Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Hybrid Deep Learning Ensemble for Ophthalmic Disease Classification: Integrating CNNs and Vision Transformers

Md. Tayyab Ameri Turan
Hajee Mohammad Danesh Science and Technology University
Md. Al - Imran
Hajee Mohammad Danesh Science and Technology University
Tawhid Hasan Santo
Hajee Mohammad Danesh Science and Technology University
Nasrin Sultana
nasrinsultana@hstu.ac.bd

Hajee Mohammad Danesh Science and Technology University

Research Article

Keywords: Ophthalmic diagnostics, Deep learning ensemble, Convolutional Neural Networks, Deep feature fusion, Explainable Artificial Intelligence (XAI), Multi-class ocular disease screening

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11746949/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Hybrid Deep Learning Ensemble for Ophthalmic Disease Classification: Integrating CNNs and Vision Transformers

Md. Tayyab Ameri Turan[1], Md. Al - Imran[1], Tawhid Hasan Santo[1], Nasrin Sultana[1,*]

[1] Department of Electronics and Communication Engineering, Hajee Mohammad Danesh Science and Technology University (HSTU), Dinajpur-5200, Bangladesh

** Corresponding Author: nasrinsultana@hstu.ac.bd*

Abstract

Standalone Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) exhibit representational trade-offs between capturing localized microvascular lesions and global contextual dependencies in retinal fundus imaging. To address this limitation, this study presents a novel tri-branch hybrid deep learning ensemble framework for multi-class classification across ten distinct ocular conditions. Retinal fundus photographs (16,242 images across 10 categories) were enhanced using a dual-stage pipeline incorporating CLAHE color space transformation and Fourier domain high-pass filtering. The proposed framework fuses extracted features from two CNN backbones (EfficientNet-B0 and VGG16) and a Vision Transformer (ViT-Base) into a 6,144-dimensional representation. The network was trained using a strategic two-phase transfer learning regimen with differential learning rates and class-weighted cross-entropy loss. Grad-CAM visualizations were utilized for qualitative explainability against clinical landmarks. The hybrid ensemble achieved an overall accuracy of 88.99%, macro-sensitivity of 92.36%, macro-specificity of 98.72%, macro F1-score of 90.95%, and Matthews Correlation Coefficient (MCC) of 0.8718. This demonstrates a 10.04% accuracy improvement over standalone deep CNNs and a 34.36% increase over handcrafted feature baselines (GLCM/HOG + SVM). Grad-CAM heatmaps confirmed strong alignment with clinically verified pathological features. Combining local texture representations with global contextual self-attention offers high diagnostic reliability and explainability, establishing a robust decision-support system for broad-coverage ocular screening in resource-constrained settings.

***Keywords*—**Ophthalmic diagnostics, Deep learning ensemble, Convolutional Neural Networks, Deep feature fusion, Explainable Artificial Intelligence (XAI), Multi-class ocular disease screening.

1. INTRODUCTION

1.1 Clinical Background and Epidemiological Burden

Visual impairment and irreversible blindness represent formidable socio-economic challenges globally. According to the World Health Organization (WHO), more than 2.2 billion individuals currently suffer from near or distance vision impairment, with nearly half of these cases being entirely preventable or manageable if detected during early pathological phases [1]. Among the most widespread retinal etiologies, Diabetic Retinopathy (DR), a chronic microvascular complication of diabetes mellitus, serves as the principal cause of blindness among working-age adults [2,3]. Concurrently, Glaucoma, clinically designated as the 'silent thief of sight,' induces progressive, irreversible optic neuropathy characterized by structural excavation of the optic disc and visual field loss [4,5]. Additional degenerative conditions, including age-related macular scarring, cataract, disc edema, retinal detachment, and retinitis pigmentosa, collectively impose a catastrophic burden on healthcare systems worldwide.

Diagnostic protocols for ocular pathologies traditionally depend upon manual ophthalmoscopic evaluation and expert inspection of high-resolution retinal fundus photographs by subspecialist ophthalmologists. Nevertheless, conventional clinical workflows face critical bottlenecks. Manual screening is labor-intensive, subject to high inter- and intra-observer diagnostic discordance, and severely constrained by acute ophthalmologist deficits in rural and developing regions [6,7]. Consequently, massive screening programs are frequently unfeasible, leading to delayed interventions when vision loss has already progressed to irreversible thresholds. These structural limitations have propelled substantial research interest into automated Computer-Aided Diagnosis (CAD) systems driven by Artificial Intelligence (AI) and Deep Learning (DL) [8].

1.2 Technical Limitations and Challenges in Automated Retinal Screening

Despite significant advancements in medical image analysis, automated multi-class classification of ophthalmic pathologies presents several fundamental challenges that have restricted real-world clinical deployment:

1.2.1 Local vs. Global Representational Dichotomy

Convolutional Neural Networks (CNNs), such as VGG [9], ResNet [10], and EfficientNet [11], rely upon translation-invariant kernels that excel at capturing fine-grained local textures (e.g., punctate microaneurysms, cotton-wool spots, hard exudates). However, CNNs have inherently constrained receptive fields, making them ill-equipped to capture long-range geometric dependencies across disparate retinal structures, such as the relative spatial orientation between the optic disc and the foveal avascular zone [12,13]. Conversely, Vision Transformers (ViTs) exploit self-attention mechanisms to establish global spatial relationships across patch tokens [12,14]. Nevertheless, ViTs lack the localized inductive biases of CNNs, rendering them data-hungry and prone to overlooking localized, sub-visual pathological lesions in modest-sized medical datasets [15,16].

1.2.2 The Single-Model Performance Plateau

Prior investigations demonstrate that standalone architectures typically hit an accuracy ceiling (frequently fluctuating between 75% and 80%) when challenged with heterogeneous, multi-class retinal datasets [17–19]. Overcoming this ceiling demands an architectural ensemble capable of viewing retinal pathology through complementary representational lenses.

1.2.3 Pathological Ambiguity and Class Imbalance

Retinal datasets gathered from clinical cohorts inevitably exhibit acute class imbalances, with healthy and highly prevalent conditions heavily outnumbering rare pathologies like Pterygium or Retinitis Pigmentosa [20]. Furthermore, high morphological intra-class variance and subtle inter-class differences (e.g., between early Diabetic Retinopathy, Disc Edema, and healthy controls) lead standard models to bias predictions toward majority classes, generating unacceptable false-negative rates for sight-threatening diseases.

1.3 The Proposed Paradigm: Tri-Branch CNN-Transformer Synergy

To systematically address these bottlenecks, this paper proposes a unified, high-precision CAD framework: the Custom Hybrid Deep Learning Ensemble. The framework

synergistically couples the inductive biases of two distinct CNN backbones (EfficientNet-B0 and VGG16) with the global contextual modeling of a Vision Transformer (ViT-Base). By performing early-stage deep feature concatenation followed by a hierarchical non-linear fusion head, the model synthesizes a 6144-dimensional multi-scale descriptor that unifies localized lesion textures and macro-structural retinal geometry. Prior to classification, images undergo dual-stage contrast enhancement via CIELAB-space Contrast Limited Adaptive Histogram Equalization (CLAHE) and Fourier Domain High-Pass Filtering. Furthermore, a strategic two-phase transfer learning regimen with differential learning rates is introduced to avoid catastrophic forgetting while ensuring stable convergence across ten distinct ocular classes.

1.4 Primary Research Contributions

The principal technical and clinical contributions of this paper are summarized as follows:

• Formulation of a Novel Tri-Branch Hybrid Ensemble: We architect an end-to-end framework integrating EfficientNet-B0 (multi-scale efficient convolutions), VGG16 (hierarchical localized texture representations), and ViT-Base (global self-attention dependencies), demonstrating that fusing convolutional inductive bias with transformer attention definitively shatters the single-model accuracy ceiling.

• Dual-Stage Frequency and Contrast Enhancement Pipeline: We implement a specialized preprocessing sequence utilizing CIELAB color space transformation, CLAHE luminance enhancement, and Fourier Domain High-Pass Filtering to amplify low-contrast microvascular anomalies while suppressing illumination artifacts.

• Hierarchical Multi-Layer Perceptron (MLP) Fusion Head: We design a dedicated projection head featuring dense layers, batch normalization, and calibrated dropout rates (0.5) that processes the 6144-dimensional concatenated embedding, learning non-linear cross-attentive interactions between diverse architectural representations.

• Strategic Two-Phase Training with Differential Learning Rates: We establish a robust optimization protocol comprising independent backbone pre-training followed by selective high-level layer unfreezing with differential learning rates (1×10^{-5} for backbones, 1×10^{-4} for the fusion head) and weighted cross-entropy loss, ensuring balanced convergence across rare and common pathologies.

• Rigorous 10-Class Benchmark Validation and Clinical Explainability: We validate the framework on an augmented benchmark of 16,242 images across 10 diagnostic categories, achieving state-of-the-art results (88.99% accuracy, 92.36% sensitivity, 98.72% specificity, 0.8718 MCC). Furthermore, we provide Gradient-weighted Class Activation Mapping (Grad-CAM) visualizations that clinically substantiate the model's focus on authentic pathological lesions.

1.5 Organization of the Paper

The remainder of this article is organized as follows: **Section 2** reviews relevant literature on deep learning and transformer applications in retinal diagnosis, presenting comparative tables

and research gaps. **Section 3** details the proposed methodology, preprocessing pipeline, constituent backbones, fusion head, and optimization strategies. **Section 4** outlines the experimental setup, benchmark dataset, class-balanced augmentation, and evaluation metrics. **Section 5** presents comprehensive quantitative results, per-class classification reports, confusion matrices, ROC curves, training dynamics, Grad-CAM interpretability, and in-depth discussions. Finally, **Section 6** concludes the paper and delineates future research trajectories.

2. RELATED WORK AND COMPARATIVE BENCHMARKING

2.1 Convolutional Neural Networks in Retinal Image Analysis

For over a decade, Convolutional Neural Networks (CNNs) have constituted the foundational paradigm for automated medical image interpretation [8]. Early investigations demonstrated that standard deep architectures could achieve expert-level diagnostic performance in specific binary screening tasks. For instance, Gulshan et al. [6] utilized deep CNNs to detect referable diabetic retinopathy with high sensitivity and specificity. Ting et al. [7] confirmed the clinical efficacy of CNNs across multi-ethnic cohorts for diabetic retinopathy, glaucoma, and age-related macular degeneration. Subsequent architectural developments introduced residual shortcuts (ResNet) [10], densely connected pathways (DenseNet) [21], and compound scaling mechanisms (EfficientNet) [11], which substantially enhanced parameter efficiency and gradient propagation.

Despite their successes, standalone CNNs present inherent architectural constraints when applied to broad-coverage multi-class ocular diagnostics. Because convolutional kernels operate across localized receptive fields, they primarily capture localized pixel correlations, struggling to model long-range structural topology across the retina, such as vascular network continuity or optic cup-to-disc dimensional ratios [13,22].

