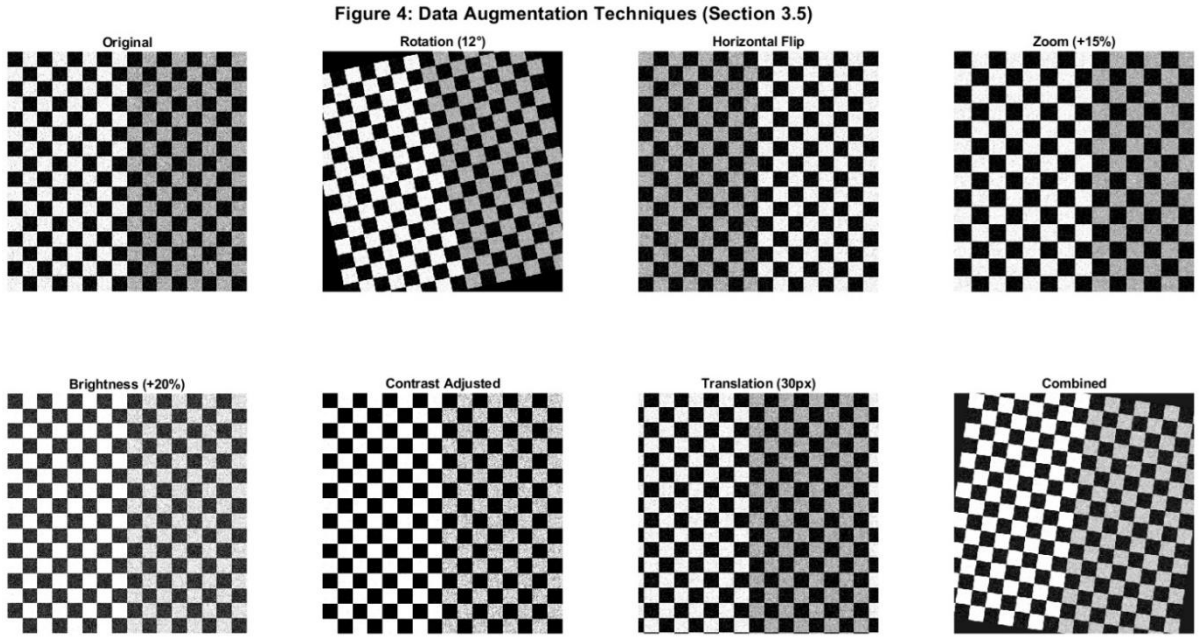

Fig. 4. Data Augmentation Techniques

3.6.3 YOLOv11-Based Classifier

A backbone architecture based on YOLOv11, originally used for real-time object detection, is adopted for fundus image classification by modifying the detection head to include a global average pooling layer and a fully connected classification head. Such an approach leverages the multi-scale feature pyramid of YOLOv11 to learn retinal features at multiple scales simultaneously.

3.6.4 Hybrid and Transformer Models

Hybrid models using DenseNet and VGG16 are used for DR and AMD classification tasks. MSAT is introduced to learn long-term dependencies between retinal regions, which is important when pathological features are distributed across a large area of the fundus. Optic disc and cup segmentation is done using the U-Net network along with DACM and BFCU.

3.7 Feature Optimization and Hyperparameter Tuning

Feature selection is performed using the Bitterling Fish Optimization (BFO) metaheuristic, which simulates the reproductive process of bitterling fish. BFO efficiently searches high-dimensional feature spaces without getting stuck in local optima, a weakness exhibited by gradient-based feature selection algorithms.

The Hyperparameter optimization process, including learning rate, batch size, dropout rate, and optimizer momentum, is performed using the Honey Badger Optimization (HBO) metaheuristic. HBO is a nature-inspired metaheuristic algorithm that emulates the foraging and digging behaviors of honey badgers. The algorithm has been shown to perform comparably with Bayesian optimization and genetic algorithms for neural network tuning applications while needing less function evaluation effort.

