RESEARCH PAPER

Artificial Intelligence and Machine Learning Based Early Detection of Diabetic Retinopathy Using Retinal Image Analysis for Enhanced Clinical Decision Support

Abhiruchi Arvind Patil Bhagat[1], Kamlesh Kumar Dhiwar[2], Makala Ramesh[3], K. Thejomoorthy[4], Presilla R[5], Jagadish S Kallimani[6] and Raja Shekar Perusomula[7]*

[1]Department of Computer Science & Engineering, Yeshwantrao Chavan College of Engineering, Nagpur-441110.

[2]Century Cement College, Baikunth, Dist.- Raipur (Chhattisgarh).

[3]Department of Computer Science & Engineering, Sri Mittapalli College of Engineering, Guntur, Andhra Pradesh.

[4]University College of Pharmaceutical Sciences, Acharya Nagarjuna University, NH16, Nagarjuna Nagar, Guntur, Andhra Pradesh 522510.

[5]Department of Computer Science and Engineering, M S Ramaiah Institute of Technology, Bangalore, India; affiliated to Visvesvaraya Technological University, Belagavi-590018, Karnataka, India

[6]Department of Artificial Intelligence and Machine Learning, M S Ramaiah Institute of Technology, Bangalore, India; affiliated to Visvesvaraya Technological University, Belagavi-590018, Karnataka, India

[7]CSRI Lab, Department of Pharmacology, Vishnu Institute of Pharmaceutical Education & Research, Sangareddy-Narsapur Rd, Narsapur, Tuljaraopet, Telangana 502313

***Corresponding Author:**

Dr. Raja Shekar Perusomula,

HOD & Associate Professor,

CSRI Lab, Department of Pharmacology,

Vishnu Institute of Pharmaceutical Education & Research,

Sangareddy-Narsapur Rd, Narsapur, Tuljaraopet, Telangana 502313

Mail id: rajashekarcology@gmail.com

ABSTRACT

Diabetic retinopathy (DR) remains one of the leading causes of preventable blindness worldwide, with an estimated 103 million individuals affected globally as of recent epidemiological surveys. Early detection and timely intervention are critical to preventing irreversible vision loss; however, the shortage of trained ophthalmologists and the high cost of traditional screening programs present significant barriers in resource-limited settings. This study presents a comprehensive artificial intelligence (AI) and machine learning (ML)-based framework for the automated early detection of diabetic retinopathy through digital retinal fundus image analysis, aimed at augmenting clinical decision support systems. A multi-tiered computational pipeline was developed and evaluated using the publicly available APTOS 2019 Blindness Detection Dataset, MESSIDOR-2, IDRiD, and EyePACS datasets, comprising over 90,000 annotated retinal fundus images. The proposed system integrates classical image preprocessing techniques including green channel extraction, contrast-limited adaptive histogram equalization (CLAHE), and Gaussian filtering with advanced deep learning architectures such as EfficientNet-B4, DenseNet-121, ResNet-50, and a custom hybrid convolutional neural network (CNN). Eight distinct algorithmic formulations were designed and benchmarked, ranging from traditional machine learning pipelines using handcrafted features to end-to-end deep learning models. The proposed hybrid EfficientNet-B4 with attention mechanism (Formulation F8) achieved the highest diagnostic accuracy of 97.3%, sensitivity of 96.8%, specificity of 97.9%, and AUC-ROC of 0.989 in a five-class grading of diabetic retinopathy severity. Grad-CAM visualization confirmed that the model reliably identifies clinically relevant pathological features including microaneurysms, hard exudates, haemorrhages, and neovascularization. The system demonstrated robust performance across diverse ethnic populations and image acquisition conditions, supporting its potential for deployment in point-of-care clinical environments. This AI-driven platform has the potential to democratize ophthalmological screening, reduce diagnostic latency, and substantially improve patient outcomes in diabetic eye disease management.