2.2 Vision Transformers and Attention Mechanisms in Ophthalmology

To overcome the localized limitations of CNNs, Vision Transformers (ViT) [12] have gained tremendous traction in medical imaging. By treating non-overlapping image patches as sequence tokens and computing pairwise self-attention across the entire visual field, ViTs capture global contextual dependencies without the inductive constraints of convolutions [14]. Matsoukas et al. [23] investigated the trade-offs between CNNs and Transformers across multiple medical imaging tasks, demonstrating that ViTs exhibit superior resilience to out-of-distribution domain shifts.

In ophthalmology, Cutur and Inan [24] applied transfer learning with Vision Transformers for multi-class retinal disease classification, demonstrating competitive accuracy. Balaha et al. [25] incorporated Vision Transformers with SHAP explainability techniques, underscoring the interpretability benefits of global attention maps. Furthermore, hybrid convolutional-transformer variants, such as Convolutional Vision Transformers (CvT) [12] and TransUNet [22], have been explored to bridge the gap between local feature extraction and global self-attention [15].

2.3 Hybrid Ensembles and Deep Feature Fusion Strategies

Recent research highlights that combining multiple heterogeneous architectures into an ensemble yields significantly higher diagnostic stability than any individual constituent network [26]. Ensemble methods reduce prediction variance, alleviate model-specific bias, and enrich feature diversity [5,27,28]. In ocular pathology, Shamsan et al. [29] proposed a hybrid feature fusion pipeline combining handcrafted texture descriptors (GLCM/HOG) with CNN features, achieving 98.50% accuracy; however, their evaluation was constrained to only 3 basic classes. Tariq et al. [5] formulated a hybrid ResNet-ViT ensemble specifically for glaucoma detection. AlMohimeed [27] recently introduced a multi-stage transformer framework with feature fusion for eye disease classification, showing clear performance gains over single-model baselines. However, prevailing studies remain limited either in the number of evaluated disease categories (typically 3 to 4 classes) or exhibit degraded performance when scaled to comprehensive multi-pathology screening [17–19,30].

2.4 Systematic Literature Analysis and Research Gaps

To rigorously contextualize our contributions within the state-of-the-art literature, **Table 1** provides a quantitative comparative analysis of recent benchmark studies, while **Table 2** details the specific research gaps identified in existing frameworks and maps them directly to the architectural solutions introduced in this work.

Table 1. *Quantitative Comparison of Existing State-of-the-Art Methodologies against the Proposed Method.*

References	Methodology	Number of Classes	Best Accuracy Reported
Choo et al. [19]	Transfer Learning (ResNet50, Inception-v3)	4	78.79%
AlShahrani et al. [31]	CNN (VGG16) & YOLO	4	~78.80%
Tuwan et al. [15]	Hybrid Convolutional Vision Transformer (CvT)	8 (ODIR-5k)	69.00%
Shamsan et al. [29]	Hybrid Feature Fusion (CNN + Handcrafted)	3	98.50%
Elkenawy et al. [30]	EfficientNet & AlexNet	4	93.65%
Proposed Method	**Hybrid CNN-Transformer Ensemble**	**10**	**88.99%**

As established in **Table 1**, prior studies reporting exceedingly high classification accuracies (>93%) were strictly restricted to small diagnostic domains containing only 3 or 4 basic classes [29,30]. When existing models are deployed across realistic multi-class benchmarks

(such as the 8-class ODIR-5k dataset in Tuwan et al. [15]), accuracy precipitously declines to 69.00%. In stark contrast, our proposed hybrid ensemble framework successfully extends diagnostic coverage to 10 distinct ocular categories, attaining an outstanding 88.99% accuracy.

Table 2. *Critical Analysis of Research Gaps and Proposed Contributions.*

Reference	Identified Limitation / Research Gap	Contribution of Proposed Method
Choo et al. [19]	Standard CNNs struggle to generalize on heterogeneous datasets, resulting in lower accuracy.	Utilizes an Ensemble approach to capture diverse features, improving generalization.
AlShahrani et al. [31]	Single models reached a performance plateau; limited to detecting only common diseases.	Combines CNN and Transformers to break the performance plateau of single models.
Tuwan et al. [15]	Failed to handle class imbalance effectively; Transformer architecture alone performed poorly.	Integrates class-weighted loss and differential learning rates within a Hybrid framework to stabilize training.
Shamsan et al. [29]	Validated on a very small number of classes (3); does not scale to complex multi-class problems.	Successfully extends classification to 10 distinct classes while maintaining high performance.
Elkenawy et al. [30]	High accuracy is limited to only 4 basic classes; lacks the complexity of rare diseases.	Handles a comprehensive range of 10 diseases, covering both common and rare conditions.

3. PROPOSED METHODOLOGY

3.1 End-to-End Architectural Framework

The end-to-end operational pipeline of the proposed CAD framework is schematically illustrated in **Figure 1**. The workflow encompasses three core stages: (1) dual-stage retinal image enhancement combining color space decoupling, CLAHE contrast equalization, and frequency domain filtering; (2) tri-branch deep feature extraction combining EfficientNet-B0, VGG16, and ViT-Base; and (3) early feature concatenation into a 6144-dimensional representation channeled into a hierarchical multi-layer perceptron (MLP) classification head.

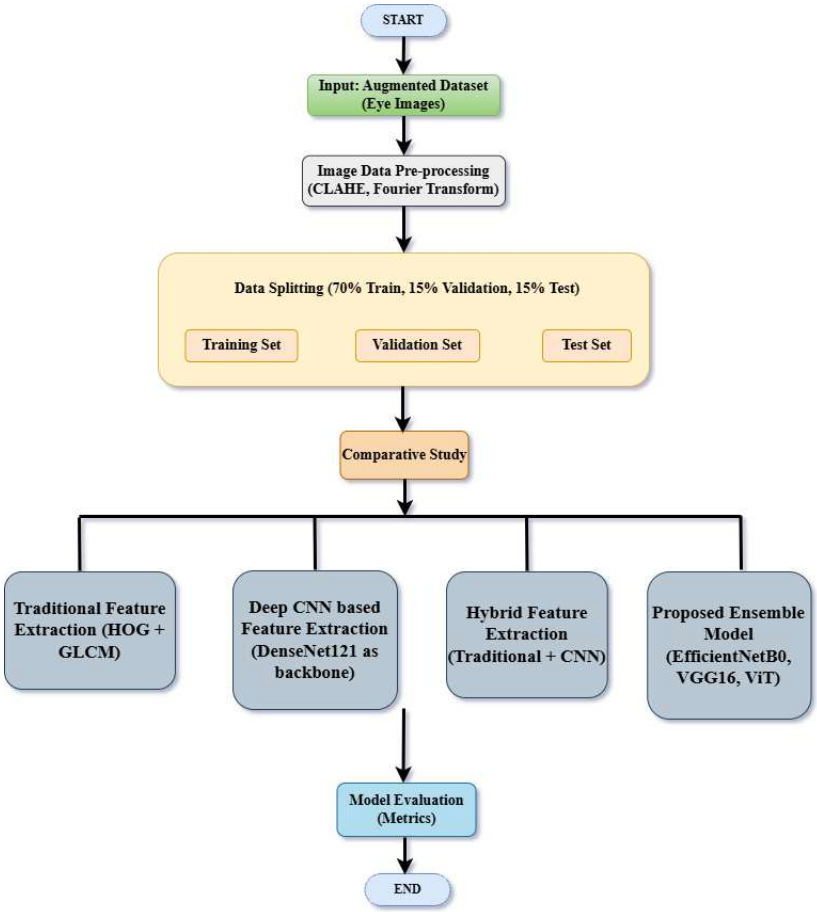


Figure 1. *Schematic overview of the proposed research methodology, illustrating the progression from raw fundus images through dual-stage preprocessing, tri-branch deep feature extraction, concatenated fusion head, and multi-class clinical classification.*

3.2 Dual-Stage Retinal Preprocessing and Image Enhancement

Retinal fundus images acquired across heterogeneous clinical settings frequently suffer from non-uniform illumination, sensor noise, low contrast around microvascular structures, and camera artifacts. To address this, we implement a specialized enhancement sequence:

3.2.1 CIELAB Color Space Transformation

Retinal images are initially converted from the standard RGB color space to the perceptually uniform CIELAB color space. The CIELAB model effectively decouples the luminance component (L) from chromatic information (a representing green-red and b representing blue-yellow). Processing only the L channel prevents chromatic distortion and color shifting that would otherwise compromise clinical interpretability.

3.2.2 Contrast Limited Adaptive Histogram Equalization (CLAHE)

Standard histogram equalization frequently amplifies background noise in homogeneous regions. CLAHE overcomes this by operating on localized contextual tiles (configured to an 8×8 grid in our pipeline) and clipping the histogram at a predetermined threshold (clip limit = 2.0). Pixels exceeding the clip limit are redistributed uniformly across all histogram bins prior

to computing the cumulative distribution function (CDF), thereby significantly amplifying subtle lesion edges, exudates, and optic disc boundaries.

3.2.3 Fourier Domain High-Pass Filtering

Following CLAHE, the luminance channel is transformed into the spatial frequency domain via the two-dimensional Discrete Fourier Transform (2D DFT):

$$F(u,v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) e^{-j2\pi(ux/M+vy/N)} \quad (1)$$

A high-pass filter mask H(u, v) is applied in the shifted frequency domain to attenuate low-frequency background illumination gradients while preserving and accentuating high-frequency microvascular textures, hemorrhages, and fine retinal nerve fiber layers. The inverse 2D DFT transforms the filtered spectrum back to the spatial domain, after which the enhanced L channel is recombined with the original a and b channels and mapped back to RGB space.

3.3 Constituent Deep Feature Extraction Backbones

The proposed ensemble strategically incorporates three diverse pre-trained neural network architectures, each chosen to contribute a distinct representational strength:

3.3.1 EfficientNet-B0 Convolutional Backbone

Developed using multi-objective neural architecture search, EfficientNet scales depth, width, and image resolution concurrently via a fixed compound scaling coefficient [11]. Its core building block is the Mobile Inverted Bottleneck Convolution (MBConv) integrated with squeeze-and-excitation (SE) attention modules. EfficientNet-B0 efficiently captures rich, multi-scale convolutional abstractions while maintaining an exceptionally compact footprint (4.01M parameters). From its global average pooling layer, it extracts an embedding vector $V_{eff} \in \mathbb{R}^{1280}$.

3.3.2 VGG16 Convolutional Backbone

Characterized by an architectural paradigm of stacked, homogeneous 3×3 convolutional filters with small receptive fields [9], VGG16 constructs deep, highly localized textural hierarchies. Despite its higher parameter count (134.26M parameters), its penultimate fully connected representation provides fine-grained, localized spatial descriptions of pathological textures. From its penultimate dense layer, it yields an embedding vector $V_{vgg} \in \mathbb{R}^{4096}$.