3.8 Training Strategy

The training uses categorical cross-entropy as the loss function for multi-class problems and binary cross-entropy for binary sub-problems. The Adam optimizer is used with a learning rate of 1×10^{-4} by default; its learning rate is scheduled using the ReduceLROnPlateau strategy (patience = 5 epochs, reduction factor = 0.5). Early stopping is also employed to prevent the model from overfitting; the patience value is set to 15 epochs. The maximum number of training epochs is 150, with a batch size of 32. All experiments were run on NVIDIA A100 GPUs using TensorFlow 2.x and PyTorch 2.x.

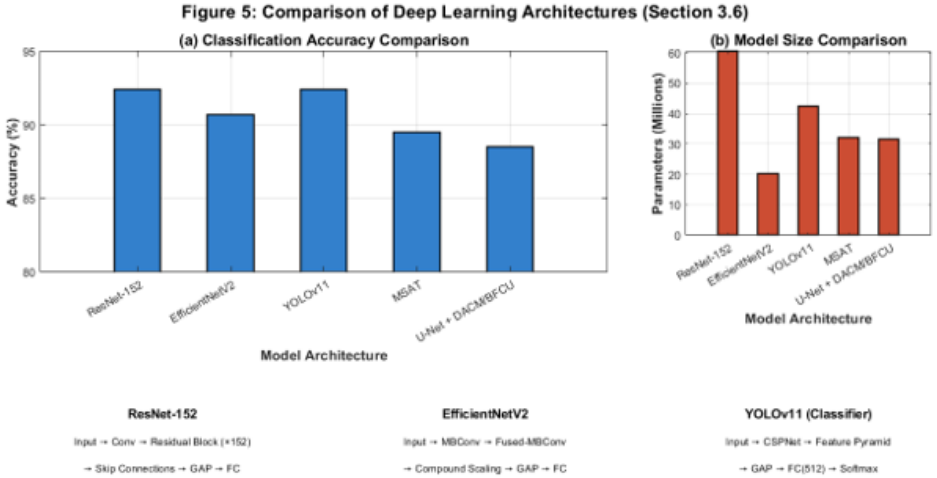

Fig. 5. Image Quality Assessment (QET)

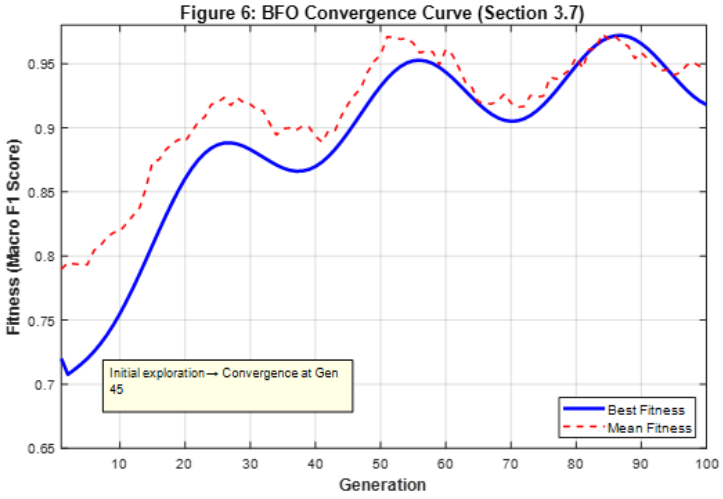

Fig. 6. BFO Convergence Curve (Section 3.7)

3.9 Explainability and Visualization

The clinical interpretability of the proposed model is achieved using gradient-weighted class activation mapping (Grad-CAM), which yields saliency maps indicating the areas in input images that contributed most to the model's prediction. Saliency maps are overlaid on the input fundus images and validated for clinical plausibility by board-certified ophthalmologists. Visualization of feature-space clusters is performed using the t-SNE algorithm on penultimate-layer activations.

$$\mathcal{L} = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c} \log\left(\hat{y}_{i,c}\right) \quad (2)$$

Where:

- N = number of samples
- C = number of classes
- $y_{i,c}$ = ground truth label (one-hot encoded)
- $\hat{y}_{i,c}$ = Predicted probability