Keywords: Diabetic Retinopathy, Artificial Intelligence, Deep Learning, Convolutional Neural Network, Retinal Image Analysis, EfficientNet, Clinical Decision Support, Fundus Photography, CLAHE, Transfer Learning, Grad-CAM, Medical Image Processing

How To Cite This Article: Bhagat AAP, Dhiwar KK, Ramesh M, Thejomoorthy K, Presilla R, Kallimani JS, Perusomula RS. Artificial intelligence and machine learning based early detection of diabetic retinopathy using retinal image analysis for enhanced clinical decision support. Int J Drug Deliv Technol. 2026;16(9s): 872-881; Doi: 10.25258/Ijddt.16.9s.91

1. INTRODUCTION

Diabetes mellitus is a chronic metabolic disorder characterized by persistent hyperglycaemia resulting from defects in insulin secretion, insulin action, or both. According to the International Diabetes Federation (IDF) Diabetes Atlas, the global prevalence of diabetes in adults aged 20–79 years reached 537 million in 2021, and this figure is projected to rise to 783 million by 2045[1]. Among the multifactorial microvascular complications of diabetes, diabetic retinopathy (DR) stands out as one of the most devastating, constituting the leading cause of new cases of blindness in working-age adults across industrialized nations. DR is a progressive disease affecting the microvasculature of the retina and, if left untreated, leads to severe visual impairment or complete blindness.

The pathophysiology of DR involves pericyte loss, basement membrane thickening, blood-retinal barrier breakdown, and eventual retinal ischaemia, culminating in proliferative neovascularization. Clinically, DR is classified into non-proliferative diabetic retinopathy (NPDR), which includes mild, moderate, and severe stages, and proliferative diabetic retinopathy (PDR), the most advanced stage associated with retinal detachment and vitreous haemorrhage[2]. The Early Treatment Diabetic Retinopathy Study (ETDRS) grading scale and the modified International Clinical Diabetic Retinopathy (ICDR) severity scale are widely used by clinicians to standardize disease classification.

Traditional screening programs rely on manual grading of fundus photographs by trained retinal specialists, a process that is time-consuming, subjective, costly, and not scalable to meet the growing global burden of diabetes. In developing countries such as India, where diabetic prevalence is alarmingly high exceeding 77 million cases access to specialized ophthalmological care remains profoundly inequitable[3]. Telemedicine and mobile health solutions have partially bridged this gap, yet the bottleneck of expert interpretation persists. Consequently, there is an urgent clinical need for automated, reliable, and scalable screening systems that can detect DR at its earliest treatable stages and integrate seamlessly into existing healthcare workflows.

The emergence of artificial intelligence (AI), and specifically deep learning, has catalysed a paradigm shift in medical image analysis. Convolutional Neural Networks (CNNs) have demonstrated remarkable performance in detecting and grading DR from fundus photographs, often matching or exceeding the diagnostic accuracy of expert ophthalmologists[4]. Landmark studies have demonstrated that AI-based systems can achieve sensitivity and specificity values exceeding 90% for DR detection, lending credence to their potential as autonomous screening tools. The U.S. Food and Drug Administration (FDA) approval of the IDx-DR system in 2018 marked a historic milestone, representing the first autonomous AI diagnostic system approved for clinical use without physician interpretation[5].

However, several challenges remain before widespread clinical deployment is feasible. These include model generalizability across diverse patient populations, imaging devices and environmental conditions; the need for explainable AI (XAI) to build physician trust; class imbalance in clinical datasets; and the integration of multi-modal data including optical coherence tomography (OCT) and fluorescein angiography. Furthermore, most published studies have evaluated models on single datasets, limiting their external validity[6]. The present study addresses these limitations by developing and rigorously evaluating a multi-formulation AI pipeline across four independent, diverse retinal image datasets.

This research makes the following key contributions: (1) design and systematic evaluation of eight distinct ML/DL formulations for DR grading, each employing different preprocessing strategies and model architectures; (2) comprehensive benchmarking on four publicly available datasets to assess cross-dataset generalizability; (3) integration of Gradient-weighted Class Activation Mapping (Grad-CAM) for explainability; (4) development of a clinical decision support module incorporating confidence scoring; and (5) comparison of computational efficiency metrics relevant to real-world deployment. The results demonstrate that the proposed framework achieves state-of-the-art performance while maintaining clinical interpretability, making it a viable candidate for integration into ophthalmological screening programs.

2. MATERIALS

The present study utilized four publicly available annotated retinal fundus image datasets widely recognized in the ophthalmological AI research community. The *APTOS 2019 Blindness Detection Dataset*, released by Aravind Eye Hospital (India) and hosted on Kaggle, constitutes the primary dataset comprising 3,662 high-resolution colour fundus images graded on a five-level ICDR severity scale (0: No DR, 1: Mild NPDR, 2: Moderate NPDR, 3: Severe NPDR, 4: PDR)[7]. The images were acquired using Topcon TRC-50DX and Zeiss Visucam fundus cameras under standardized conditions, with pixel dimensions ranging from 1024×1024 to 4288×2848. The dataset exhibits significant class imbalance, with Grade 0 constituting approximately 49.3% of samples, necessitating oversampling strategies.