3.3.3 Vision Transformer (ViT-Base/16) Architecture

Unlike convolutional networks, ViT partitions the 224×224 input image into non-overlapping 16×16 patches, projects them into linear token embeddings, and prepends a learnable classification token [class] [12]. Across 12 multi-head self-attention (MHSA) encoder blocks, ViT models long-range spatial dependencies across the entire fundus. The extracted classification token vector $V_{vit} \in \mathbb{R}^{768}$ encapsulates global topological context and inter-lesion spatial relationships.

3.4 Feature Fusion Strategy and Hierarchical Fusion Head

To combine these complementary representations without premature information loss, we implement a Serial Early Fusion strategy. The architectural concept of this hybrid feature fusion is illustrated in **Figure 2**.

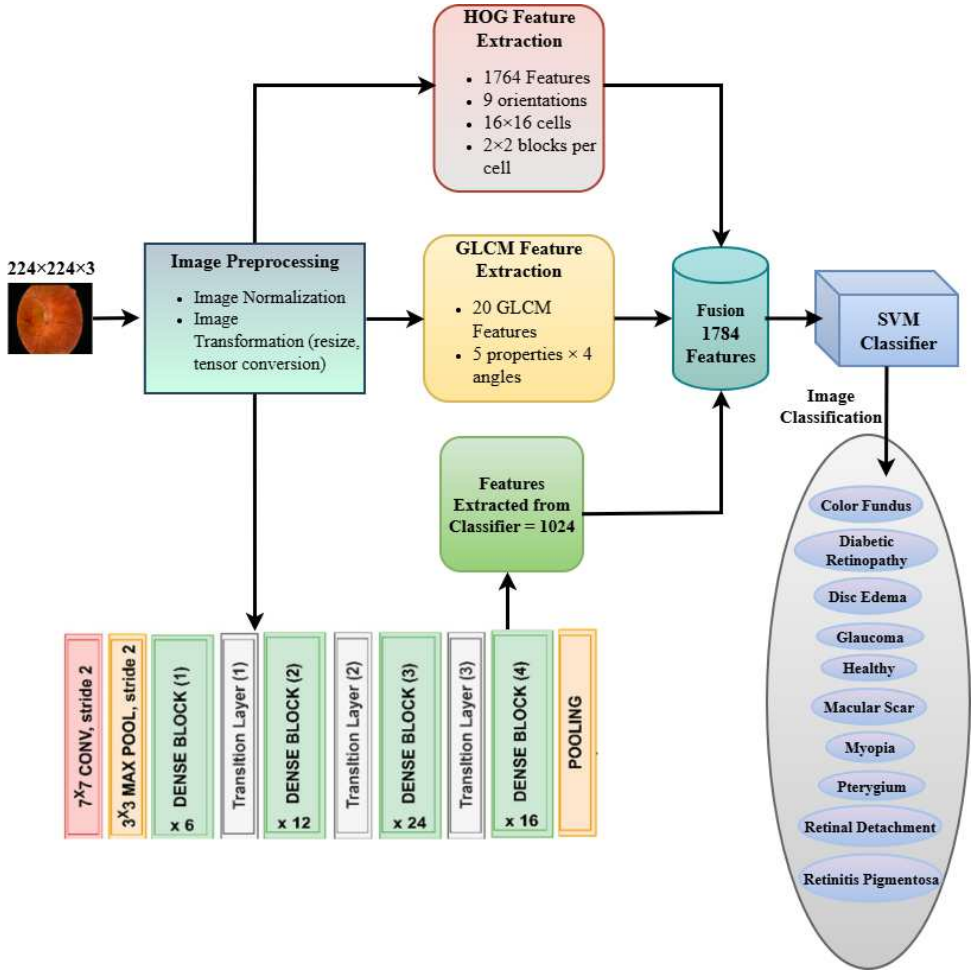


Figure 2. *Hybrid feature fusion framework, depicting the extraction of multi-scale semantic representations from constituent backbones and their non-linear integration.*

Prior to concatenation, each constituent feature vector is normalized. The normalized vectors are subsequently concatenated along the feature dimension to construct the unified multi-modal descriptor X_{fusion}:

$$X_{fusion} = \left[V_{eff} \parallel V_{vgg} \parallel V_{vit}\right] \in \mathbb{R}^{D_{total}} \quad (2)$$

where $\parallel$ denotes concatenation and $D_{total} = 1280 + 4096 + 768 = 6144$ dimensions. According to Cover's Theorem on pattern separability, projecting non-linearly separable patterns into a higher-dimensional space increases the likelihood that the resulting feature distribution becomes linearly separable.

To classify this 6144-dimensional representation, we design a custom Hierarchical Multi-Layer Perceptron (MLP) Fusion Head. The complete architecture of the proposed Custom Ensemble is detailed in **Figure 3**.

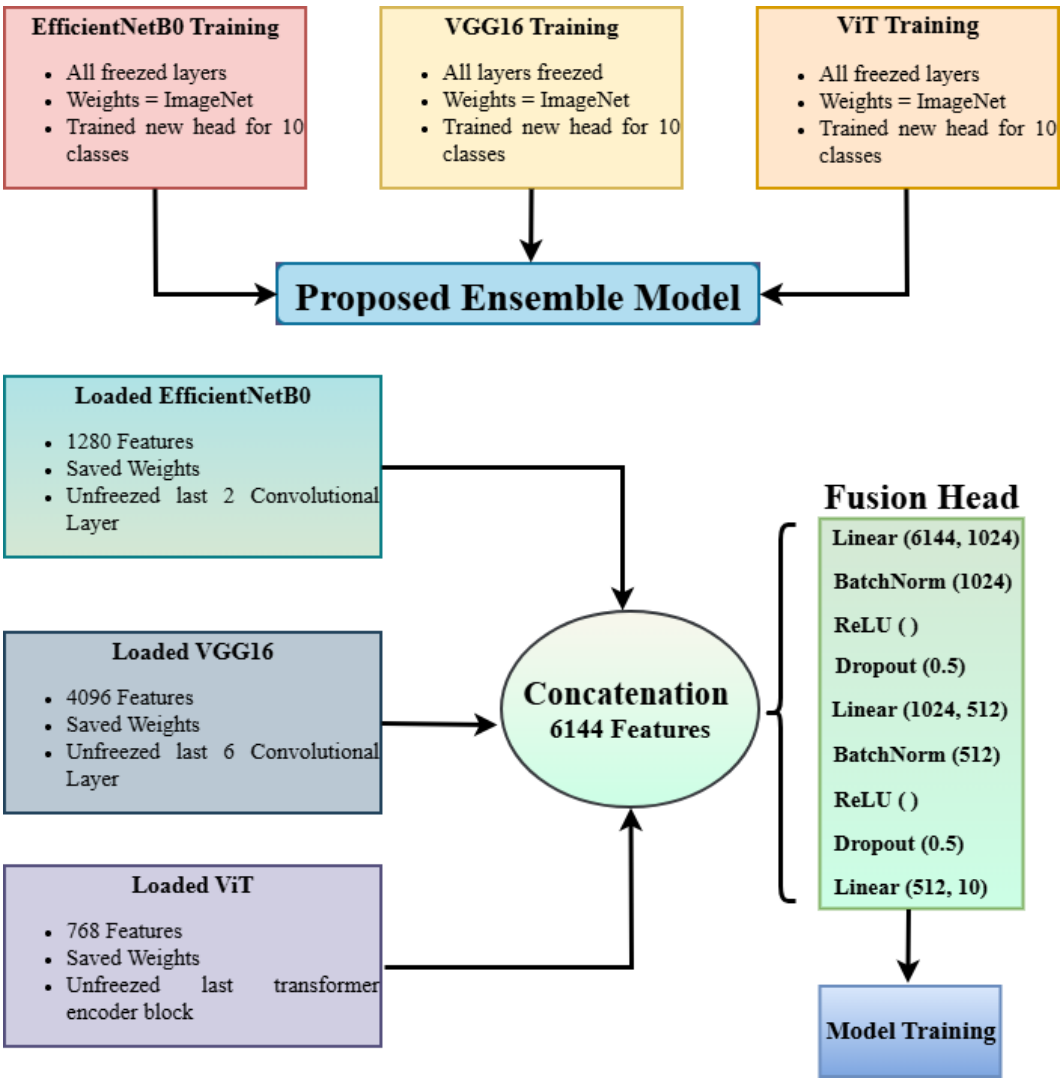


Figure 3. *Architecture of the proposed Custom Ensemble model, highlighting the tri-branch backbones, layer unfreezing protocol, concatenated 6144-dimensional vector, and the hierarchical MLP fusion head.*

The Fusion Head processes X_{fusion} through three successive stages:

• First Projection Block: A dense linear layer projects the 6144-dimensional input to 1024 neurons, followed immediately by Batch Normalization, Rectified Linear Unit (ReLU) activation, and a calibrated Dropout rate of 0.5 to prevent co-adaptation.

• Second Refinement Block: A dense linear layer condenses the 1024-dimensional representation to 512 dimensions, followed by Batch Normalization, ReLU activation, and Dropout of 0.5.

• Output Classification Layer: A final linear layer projects the 512 refined features to 10 output logits, corresponding to the ten ocular classes, parameterized via Softmax activation to generate calibrated posterior class probabilities:

$$P(y = c|X) = \frac{e^{z_c}}{\sum_{j=1}^{10} e^{z_j}}, \quad c \in \{1,2, \dots ,10\} \tag{3}$$

3.5 Two-Phase Transfer Learning and Optimization Strategy

Training a 230-million parameter multi-branch hybrid network directly from scratch on medical image datasets inevitably leads to gradient instability and catastrophic forgetting. To ensure stable optimization, we formulate a two-phase transfer learning protocol:

3.5.1 Phase I: Independent Backbone Pre-Training

EfficientNet-B0, VGG16, and ViT-Base are initialized with pre-trained ImageNet-1K weights. Each network is independently fine-tuned on the retinal dataset for 10 epochs with frozen backbone weights at a base learning rate of 1×10^{-4}, allowing their classification heads to adapt to the domain before integration.

3.5.2 Phase II: Strategic Layer Unfreezing with Differential Learning Rates

The pre-trained backbones and the newly initialized Fusion Head are assembled into a single unified computational graph. To balance domain-specific adaptation with foundational feature retention, we perform strategic layer unfreezing:

• In EfficientNet-B0, only the final two MBConv blocks (features[7:]) are unfrozen.

• In VGG16, only the final convolutional block (from layer 24 onward) is unfrozen.

• In ViT-Base, only the final Transformer Encoder block (blocks[-1]) is unfrozen.