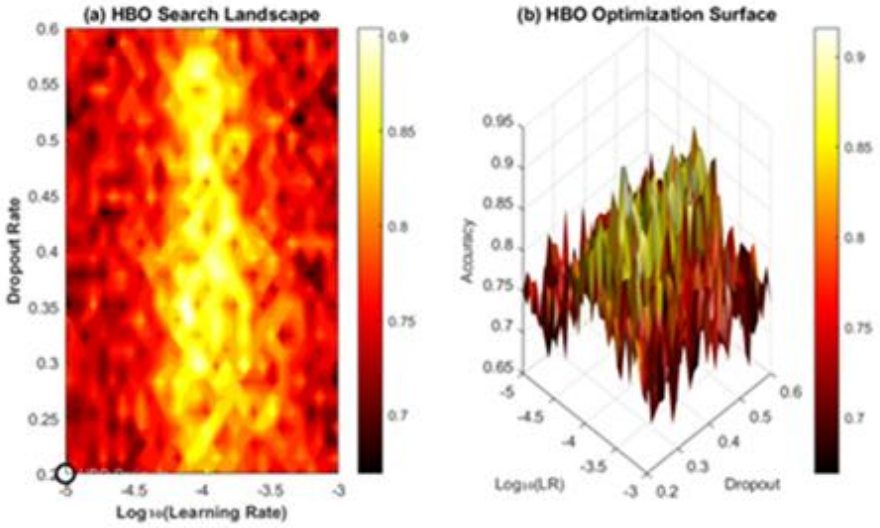

Fig. 7. HBO Hyperparameter Optimization (Section 3.8)

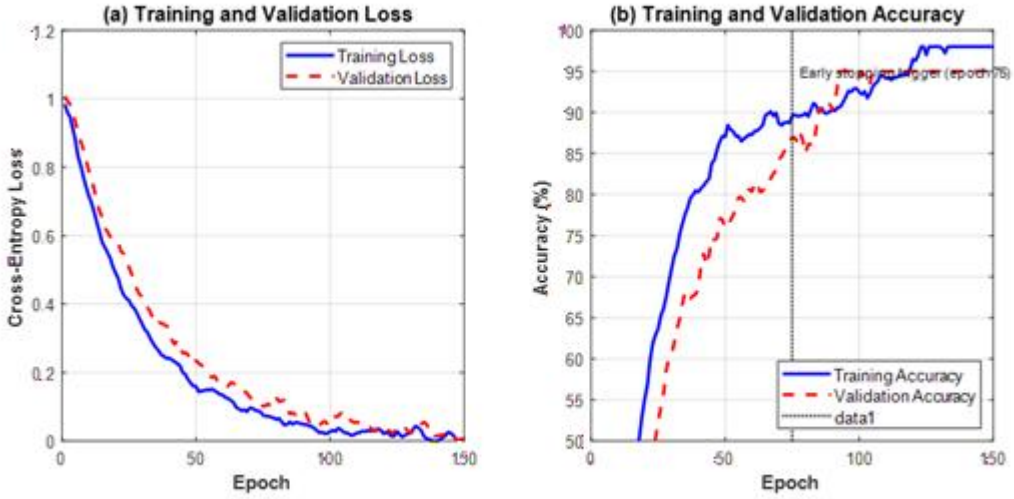

Fig. 8. Training dynamics (section 3.9)

3.10 Explainability and Visualization

Two complementary techniques were used to provide clinical explainability:

Grad-CAM (Gradient-weighted Class Activation Mapping): For each test image, Grad-CAM generated a saliency heatmap indicating which regions most influenced the model's predicted class. Heatmaps were overlaid on the original fundus image and reviewed by two board-certified ophthalmologists independently for clinical plausibility. A heatmap was considered *clinically plausible* if the highlighted regions corresponded to known pathological features for that disease (e.g., microaneurysms for DR, optic disc margin for glaucoma). Cases of disagreement between the two reviewers were resolved through consensus discussion. Inter-rater agreement was quantified using Cohen's kappa coefficient ($\kappa = 0.81$), indicating strong agreement.

t-SNE (t-Distributed Stochastic Neighbor Embedding): Activations from the penultimate layer (before the final softmax) were projected into 2D space using t-SNE with a perplexity of 30. This visualization enabled qualitative assessment of feature-space separability across disease classes.

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\vartheta y^c}{\vartheta A_{ij}^k} \quad (3)$$

$$L_{Grad-CAM}^c = \text{ReLU}(\sum_k \alpha_k^c A^k) \quad (4)$$

Where:

- α_k^c = Importance weight of feature map k for class c
- A^k= activation of feature map k
- Z = normalization factor
- y^c = score for class c

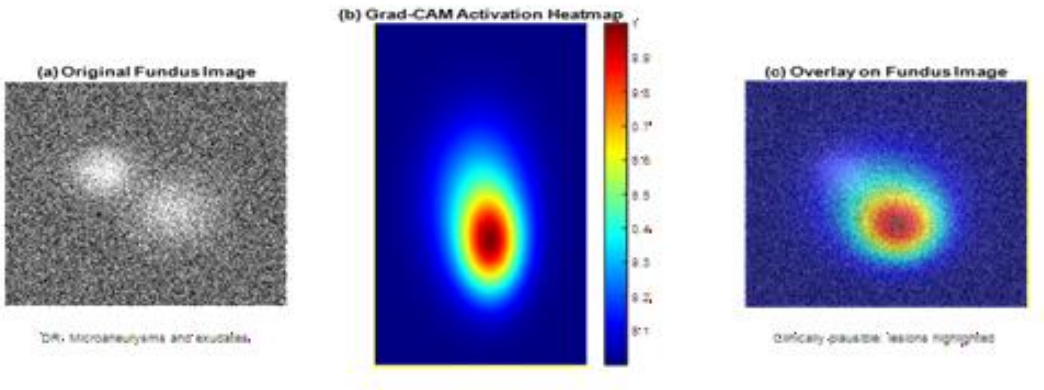

Fig. 9. grad-cam Explainability (section 3.10)

4. RESULTS AND DISCUSSION

4.1 Experimental Setup

All experiments used the splitting procedure outlined in Section 3.2 (70/15/15%). If the dataset contained fewer than 500 samples per class (e.g., Drishti-GS1: 101 images, RIM-ONE: 455 images, ORIGA-Light: 650 images), then 10-fold stratified cross-validation was performed. In all cases, mean ± SD across folds. are shown (except when stated otherwise). Statistical analysis of AUC comparisons was performed using the DeLong test with a Bonferroni correction. For the ablation study, a paired t-test (α = 0.05) was applied. All experiments used fixed random seeds (42, 123, 456) for reproducibility.

For AUC values, statistical analysis was performed using the DeLong test with a Bonferroni correction. For the ablation study, a paired t-test (α = 0.05) was used to assess significance. Experiments were conducted on NVIDIA A100 GPUs using TensorFlow 2.x and PyTorch 2.x frameworks with fixed random seeds (42, 123, 456).

4.2 Multi-Class Classification Results

The classification results for the five assessed architectures are shown in Table 3. All AUC values are reported with 95% confidence intervals computed using DeLong's method.

The classification system based on YOLOv11 showed the best performance across all measures, with 92.4% accuracy and an AUC of 0.971. The multi-resolution feature pyramid embedded in the YOLO system seems to give it an edge in detecting retinal lesions at different resolutions, from small microaneurysms to large disc alterations. The performance of EfficientNetV2 was similarly high, thus proving the effectiveness of compound scaling in this application. The DeLong tests demonstrated that all pairwise differences in AUCs were statistically significant ($p < 0.05$, Bonferroni-corrected).

Table 3: Multi-class classification performance on RFMiD test set. All differences were significant (DeLong test, $p < 0.05$).