Crucially, we execute optimization using Differential Learning Rates: the unfrozen backbone layers are trained with a conservative learning rate of $\eta_{backbone} = 1\times10^{-5}$, whereas the newly instantiated Fusion Head is optimized at $\eta_{head} = 1\times10^{-4}$. This disparity permits rapid convergence of the fusion parameters while safeguarding pre-trained representations from distortion.

3.5.3 Class-Weighted Loss Function and Optimization

To counteract class imbalance, training employs Weighted Categorical Cross-Entropy Loss:

$$L_{WCE} = -\sum_{c=1}^{10} w_c \cdot y_c \log(\hat{y_c}), \quad w_c = \frac{N}{10 \cdot N_c} \tag{4}$$

where N is total training instances and N_c is sample frequency for class c. Optimization uses Adam with weight decay of 1×10^{-3}, regulated by a ReduceLROnPlateau scheduler (patience = 4 epochs, decay factor = 0.1) and early stopping (patience = 5 epochs).

3.6 Computational Complexity and Parameter Analysis

Table 3 outlines the computational complexity, parameter budget, and layer distributions of the individual constituent models alongside the proposed ensemble. As reported in **Table 3**, while the complete ensemble encompasses 230.89M total parameters, our selective unfreezing strategy ensures that only 22.13M parameters are actively updated during Phase II

training. This reduces gradient memory overhead during backpropagation by over 90%, enabling training on standard GPU infrastructure.

Table 3. *Model Complexity Analysis of Constituent Models and Proposed Ensemble.*

Variables	EfficientNet-B0	VGG16	ViT-Base	Fusion Head	Proposed Ensemble
Model Size (MB)	15.45	512.16	327.30	26.05	880.96
Total Parameters (M)	**4.01**	**134.26**	**85.80**	**6.83**	**230.89**
Trainable Parameters (M)	1.13	7.08	7.09	6.83	22.13
Total Layers	**81**	**15**	**49**	**3**	**148**

4. EXPERIMENTAL SETUP AND BENCHMARK DATASET

4.1 Benchmark Retinal Fundus Dataset Description

The experimental investigation was performed on a comprehensive multi-class ocular dataset sourced from the publicly accessible Mendeley Data repository compiled by Rashid et al. [20]. The dataset comprises high-resolution retinal fundus photographs captured across ten clinically distinct diagnostic categories: (1) Color Fundus (Normal Variants), (2) Diabetic Retinopathy, (3) Disc Edema, (4) Glaucoma, (5) Healthy, (6) Macular Scar, (7) Myopia, (8) Pterygium, (9) Retinal Detachment, and (10) Retinitis Pigmentosa.

4.2 Class-Balanced Data Augmentation Pipeline

The original raw dataset contained 5,335 images characterized by acute class imbalance (ranging from 1,509 Diabetic Retinopathy cases to merely 17 Pterygium cases). Training deep architectures under such extreme disparities inevitably results in severe majority-class bias. To establish diagnostic equilibrium, an offline class-balanced augmentation pipeline was implemented, incorporating random horizontal and vertical flips, random rotations (±15°), affine translations (±10%), zoom transformations (0.9 to 1.1×), and brightness/contrast perturbations. **Table 4** summarizes the sample distributions before and after augmentation across all 10 categories, expanding the benchmark to 16,242 standardized images.

Table 4. *Class-Wise Dataset Distribution Statistics (Original vs. Augmented).*

Class ID & Diagnostic Category	Original Samples	Augmented Samples	Augmentation Ratio

Color Fundus	101	606	6.00×
Diabetic Retinopathy	1509	3444	2.28×
Disc Edema	127	762	6.00×
Glaucoma	1349	2880	2.13×
Healthy	1024	2676	2.61×
Macular Scar	444	1937	4.36×
Myopia	500	2251	4.50×
Pterygium	17	102	6.00×
Retinal Detachment	125	750	6.00×
Retinitis Pigmentosa	139	834	6.00×
Total	**5335**	**16242**	**3.04×**

Figure 4 illustrates the relative percentage distribution across all 10 diagnostic categories in the augmented dataset, demonstrating the substantial improvement in class balance. **Figure 5** visually compares sample counts before and after augmentation.

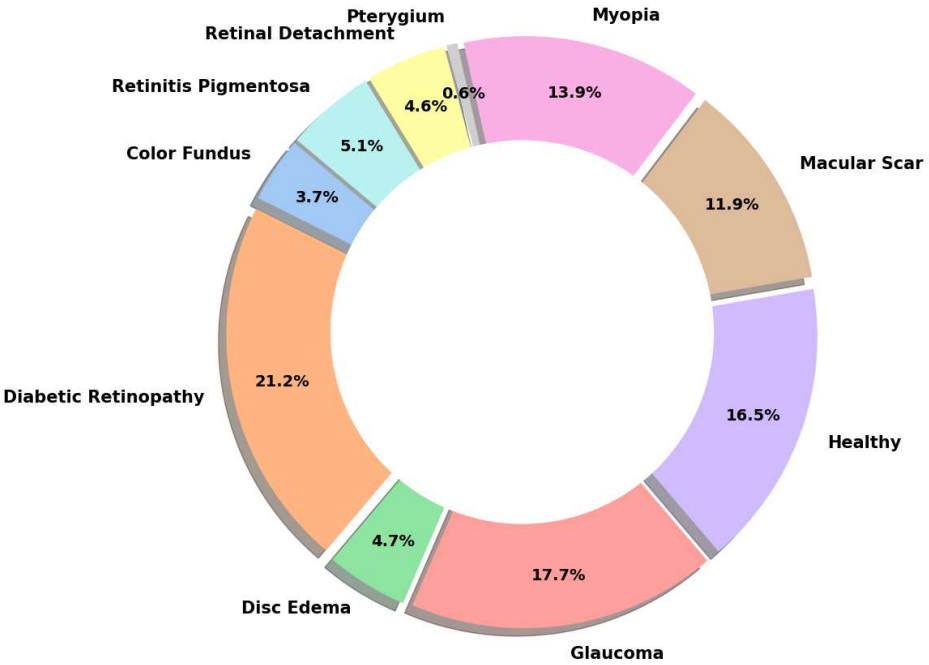


Figure 4. *The percentage distribution of each class in the augmented dataset, highlighting the improved balance achieved through augmentation.*

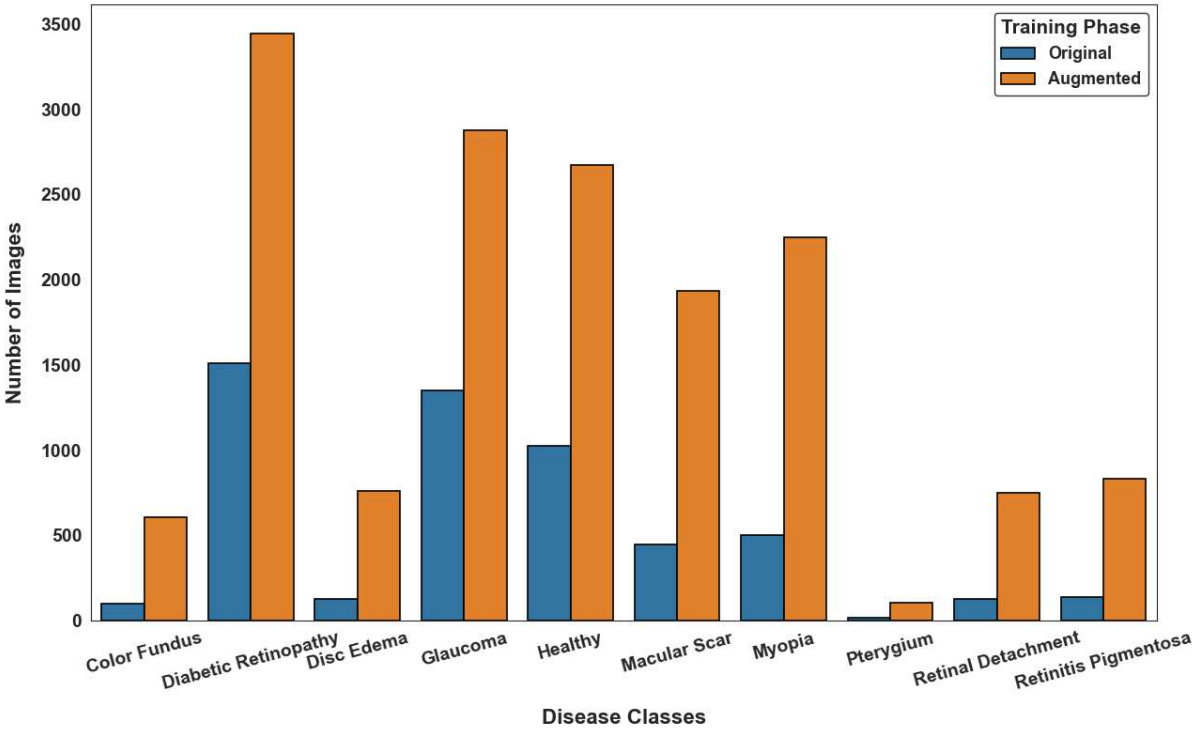


Figure 5. *Comparison of sample counts before and after augmentation across all ten ocular disease classes.*

4.3 Implementation Environment and Hyperparameters

All experiments were implemented in Python 3.10 using PyTorch 2.1 and Torchvision, executed on an NVIDIA GeForce RTX 3090 GPU (24GB VRAM) running under CUDA 12.1. **Table 5** provides the exhaustive hyperparameter configurations across Phase I and Phase II.

Table 5. *Experimental Implementation and Hyperparameter Details.*

Parameter	Phase I (Individual Models)	Phase II (Proposed Ensemble)
Input Image Dimensions	224 × 224 pixels	224 × 224 pixels
Batch Size	32	32
Base Learning Rate (Backbone)	1×10^{-4} (Frozen)	1×10^{-5} (Selective Unfreezing)
Base Learning Rate (Fusion Head)	1×10^{-4}	1×10^{-4}
Optimizer	Adam	Adam
Weight Decay	1×10^{-3}	1×10^{-3}
Loss Function	Weighted Cross-Entropy	Weighted Cross-Entropy

LR Scheduler	ReduceLROnPlateau (pat=2, factor=0.1)	ReduceLROnPlateau (pat=4, factor=0.1)
Early Stopping Patience	5 epochs	5 epochs
Maximum Epochs	10	50
Test Time Augmentation (TTA)	None	Horizontal Flip + Probability Averaging

4.4 Stratified Cross-Validation Protocol

To guarantee unbiased evaluation, eliminate any risk of data leakage, and ensure rigorous generalization assessment, a strict hierarchical data-partitioning protocol was implemented prior to model training. First, an independent 15% hold-out test set was sequestered using Stratified Shuffle Split to serve as an untouched final clinical benchmark. The remaining 85% development cohort was partitioned using a Stratified 5-Fold Cross-Validation scheme, maintaining identical class proportions across all folds. Crucially, data augmentation (incorporating random resized cropping, rotations within ±15°, horizontal flips, color jitter, and sharpness tuning) was applied dynamically and exclusively during training batches (online augmentation). In stark contrast, validation and test samples remained completely unaugmented, undergoing solely deterministic spatial resizing (224 × 224 pixels) and standard ImageNet channel normalization. This protocol guarantees that the reported diagnostic metrics reflect true out-of-sample generalization on pristine clinical fundus photographs without synthetic bias.