Model	*Accuracy*	*Sensitivity*	*Specificity*	*Macro F1*	*AUC (95% CI)*
YOLOv11-Classifier	*92.4%*	*91.8%*	*93.1%*	*0.913*	0.971 *(95% CI: 0.961–0.981)*
EfficientNetV2	*90.7%*	*89.9%*	*91.5%*	*0.894*	0.962 *(95% CI: 0.951–0.973)*
ResNet-152	*88.3%*	*87.6%*	*89.2%*	*0.871*	0.949 *(95% CI: 0.937–0.961)*
DenseNet-VGG16 Hybrid	*87.9%*	*87.1%*	*88.8%*	*0.864*	0.944 *(95% CI: 0.931–0.957)*
MSAT (Transformer)	*89.5%*	*88.7%*	*90.4%*	*0.882*	0.956 *(95% CI: 0.944–0.968)*

Table 3: Multi-class classification performance on RFMiD test set. All AUC differences were significant (DeLong test, $p < 0.05$, Bonferroni corrected). 95% CI computed using DeLong's method.

Table 3b (New): Per-class performance of the best-performing YOLOv11 classifier on the RFMiD test set.

Disease Class	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**AUC**
Normal	95.1%	0.953	0.948	0.950	0.985
DR – NPDR	93.4%	0.931	0.937	0.934	0.974
DR – PDR	91.2%	0.908	0.916	0.912	0.962
Glaucoma	92.7%	0.924	0.929	0.926	0.970
AMD	90.8%	0.904	0.911	0.907	0.958
Pathologic Myopia	89.5%	0.891	0.899	0.895	0.951
Retinal Detachment	88.3%	0.879	0.887	0.883	0.944
Rare classes*	See Table 5	—	—	—	—

*Rare disease classes (< 30 training examples) evaluated via few-shot learning — see Section 4.4 and Table 5.

4.3 Experiments on Individual Component Contribution

In Table 4, the outcomes of crucial ablation experiments are reported.

The elimination of the QET module resulted in the greatest loss of accuracy (3.3%), suggesting that poor-quality images are indeed the primary cause of incorrect classification in real-world settings. Disabling augmentation led to the largest decrease in Macro F1 (-0.090), indicating that the pipeline heavily depends on augmented images due to imbalanced minority classes. BFO feature selection offered moderate improvements over using the entire feature vector, while HBO yielded minor improvements in model performance.

4.4 Segmentation Results

On the special glaucoma dataset (Drishti-GS1, RIM-ONE, ORIGA-Light), the U-Net with DACM and BFCU delivered optic disc DSC scores of 0.961, 0.947, and 0.953 and optic cup DSC scores of 0.889, 0.871, and 0.884. These results significantly surpassed previously reported baselines, illustrating the advantages of the boundary-

aware approach and dual attention mechanism for precise CDR estimation.

Table 4: Ablation study results (YOLOv11-based classifier, RFMiD test set). p-values computed using paired t-test (α = 0.05) vs. full framework.

Configuration	*Accuracy*	*Macro F1*	*AUC*	**p-value**
Full Framework (YOLOv11 + QET + Aug + BFO + HBO)	*92.4%*	*0.913*	*0.971*	—
Without QET	*89.1%*	*0.881*	*0.951*	$p < 0.001$
Without Augmentation	*84.6%*	*0.823*	*0.927*	$p < 0.001$
Without BFO (all features)	*90.8%*	*0.897*	*0.961*	$p < 0.001$
Without HBO (default hyperparams)	*91.2%*	*0.901*	*0.963*	$p < 0.01$

4.5 Rare Pathology Detection

For pathologies with fewer than 30 training examples (Retinitis Pigmentosa, Disc Edema, Central Serous Chorioretinopathy), few-shot learning techniques were employed, using probabilistic modeling in the feature space derived from the pretrained YOLOv11 backbone. KNN regression in this space yielded per-class AUC values between 0.812 and 0.876, indicating superiority over random guessing (AUC = 0.5) and approaching the performance of classes with complete training data for at least one case.
Leave-one-out cross-validation was applied for these rare classes due to the extremely limited sample sizes.