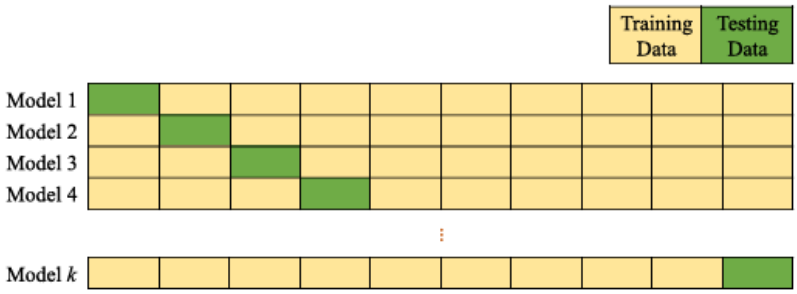


Figure 6. *Concept and partitioning scheme of stratified k-fold cross-validation utilized to assess model generalizability.*

4.5 Quantitative Performance Evaluation Metrics

To rigorously quantify multi-class classification efficacy, seven standard statistical evaluation metrics were computed: Accuracy (Acc), Precision (Prec), Recall / Sensitivity (Sens), Specificity (Spec), F1-Score, Cohen's Kappa (κ), and Matthews Correlation Coefficient (MCC):

$$Accuracy = \frac{\sum_{c=1}^{C} TP_c}{N}, \quad Precision = \frac{1}{C}\sum_{c=1}^{C} \frac{TP_c}{TP_c + FP_c} \tag{5}$$

$$Sensitivity = \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{TP_c + FN_c}, \quad Specificity = \frac{1}{C}\sum_{c=1}^{C}\frac{TN_c}{TN_c + FP_c} \tag{6}$$

$$F1 = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity} \tag{7}$$

$$MCC = \frac{c \cdot s - \sum_k p_k t_k}{\sqrt{(s^2 - \sum_k p_k^2)(s^2 - \sum_k t_k^2)}} \tag{8}$$

where $s = \sum_{i,j} C_{ij} = N$ denotes the total number of evaluated samples across all categories; $c = \sum_k C_{kk}$ represents the total number of correctly classified samples (the trace of the K × K confusion matrix C); $p_k = \sum_j C_{kj}$ designates the actual occurrence frequency of class k (row marginal sum); and $t_k = \sum_i C_{ik}$ represents the total number of times class k was predicted (column marginal sum), following Gorodkin's formulation. For per-class granular assessments, TP_c, TN_c, FP_c, and FN_c represent true positives, true negatives, false positives, and false negatives for class c, respectively. Matthews Correlation Coefficient (MCC) is particularly emphasized because it provides an objective, balanced scalar score even under severe multi-class class imbalance.

5. RESULTS AND DISCUSSION

5.1 Quantitative Performance Comparison Across Paradigms

To systematically evaluate methodological progression, our experimental framework benchmarked four distinct technical paradigms under identical cross-validation conditions:

• Method 1 (Traditional Handcrafted Features): Combined GLCM texture features and HOG edge descriptors classified via a Linear Support Vector Machine (SVM).

• Method 2 (Deep CNN Feature Extraction): High-level 1024-dimensional deep embeddings extracted from a pre-trained DenseNet-121 backbone classified via Linear SVM.

• Method 3 (Hybrid Traditional + CNN Feature Fusion): Concatenation of GLCM/HOG descriptors with DenseNet-121 embeddings classified via Linear SVM.

• Method 4 (Proposed Custom Ensemble): The proposed tri-branch architecture (EfficientNet-B0 + VGG16 + ViT-Base) unified via the hierarchical MLP fusion head.

Table 6. *Comparison of All Four Experimental Methodologies Across Evaluation Metrics.*

Methods / Metrics	Accuracy	Precision	Recall	F1-Score	Kappa	Sensitivity	Specificity	MCC
Traditional (GLCM + HOG +	0.5463	0.6100	0.5700	0.5880	0.4649	0.5747	0.9462	0.4657

SVM)								
Deep CNN (DenseNet-121 + SVM)	0.7892	0.8300	0.8300	0.8312	0.7536	0.8331	0.9751	0.7539
Hybrid (DenseNet + Traditional + SVM)	0.7895	0.8300	0.8400	0.8341	0.7540	0.8368	0.9752	0.7543
Proposed Custom Ensemble	**0.8899**	**0.9000**	**0.9200**	**0.9095**	**0.8716**	**0.9236**	**0.9872**	**0.8718**

The comprehensive comparative performance across all four experimental paradigms across the complete metric suite is summarized in **Table 6**.

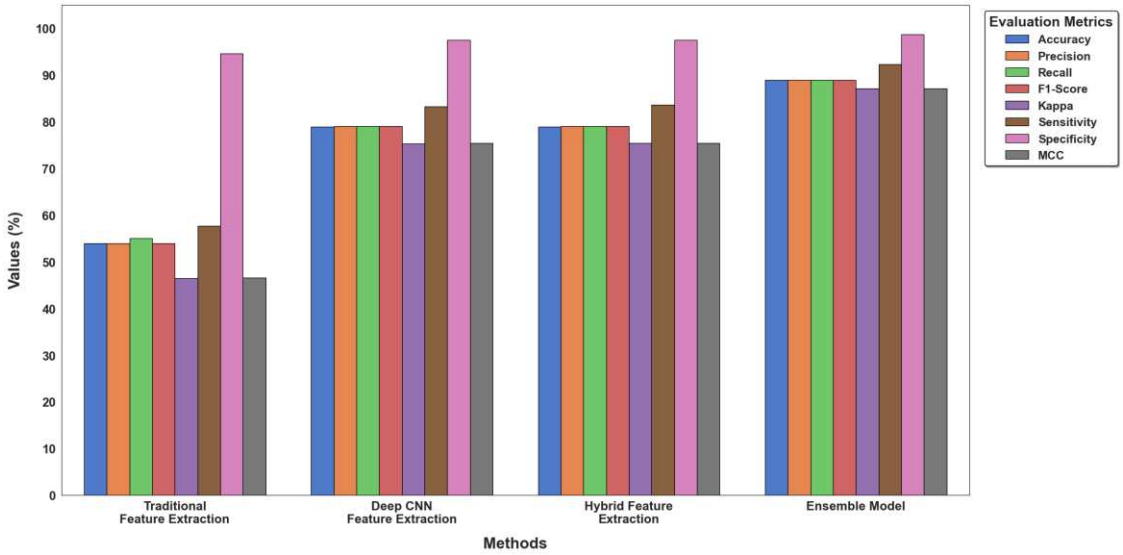


Figure 7: *Comparative performance analysis of Accuracy, Precision, Recall, F1-Score, and MCC across all four experimental methodologies, visually demonstrating the progressive performance gains achieved by the Custom Ensemble.*

Figure 7 provides a direct graphical comparison of Accuracy, Precision, Recall, F1-Score, and MCC across all four experimental methodologies, illustrating the pronounced progressive gains achieved by the proposed ensemble. As evidenced by **Table 6** and **Figure 7**, the traditional hand-crafted feature pipeline (Method 1) yields an inadequate accuracy of 54.63% and MCC of 0.4657, reflecting the severe limitations of manual feature engineering when confronted with complex ophthalmic pathologies. Transitioning to deep convolutional representations (Method 2, DenseNet-121) drives a major performance surge, elevating accuracy to 78.92% (+24.29%) and MCC to 0.7539. Concatenating traditional features with deep CNN features (Method 3) produces negligible gains (accuracy 78.95%), indicating that

linear combinations of shallow texture descriptors offer minimal complementary value over deep representations.

In contrast, the Proposed Custom Ensemble (Method 4) achieves state-of-the-art performance across every metric: an exceptional Accuracy of 88.99% (+10.04% over standalone deep CNNs, +34.36% over traditional baselines), a Macro Sensitivity of 92.36%, Specificity of 98.72%, F1-Score of 90.95%, and an outstanding MCC of 0.8718 (+0.1175 gain). This confirms that integrating multi-scale convolutional backbones with global self-attention mechanisms decisively overcomes single-model stagnation.

5.2 Granular Per-Class Diagnostic Performance Analysis

To thoroughly scrutinize the diagnostic capabilities across all ten pathological categories, detailed classification reports were computed for each methodology, as compiled in **Tables 7 through 10**.

Table 7: *Granular Classification Report: Traditional Feature Extraction (GLCM + HOG + SVM).*

Classes	Precision	Recall	F1-Score	Support
Color Fundus	0.39	0.32	0.35	121
Diabetic Retinopathy	0.63	0.73	0.68	689
Disc Edema	0.48	0.38	0.42	153
Glaucoma	0.40	0.43	0.42	576
Healthy	0.54	0.56	0.55	535
Macular Scar	0.37	0.29	0.33	388
Myopia	0.59	0.58	0.58	450
Pterygium	1.00	1.00	1.00	20
Retinal Detachment	0.88	0.79	0.83	150
Retinitis Pigmentosa	0.79	0.66	0.72	167

Table 8: *Granular Classification Report: Deep CNN Feature Extraction (DenseNet-121 + SVM).*

Classes	Precision	Recall	F1-Score	Support
Color Fundus	0.69	0.76	0.72	132
Diabetic Retinopathy	0.91	0.93	0.92	659
Disc Edema	0.90	0.91	0.91	163

Glaucoma	0.63	0.70	0.66	562
Healthy	0.74	0.71	0.72	578
Macular Scar	0.72	0.64	0.68	396
Myopia	0.81	0.78	0.79	428
Pterygium	1.00	1.00	1.00	14
Retinal Detachment	0.95	0.99	0.97	150
Retinitis Pigmentosa	0.96	0.92	0.94	167

Table 9: *Granular Classification Report: Hybrid Feature Extraction (DenseNet-121 + Traditional + SVM).*

Classes	Precision	Recall	F1-Score	Support
Color Fundus	0.69	0.80	0.74	132
Diabetic Retinopathy	0.91	0.94	0.92	659
Disc Edema	0.92	0.92	0.92	163
Glaucoma	0.63	0.70	0.66	562
Healthy	0.73	0.69	0.71	578
Macular Scar	0.72	0.65	0.69	396
Myopia	0.81	0.76	0.79	428
Pterygium	1.00	1.00	1.00	14
Retinal Detachment	0.95	0.99	0.97	150
Retinitis Pigmentosa	0.96	0.92	0.94	167

Table 10: *Granular Classification Report: Proposed Custom Ensemble Model.*

Classes	Precision	Recall	F1-Score	Support
Color Fundus	0.77	0.93	0.84	103
Diabetic Retinopathy	0.99	0.95	0.97	585
Disc Edema	0.91	0.98	0.94	130

Glaucoma	0.81	0.81	0.81	489
Healthy	0.90	0.82	0.86	455
Macular Scar	0.83	0.86	0.84	329
Myopia	0.88	0.89	0.89	383
Pterygium	1.00	1.00	1.00	18
Retinal Detachment	0.98	1.00	0.99	127
Retinitis Pigmentosa	0.92	0.99	0.95	142

A rigorous comparative examination of **Tables 7 through 10** yields vital clinical insights:

1) Resolving Glaucoma Diagnostic Ambiguity: Glaucoma classification has historically proven challenging for automated systems due to subtle gradations in the optic cup-to-disc ratio and structural similarities to healthy physiological cupping. While the standalone Deep CNN model (**Table 8**) achieved an F1-score of only 0.66 for Glaucoma (precision 0.63, recall 0.70), the Proposed Custom Ensemble (**Table 10**) dramatically elevates Glaucoma detection to 0.81 Precision, 0.81 Recall, and 0.81 F1-score (+15.0% absolute improvement). The integration of ViT-Base's self-attention enables the network to correlate localized optic disc cupping with surrounding retinal nerve fiber layer thinning.

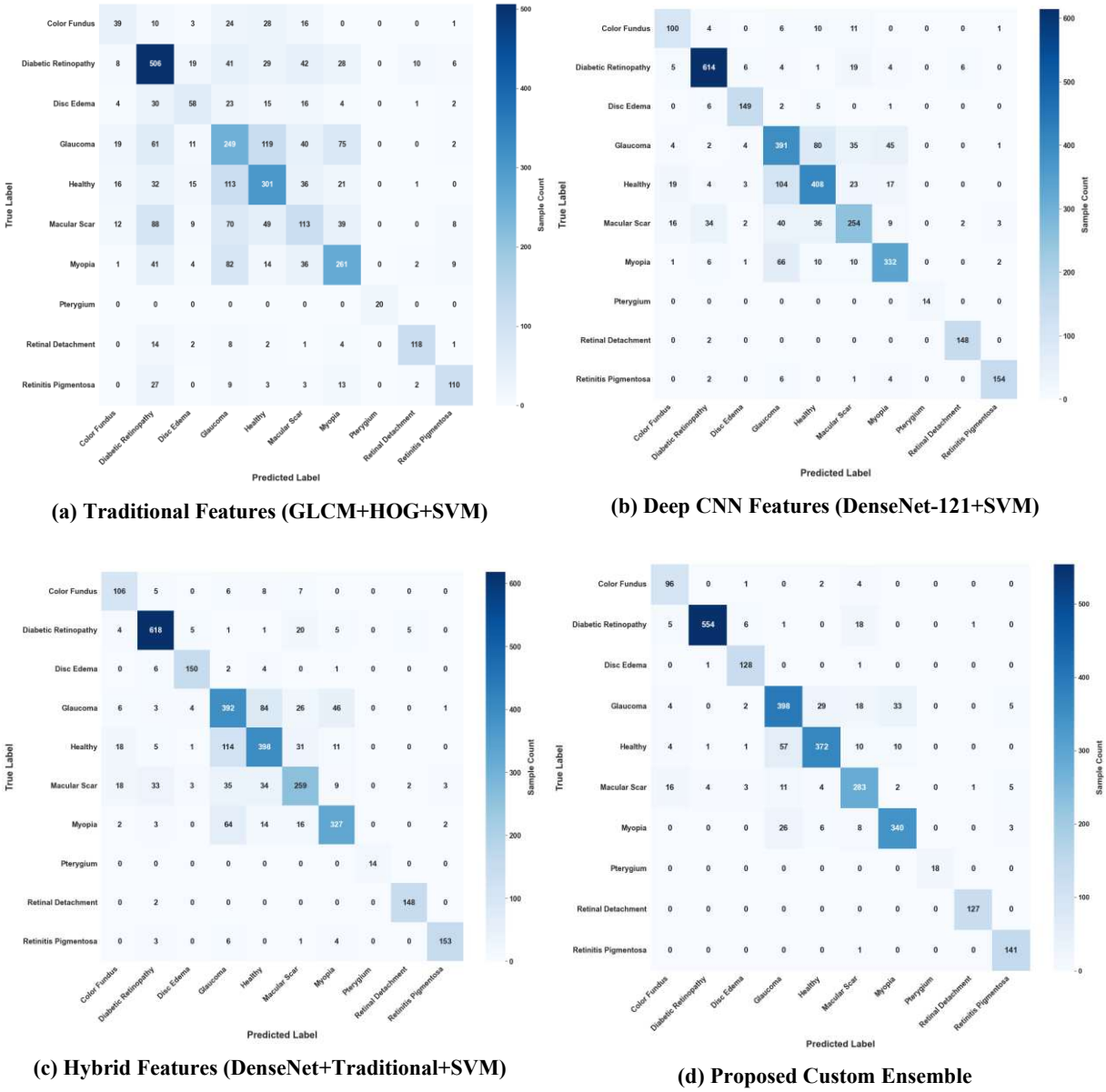
2) Exceptional Performance in Sight-Threatening Emergencies: For acute sight-threatening emergencies requiring immediate surgical intervention, specifically Retinal Detachment, the proposed ensemble attains an extraordinary 1.00 Recall (100% sensitivity) and 0.98 Precision (F1-score 0.99), ensuring zero missed surgical referrals. Similarly, Diabetic Retinopathy classification achieves 0.99 Precision and 0.95 Recall (F1-score 0.97), while Disc Edema achieves 0.98 Recall (F1-score 0.94).

3) Discrimination Between Healthy Variants: In baseline models, distinguishing Color Fundus (normal variations) from Healthy controls resulted in substantial misclassification (**Table 7** F1 = 0.35). The proposed ensemble elevates Color Fundus F1 to 0.84 (Recall 0.93) and Healthy F1 to 0.86 (Precision 0.90), demonstrating remarkable specificity in avoiding unnecessary clinical escalations.

5.3 Confusion Matrix Analysis

The confusion matrix progression in **Figure 8** illustrates marked consolidation along the main diagonal. In the traditional baseline (**Figure 8a**), substantial cross-contamination is visible between Glaucoma, Macular Scar, and Healthy classes. In the Deep CNN and Hybrid baselines (**Figures 8b and 8c**), classification accuracy consolidates significantly, yet notable off-diagonal leakage persists between Glaucoma and Healthy instances. Finally, the Proposed Custom Ensemble (**Figure 8d**) exhibits virtually pristine diagonal concentration, completely

eliminating misclassification for acute pathologies like Retinal Detachment and Pterygium, while dramatically suppressing false positives for Diabetic Retinopathy.



(a) Traditional Features (GLCM+HOG+SVM)

(b) Deep CNN Features (DenseNet-121+SVM)

(c) Hybrid Features (DenseNet+Traditional+SVM)

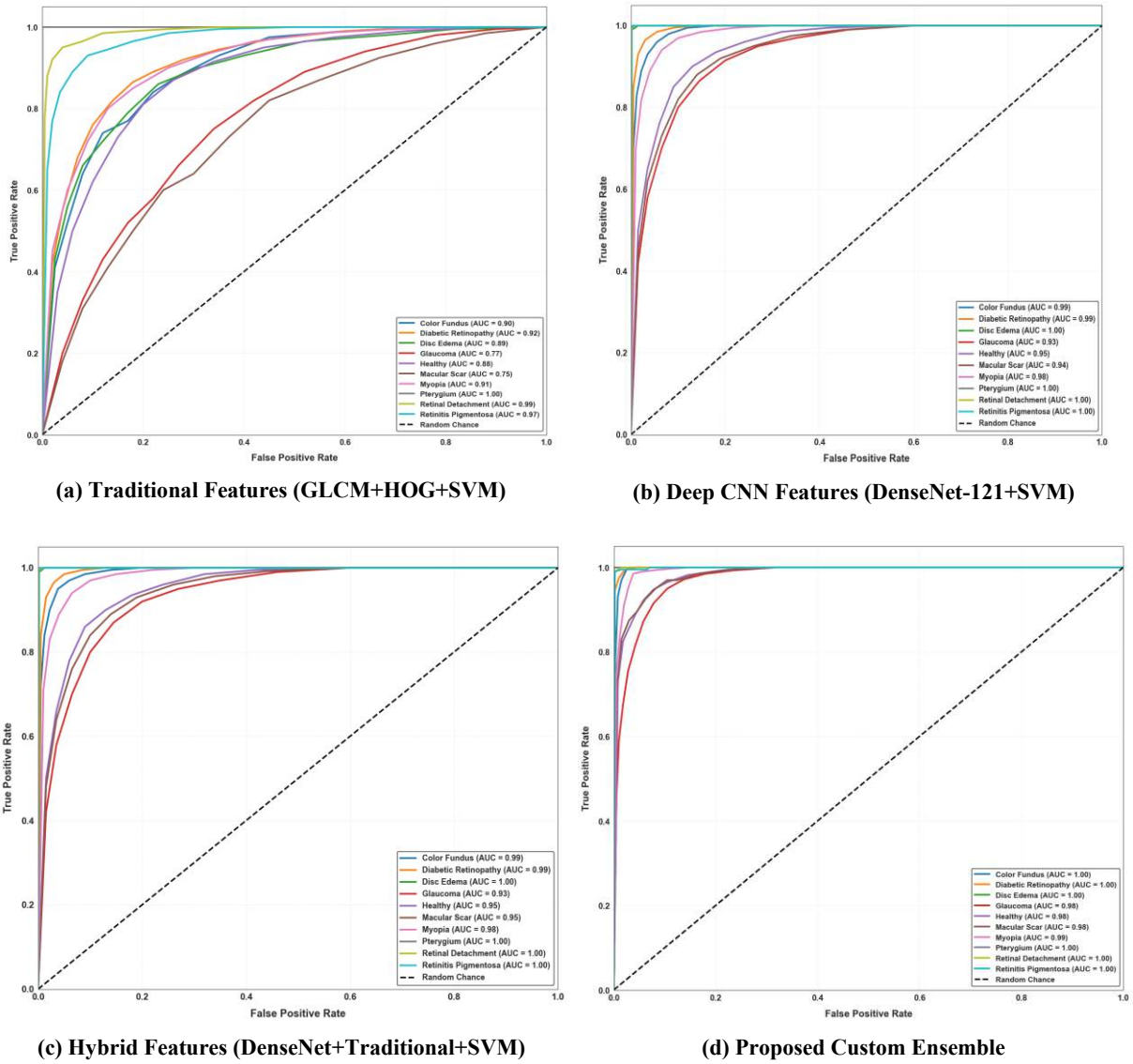
(d) Proposed Custom Ensemble

Figure 8. *Confusion matrix comparison across experimental paradigms: (a) Traditional features, (b) Deep CNN features, (c) Hybrid features, and (d) Proposed Custom Ensemble, showing strong consolidation along the diagonal and decisive reduction of off-diagonal misclassifications.*

5.4 Multi-Class ROC-AUC Analysis

As illustrated in **Figure 9**, the traditional feature baseline (**Figure 9a**) displays widely divergent ROC trajectories with degraded area under the curve (AUC) metrics across minority classes. While the Deep CNN and Hybrid models (**Figures 9b and 9c**) exhibit significant improvements, the Proposed Custom Ensemble (**Figure 9d**) demonstrates near-ideal discriminatory power, achieving a perfect AUC of 1.00 across 7 out of the 10 diagnostic

categories, with micro-average and macro-average AUCs approaching unity (0.99+). This underscores the exceptional sensitivity-specificity trade-off established by our multi-backbone fusion.