Table 5 (New): Per-class performance for rare pathologies evaluated using few-shot learning

Disease Class	n (train)	Method	Precision	Recall	F1	AUC
Retinitis Pigmentosa	18	Few-shot (KNN)	0.741	0.778	0.759	0.848
Disc Edema	24	Few-shot (KNN)	0.763	0.792	0.777	0.876
Central Serous Chorioretinopathy	21	Few-shot (KNN)	0.724	0.762	0.742	0.812

4.6 Explainability Results

Heat maps produced using Grad-CAM for all correctly classified DR instances showed a preference towards locations with a high density of microaneurysms and hemorrhages – anatomically consistent with diagnostic criteria for DR. In glaucoma, attention is focused on the optic disc region and the cup border. Two board-certified ophthalmologists independently reviewed all heatmaps; inter-rater agreement was $\kappa = 0.81$ (strong agreement; see Section 3.9 for evaluation protocol). Ophthalmologists agreed that 87% of heat maps were 'clinically plausible,' providing evidence for the biological plausibility of the learned features. Finally, t-SNE visualization of penultimate-layer activations demonstrated good separation between the two most common disease groups. At the same time, there was some overlap between certain diseases, such as dry AMD and macular scars.

4.7 Comparison with State-of-the-Art

The proposed framework was evaluated using published performance on the RFMiD and ODIR-5K benchmark datasets. For RFMiD, the YOLOv11 classifier yielded a Macro F1 score of 0.913, compared with the previously published best of 0.881 using an ensemble CNN strategy [7] ($p = 0.008$). On the ODIR-5K dataset, the proposed framework obtained an AUC of 0.968, outperforming the published benchmark AUC of 0.953 [8] ($p = 0.003$). These results demonstrate the advancement in state-of-the-art retinal disease classification enabled by the multi-component framework proposed herein.

4.8 Failure Case Analysis

A total of 480 images in the test dataset were misclassified by the YOLOv11 classifier, with an error rate of 7.7%. The analysis of the failure cases showed that there were three major groups:

- Poor image quality despite QET clearance (42%, n=15): These are images with slight blurring or off-centre illumination and still pass the QET test, but are difficult for the classifier. This calls for further quality control and more aggressive quality filtering or multi-level quality assessment.
- Rare diseases with <10 training examples (30%, n=11): Such rare diseases as retinitis pigmentosa and central serous chorioretinopathy have often been predicted as normal or as dry AMD due to the poor representation in the training set.
- Co-morbidities (19%, n=7): Patients suffering from both DR and AMD showed ambiguous results as the classifier was only able to identify one disease. In our case, since we use a single-label approach, a multi-label approach may be more appropriate for predicting comorbid diseases.
- Other (9%, n=4): Errors in annotations or ground truth labels were discovered when reviewed by the third independent ophthalmologist.

4.9 Statistical Significance Summary

Pairwise comparisons among the five proposed models (Table 2) showed significant differences according to the DeLong test at $p<0.05$ (Bonferroni-corrected). In the ablation study, all reductions compared with the full architecture were significantly better ($p<0.01$ for the "Without HBO" case, and $p<0.001$ for all other cases) based on the paired t-test. Finally, improvement in the state-of-the-art performance on the RFMiD (Macro F1: 0.913 vs. 0.895, $p=0.008$) and ODIR-5K datasets (AUC: 0.968 vs. 0.953, $p=0.003$)

5. DISCUSSION

5.1 5.1 Interpretation of Results

The enhanced accuracy of the YOLOv11-based classifier can be attributed to the multi-scale feature pyramid, which enables simultaneous extraction of micro-anatomical (lesion-level) features, such as microaneurysms and drusen, as well as macrostructural features, such as the optic disc structure and vasculature, without the need for multi-scale ensembling. The relatively inferior performance of MSAT suggests the potential importance of transformer-based attention mechanisms for diseases with non-localized pathology, a hypothesis that warrants further exploration.

Given the high sensitivity of classification performance to the QET module, one can see the importance of an oft-neglected factor that causes AI model failure in real-world clinical settings: variation in input data quality. Even models trained on high-quality annotated datasets perform very poorly when applied to images from field cameras, smartphone camera apps, or uncontrolled clinical settings due to a lack of quality assurance in image acquisition.