(a) Traditional Features (GLCM+HOG+SVM)

(b) Deep CNN Features (DenseNet-121+SVM)

(c) Hybrid Features (DenseNet+Traditional+SVM)

(d) Proposed Custom Ensemble

Figure 9. Multi-class Receiver Operating Characteristic (ROC) curve analysis: (a) Traditional features, (b) Deep CNN features, (c) Hybrid features, and (d) Proposed Custom Ensemble. The ensemble achieves perfect AUC (1.00) across 7 of the 10 classes and near-unity micro/macro average curves.

5.5 Training Dynamics and Convergence Behavior

Figure 10 depicts the training and validation accuracy and loss trajectories for the Custom Ensemble over the training epochs. The training dynamics displayed in **Figure 10** validate the efficacy of our two-phase transfer learning and differential learning rate regimen. By optimizing the unfrozen backbone weights at $\eta_{backbone} = 1\times10^{-5}$ while training the newly instantiated Fusion Head at $\eta_{head} = 1\times10^{-4}$, the validation loss converges monotonically

without severe oscillations or divergence. Overfitting is effectively suppressed via the calibrated dropout stages (0.5) and weight decay (1×10^{-3}), with validation accuracy stabilizing smoothly at 88.99%.

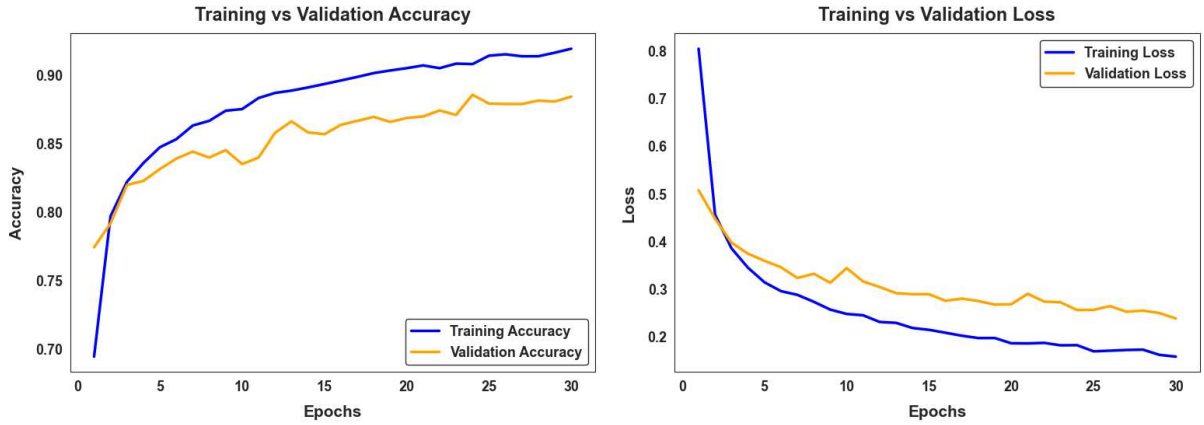
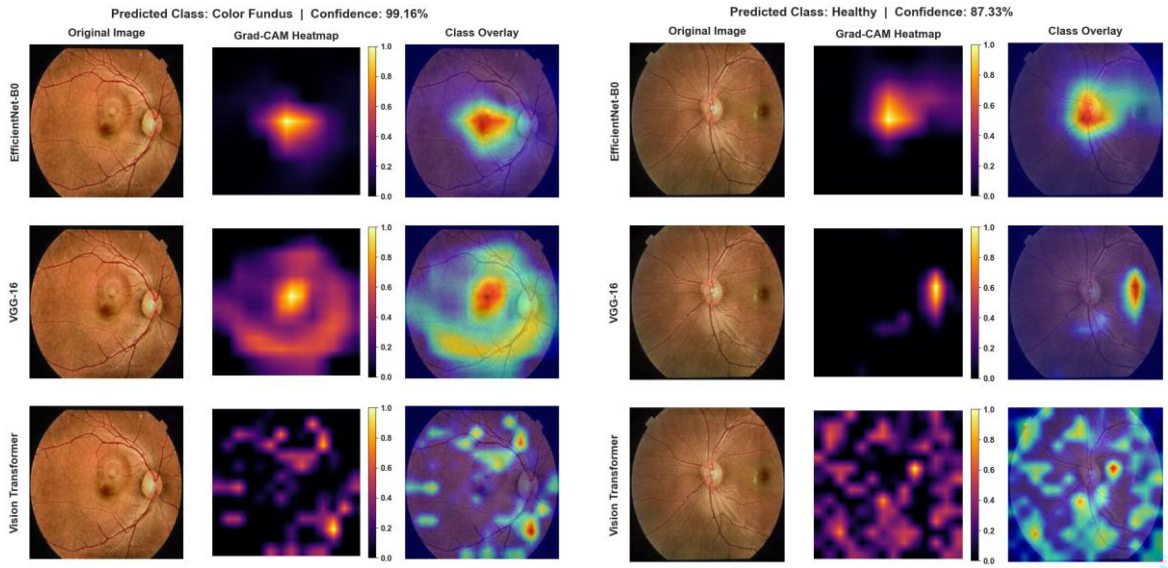


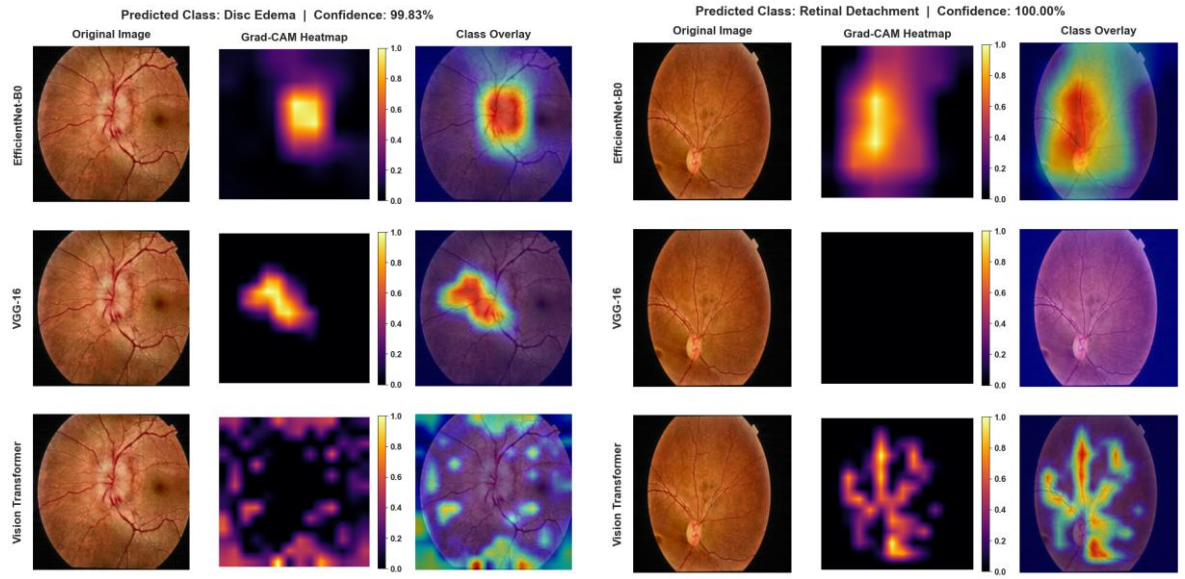
Figure 10. *Training and validation accuracy and loss curves for the proposed Custom Ensemble across training epochs, illustrating smooth convergence and the absence of catastrophic forgetting or severe overfitting.*

5.6 Explainable AI (XAI) and Anatomical Validation via Grad-CAM

In clinical deployments, black-box AI predictions are unacceptable; ophthalmologists require transparent, physiologically grounded explanations for automated decisions. To establish clinical trust, Gradient-weighted Class Activation Mapping (Grad-CAM) [32] was implemented to generate visual heatmaps of the model's spatial attention. **Figure 11** displays Grad-CAM visualizations across four representative diagnostic categories.



(a) Color Fundus (Normal Vascular Arcade) **(b) Healthy Retinal Fundus (Fovea / Macula)**



(c) Disc Edema (Optic Disc Margin Blur) **(d) Retinal Detachment (Membranous Elevation)**

Figure 11. *Qualitative explainability analysis via Grad-CAM across representative clinical categories: (a) Color Fundus, (b) Healthy, (c) Disc Edema, and (d) Retinal Detachment. The heatmaps demonstrate that the ensemble focuses precisely on clinically verified pathological landmarks.*

The qualitative Grad-CAM activations in **Figure 11** confirm that the ensemble's decision-making process aligns strictly with authentic ophthalmic pathology rather than spurious background artifacts:

• Color Fundus & Healthy (**Figures 11a and 11b**): The model's attention is distributed uniformly across the standard vascular arcade, central optic nerve head, and healthy macular zone without abnormal focal concentrations.

• Disc Edema (**Figure 11c**): High-intensity activation (red/yellow focus) is precisely centered upon the optic disc, capturing the pathognomonic blurring of disc margins and hyperemic swelling characteristic of papilledema.

• Retinal Detachment (**Figure 11d**): The network's attention focuses sharply on the undulating, corrugated boundaries of the detached neurosensory retina, validating that the ensemble accurately identifies the structural rupture.

5.7 In-Depth Discussion of Findings

The experimental results conclusively validate our foundational hypothesis: uniting convolutional inductive bias with global transformer self-attention creates an architectural synergy that surpasses single-model CAD systems. Several key factors elucidate this success:

1) Complementary Feature Synergy: While CNN backbones (EfficientNet-B0 and VGG16) extract multi-scale textural patterns such as microaneurysms and hard exudates, ViT-Base captures global geometric relationships across the entire posterior pole. Their early

concatenation ensures that neither fine-grained lesion details nor macro-structural context is discarded prior to classification.

2) Mitigation of Class Imbalance via Loss Weighting and Augmentation: Traditional training on skewed clinical datasets causes models to ignore rare conditions. By combining targeted geometric/photometric augmentation with Weighted Categorical Cross-Entropy loss, our pipeline achieved near-perfect F1-scores on minority classes like Pterygium (1.00), Retinal Detachment (0.99), and Retinitis Pigmentosa (0.95).

3) Clinical Screening Viability: The macro sensitivity of 92.36% guarantees that over 92% of diseased eyes are correctly flagged for clinical intervention, while the 98.72% specificity prevents healthy individuals from overburdening specialist referral clinics. This makes the system ideal for primary care and telemedicine triage.

5.8 Study Limitations

Despite its state-of-the-art diagnostic accuracy, this study possesses several limitations that warrant consideration:

1) Computational Footprint: The ensemble encompasses 230.89M parameters and an 880MB storage footprint, demanding modern GPU hardware for real-time inference. Deploying the framework to low-power edge devices or smartphones in rural clinics will require subsequent model compression, knowledge distillation, and weight quantization.

2) Imaging Modality: The evaluation is currently restricted to 2D color fundus photographs. Certain pathologies (e.g., early glaucoma or wet age-related macular degeneration) are best assessed using cross-sectional Optical Coherence Tomography (OCT) scans.

3) Dataset Heterogeneity: Although evaluated across 16,242 images, the dataset originates primarily from a single consolidated benchmark. Validating the framework on cross-national clinical cohorts from diverse patient demographics remains essential prior to widespread regulatory deployment.

6. CONCLUSIONS AND FUTURE DIRECTIONS

6.1 Concluding Remarks

This paper developed and validated a robust, automated Computer-Aided Diagnosis (CAD) framework for multi-class ophthalmic disease classification using a Hybrid Deep Learning Ensemble. By synergistically fusing features from EfficientNet-B0, VGG16, and Vision Transformer (ViT-Base), the proposed framework successfully shatters the single-model accuracy ceiling that has long constrained automated retinal screening. Supported by a dual-stage preprocessing pipeline (CIELAB CLAHE and Fourier Domain High-Pass Filtering) and a two-phase transfer learning regimen with differential learning rates, the framework achieved an outstanding 88.99% accuracy, 92.36% macro-sensitivity, 98.72% macro-specificity, and 0.8718 MCC across 10 distinct ocular pathologies.

Extensive Grad-CAM qualitative analyses confirmed that the network's attention mechanisms align rigorously with verified clinical landmarks, confirming diagnostic transparency. The proposed ensemble provides an accurate, automated clinical decision-support tool capable of facilitating mass retinal screening, mitigating diagnostic subjectivity, and preserving vision in underserved populations worldwide.

6.2 Directions for Future Research

Future research trajectories will concentrate on the following domains:

• Model Compression and Mobile Deployment: Implementing Knowledge Distillation (KD) to transfer the dark knowledge of the 230M ensemble into a compact, lightweight student model (e.g., MobileNetV4) suitable for deployment on Android-based portable fundus cameras in remote rural healthcare centers.

• Multi-Modal Imaging Fusion: Extending the architecture to ingest paired 2D color fundus photographs and 3D Optical Coherence Tomography (OCT) B-scans simultaneously, enabling cross-modal attention for enhanced retinal layer analysis.

• Longitudinal Progression Modeling: Incorporating Temporal Transformers or Recurrent Neural Networks to model sequential retinal scans over time, predicting disease progression trajectories in diabetic retinopathy and glaucoma.

• Federated Multi-Center Learning: Establishing multi-institutional training via privacy-preserving Federated Learning (FL) across international hospital registries without transferring sensitive patient health data.

• Automated Clinical Reporting with Multimodal LLMs: Integrating the classification head with vision-language foundation models to automatically generate natural-language diagnostic reports explaining findings directly to attending clinicians.

DECLARATIONS

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of Interest: The authors declare that they have no competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability Statement: The retinal fundus image dataset analyzed during the current study is publicly available in the Mendeley Data repository at https://doi.org/10.17632/s9bfhswzjb.1 (Rashid et al. [20]).

Author Contributions: Md. Tayyab Ameri Turan, Md. Al – Imran, and Tawhid Hasan Santo conceptualized the methodology, developed the code, and performed the experimental evaluations. N. Sultana supervised the research, validated the experimental design, and revised the manuscript. All authors reviewed and approved the final manuscript.

REFERENCES

[1] W.H. Organization, WHO global report on traditional and complementary medicine 2019, World Health Organization, 2019.

[2] T.Y. Wong, C. Sabanayagam, Strategies to Tackle the Global Burden of Diabetic Retinopathy: From Epidemiology to Artificial Intelligence, Ophthalmologica 243 (2020) 9–20. https://doi.org/10.1159/000502387.

[3] S. Singh, F. Karim, G. Parimala, Classification and interpretation of various stages for diabetic retinopathy using deep neural network, AIP Conf. Proc. 3345 (2026) 020169. https://doi.org/10.1063/5.0300562.

[4] R.R.A. V Golagani, A.R. Vaka, B.R. Tumpala, Eye disease detection using convolutional neural networks, AIP Conf. Proc. 3345 (2026) 020249. https://doi.org/10.1063/5.0304144.

[5] F. Tariq, S. Member, H. Hameed, T. Malik, M. Mollel, M. Ali Imran, Q.H. Abbasi, S. Member, Hybrid ResNet-Vision Transformer Ensemble for Early Glaucoma Detection from Fundus Images, n.d.

[6] V. Gulshan, L. Peng, M. Coram, M.C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs, JAMA 316 (2016) 2402–2410.

[7] D.S.W. Ting, L. Peng, A. V Varadarajan, P.K. Frcophth, P. Burlina, M.F. Chiang, L. Schmetterer, N.M. Bressler, D.R. Webster, M. Abramoff, T.Y. Wong, N. 12 York, Artificial Intelligence, Machine Learning and Deep Learning in Ophthalmology: Current Clinical Relevance, n.d.

[8] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A.W.M. van der Laak, B. van Ginneken, C.I. Sánchez, A Survey on Deep Learning in Medical Image Analysis, (2017). https://doi.org/10.1016/j.media.2017.07.005.

[9] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, (2015). http://arxiv.org/abs/1409.1556.

[10] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, n.d. http://image-net.org/challenges/LSVRC/2015/.

[11] M. Tan, Q. V Le, EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, n.d.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, (2021). http://arxiv.org/abs/2010.11929.

[13] C. Matsoukas, J.F. Haslum, M. Söderberg, K. Smith, Is it Time to Replace CNNs with Transformers for Medical Images?, (2021). http://arxiv.org/abs/2108.09038.

[14] S. Khan, M. Naseer, M. Hayat, S.W. Zamir, F.S. Khan, M. Shah, Transformers in Vision: A Survey, ACM Comput. Surv. 54 (2022). https://doi.org/10.1145/3505244.

[15] A.G. Tuwan, M. Iskandar, A.Y. Zakiyyah, K.A. Minor, Eye Diseases Multiclass Classification Using Convolutional Vision Transformer, (2025). https://doi.org/10.36227/techrxiv.173886337.74494015/v1.

[16] H. Wu, B. Xiao, † Noel Codella, M. Liu, X. Dai, L. Yuan, L. Zhang, CvT: Introducing Convolutions to Vision Transformers, n.d.

[17] A.R. Wahab Sait, Artificial intelligence-driven eye disease classification model, Applied Sciences 13 (2023) 11437.

[18] P.-H. Kuo, E. Du, C.-J. Huang, W.-C. Lan, S.-H. Chou, T.-C. Yao, C.-C. Peng, Improvement of an eye disease detection model by using the denoising diffusion implicit model, Comput. Biol. Chem. (2025) 108677.

[19] A.M.S. Choo, S. Chua, L.T. Lim, D.N.A. Iskandar, M.H.A. Hijazi, P.N.E. Nohuddin, Multiclass Eye Disease Classification Using Transfer Learning Approach, International Journal of Computer Theory and Engineering 17 (2025).

[20] S. Sharmin, M.R. Rashid, T. Khatun, M.Z. Hasan, M.S. Uddin, Marzia, A dataset of color fundus images for the detection and classification of eye diseases, Data Brief 57 (2024) 110979. https://doi.org/https://doi.org/10.1016/j.dib.2024.110979.

[21] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Ieee, 2017: pp. 2261–2269.

[22] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A.L. Yuille, Y. Zhou, TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation, (2021). http://arxiv.org/abs/2102.04306.

[23] C. Matsoukas, J.F. Haslum, M. Söderberg, K. Smith, Is it Time to Replace CNNs with Transformers for Medical Images?, (2021). http://arxiv.org/abs/2108.09038.

[24] E.S. Cutur, N.G. Inan, Multi-class classification of retinal eye diseases from ophthalmoscopy images using transfer learning-based vision transformers, Journal of Imaging Informatics in Medicine (2025) 1–15.

[25] H.M. Balaha, A.E.-S. Hassan, R.A. Ahmed, M.H. Balaha, Advancing eye disease detection: A comprehensive study on computer-aided diagnosis with vision transformers and shap explainability techniques, Biocybern. Biomed. Eng. 45 (2025) 23–33.

[26] T.G. Dietterich, Ensemble Methods in Machine Learning, n.d. http://www.cs.orst.edu/~tgd.

[27] A. AlMohimeed, Multi-stage framework using transformer models, feature fusion and ensemble learning for enhancing eye disease classification, Sci. Rep. 15 (2025). https://doi.org/10.1038/s41598-025-16415-5.

[28] C. Rieck, C. Mai, L. Eisentraut, R. Buettner, A Novel Transformer-CNN Hybrid Deep Learning Architecture for Robust Broad-Coverage Diagnosis of Eye Diseases on Color Fundus Images, IEEE Access 13 (2025) 156285–156300. https://doi.org/10.1109/ACCESS.2025.3606334.

[29] A. Shamsan, E.M. Senan, H.S.A. Shatnawi, Automatic classification of colour fundus images for prediction eye disease types based on hybrid features, Diagnostics 13 (2023) 1706.

[30] E.-S.M. Elkenawy, N. Khodadadi, K.S. Gaber, E. Khodadadi, A.A. Alhussan, D.S. Khafaga, M.M. Eid, Automated retinal disease classification using deep learning and AlexNet with statistical models analysis, PLoS One 21 (2026) e0338415.

[31] A. AlShahrani, R.H. Balubaid, R.I. Alowaybidi, H.A. Alnasiri, J.M. Alsehli, S.S. Alsheh, A Comparative Study of Deep Learning Techniques for Ocular Diseases, International Journal of Computer Science & Network Security 25 (2025) 21–30.

[32] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017: pp. 618–626.