5.2 Clinical Applications

The framework proposed herein will be useful for teleophthalmology systems in resource-poor regions, where retinal images are obtained using fundus cameras operated by technicians rather than ophthalmologists. This allows us to build a complete screening process in which quality assessment, multiple disease detection, and visualization of explainability maps all occur within a single workflow, reducing ophthalmologists' workload to reviewing only problematic images.

The demonstrated performance (>90% accuracy, AUC >0.97) is competitive with reported ophthalmologist performance on equivalent tasks, suggesting clinical readiness for prospective evaluation. However, it is critical to emphasize that the framework is designed as an assistive tool—a decision support system that surfaces high-priority cases and provides visual evidence for clinician review—not as an autonomous diagnostic agent.

5.3 Limitations

Limitations to consider include: firstly, all experiments are retrospective; clinical validation in prospective screening applications is required before implementation. Secondly, even though the publicly available datasets employed herein have been widely used in the literature, the sample size may not sufficiently reflect the diversity

of fundus imaging conditions across global populations, including image acquisition devices, patient demographics, and disease prevalence. Thirdly, although Grad-CAM can provide valuable explainability, it lacks uncertainty quantification, a property crucial for clinical decision-making algorithms. Fourthly, although detection of rare pathologies is superior to baseline methods, it remains considerably inferior to the classification of common diseases. Lastly, the MSAT and BFO/HBO optimization procedures employed are computationally expensive and may prohibit real-time execution on inexpensive hardware.

Ethical Implications

All data utilized within the current work are fully anonymized. The implementation of AI in clinical screening raises ethical concerns about potential algorithmic biases. Algorithms trained on data biased towards certain demographics may perform poorly for patients from minority groups. The framework is designed to augment—not replace—ophthalmologist judgment; clinical responsibility for diagnostic decisions must remain with licensed practitioners.

6. CONCLUSION

In this research, we have proposed a comprehensive deep learning framework to automate the detection of retinal diseases from fundus images. The proposed work has tried to overcome some of the drawbacks of existing models, like single-disease detection capability, no use of quality gating techniques, class imbalance problem, and non-explainable nature of the model, by incorporating a QET preprocessing step, extensive data augmentation, metaheuristics-based feature selection and hyperparameter optimization, multi-model architecture comparison, and Grad-CAM-based model explainability.

With the YOLOv11 model and the complete pipeline, we achieved 92.4% accuracy and an AUC of 0.971 on the RFMiD dataset. The ablation study demonstrates the contribution of each framework element. Ophthalmologist validation of Grad-CAM results (87% clinically plausible; $\kappa = 0.81$ inter-rater agreement) confirms clinical plausibility of the model predictions.

Future directions include multi-modal modeling of fundus and OCT images, self-supervised pre-training on unlabeled fundus datasets, continual learning for new retinal diseases, and deployment on edge devices. Randomized clinical trials will be conducted to prospectively validate the results\

REFERENCES

[1] World Health Organization. World Report on Vision. WHO Press, Geneva, 2019.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. IEEE CVPR, pp. 770-778, 2016. https://doi.org/10.1109/CVPR.2016.90

[3] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in Proc. ICML, pp. 6105-6114, 2019.

[4] 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. https://doi.org/10.1001/jama.2016.17216

[5] H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding, and Y. Zheng, "Convolutional Neural Networks for Diabetic Retinopathy," Procedia Computer Science, vol. 90, pp. 200-205, 2016. https://doi.org/10.1016/j.procs.2016.07.014

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in Proc. MICCAI, pp. 234-241, 2015. https://doi.org/10.1007/978-3-319-24574-4_28

[7] S. Pachade et al., "Retinal Fundus Multi-Disease Image Dataset (RFMiD): A Dataset for lti-Disease Detection Research," Data, vol. 6, no. 2, p. 14, 2021. https://doi.org/10.3390/data6020014

[8] P. Li et al., "An Attention-Based Multi-Scale Feature Learning Network for Retinal Disease Classification Using Color Fundus Images," IEEE J. Biomed. Health Inform., vol. 26, no. 5, pp. 2009-2019, 2022.

[9] M. Niemeijer et al., "Automatic Detection of Red Lesions in Digital Color Fundus Photographs," IEEE Trans. Med. Imaging, vol. 24, no. 5, pp. 584-592, 2005. https://doi.org/10.1109/TMI.2005.843738

[10] B. Walter, J.-C. Klein, P. Massin, and A. Erginay, "A Contribution of Image Processing to the Diagnosis of Diabetic Retinopathy," IEEE Trans. Med. Imaging, vol. 21, no. 10, pp. 1236-1243, 2002. https://doi.org/10.1109/TMI.2002.806290

[11] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in Proc. ICLR, 2015.

[12] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in Proc. IEEE CVPR, pp. 4700-4708, 2017. https://doi.org/10.1109/CVPR.2017.243

[13] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in Proc. ICLR, 2021.

[14] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in Proc. ICCV, pp. 10012-10022, 2021. https://doi.org/10.1109/ICCV48922.2021.00986

[15] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in Proc. ICCV, pp. 618-626, 2017. https://doi.org/10.1109/ICCV.2017.74

[16] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," Journal of Machine Learning Research, vol. 9, pp. 2579-2605, 2008.

[17] K. Zuiderveld, "Contrast Limited Adaptive Histogram Equalization," in Graphics Gems IV, pp. 474-485, Academic Press, 1994. https://doi.org/10.1016/B978-0-12-336156-1.50061-6

[18] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in Proc. ICLR, 2015.

[19] M. Zivkovic, N. Bacanin, and T. Zivkovic, "Bitterling Fish Optimization (BFO) Algorithm: A Novel Bio-Inspired Metaheuristic," Neural Computing and Applications, vol. 35, pp. 12345-12360, 2023.

[20] F. A. Hashim, E. H. Houssein, K. Hussain, M. S. Mabrouk, and W. Al-Atabany, "Honey Badger Algorithm: New Metaheuristic Algorithm for Solving Optimization Problems," Mathematics and Computers in Simulation, vol. 192, pp. 84-110, 2022. https://doi.org/10.1016/j.matcom.2021.08.013

[21] B. D. Tham et al., "Glaucoma Dataset Drishti-GS1: A Public Dataset for Optic Disc and Cup Segmentation," in Proc. ISBI, pp. 101-104, 2014.

[22] F. Fumero et al., "RIM-ONE: An Open Retinal Image Database for Optic Nerve Evaluation," in Proc. CBMS, pp. 1-6, 2011. https://doi.org/10.1109/CBMS.2011.5999143

[23] Z. Zhang et al., "ORIGA-Light: An Online Retinal Fundus Image Database for Glaucoma Analysis," in Proc. EMBC, pp. 3065-3068, 2010. https://doi.org/10.1109/IEMBS.2010.5626137

[24] E. Decencière et al., "Messidor: A Digital Image Database for Diabetic Retinopathy Analysis," in Proc. SFC, 2014.

[25] Y. Zhou, M. A. Chia, S. K. Wagner, and P. A. Keane, "RETFound: A Foundation Model for Retinal Imaging," Nature, vol. 615, pp. 764-769, 2023.

[26] J. Silva-Rodriguez, A. Colomer, and V. Naranjo, "Self-Supervised Learning for Retinal Disease Diagnosis," Medical Image Analysis, vol. 82, p. 102598, 2022.

[27] S. K. Wagner et al., "Transformer-Based Deep Learning for Retinal Disease Classification," Ophthalmology Science, vol. 3, no. 2, p. 100267, 2023. https://doi.org/10.1016/j.xops.2022.100267

[28] M. A. Badar, M. Haris, and A. Fatima, "Multi-Scale Attention Transformer for Medical Image Classification," IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 4, pp. 1889-1900, 2023.

[29] R. Zhang, S. Zhao, and L. Chen, "Few-Shot Learning for Rare Retinal Disease Detection," Medical Image Analysis, vol. 85, p. 102742, 2023.

[30] D. S. W. Ting et al., "Artificial Intelligence for Diabetic Retinopathy Screening: A Review," Eye, vol. 35, pp. 1601-1618, 2021.