The *MESSIDOR-2 Dataset*, an extension of the original MESSIDOR (Methods to Evaluate Segmentation and Indexing Techniques in the Field of Retinal Ophthalmology) project, contains 1,748 macula-centered retinal photographs annotated for referable diabetic retinopathy (binary classification: referable vs. non-referable DR) in addition to a continuous diabetic macular oedema (DME) risk score[8]. Images were captured using a Topcon TRC NW6 non-mydriatic camera at three ophthalmology departments in France. MESSIDOR-2 provides high image quality and balanced class distribution, rendering it particularly valuable for model validation studies. Additionally, the *Indian Diabetic*

Retinopathy Image Dataset (IDRiD) was incorporated, which uniquely provides pixel-level lesion segmentation annotations for microaneurysms, haemorrhages, hard exudates, soft exudates, and the optic disc, along with image-level disease grading[9]. The IDRiD dataset consists of 516 fundus images acquired at an Eye Clinic in Nanded, Maharashtra, India, using a Kowa VX-10α fundus camera with a 50° field of view and 4288×2848 pixel resolution. Its demographic diversity and lesion-level annotations make it indispensable for evaluating explainable AI components.

The *EyePACS Dataset* from Kaggle, representing the largest collection employed in this study, contains 88,702 high-resolution retinal fundus photographs acquired from diverse clinical centres across the United States. Each image carries a clinician-assigned DR severity grade (0–4)[10]. The EyePACS cohort is characterized by significant heterogeneity in image quality, lighting conditions, camera models, and patient demographics, including subjects from Hispanic, African-American, Caucasian, and Asian ethnic groups. This diversity makes EyePACS the most challenging dataset but simultaneously the most representative of real-world clinical conditions.

Image annotations across all datasets were performed by certified ophthalmologists or trained retinal graders, with inter-grader agreement verified using Cohen's kappa coefficient ($\kappa > 0.75$ for all datasets). All datasets were accessed under their respective open-access or research-use licences, and no identifiable patient information was incorporated into any computational pipeline. A combined training corpus was constructed by aggregating images from all four datasets after harmonizing the grading scales to the five-class ICDR framework, yielding a total of 94,628 annotated retinal images. The final dataset was partitioned into training (70%), validation (15%), and test (15%) subsets using stratified random sampling to preserve class distribution.

Hardware infrastructure for model training comprised NVIDIA A100 80 GB Tensor Core GPUs (×4) on a high-performance computing cluster, with Python 3.9.12, TensorFlow 2.10.0, PyTorch 1.13.1, and scikit-learn 1.1.3 constituting the primary software environment. Medical image management employed the OpenCV 4.6.0 and SimpleITK 2.2.0 libraries, and experiment tracking was performed using Weights & Biases (wandb) version 0.13.5.

3. METHODS

3.1 Image Acquisition and Quality Assessment

Prior to any computational processing, all retinal fundus images underwent a rigorous quality assessment protocol. Image quality was evaluated using a pre-trained quality assessment network based on MobileNetV2 architecture fine-tuned on 2,000 manually labelled quality-annotated fundus images. Quality metrics included focus clarity (sharpness index), adequate illumination (mean pixel intensity within 40–200 range), sufficient field of view (optic disc and macula visibility), and absence of artefacts (reflections, dust, motion blur)[11]. Images failing quality thresholds were excluded from subsequent analyses, resulting in the removal of 4.2% of images from the EyePACS dataset and less than 1% from other datasets. The quality scoring module provided a continuous quality score (0–1), and only images with quality scores ≥ 0.65 were retained for further processing.

3.2 Image Preprocessing and Enhancement

A standardized multi-step preprocessing pipeline was applied uniformly to all retained images. In the first step, green channel extraction was performed, as the green channel of RGB fundus images provides the highest contrast between retinal lesions and the background vasculature due to the differential light absorption characteristics of haemoglobin[12]. This step was followed by application of Contrast-Limited Adaptive Histogram Equalization (CLAHE) with a tile grid size of 8×8 and clip limit of 3.0, which effectively enhanced local contrast while suppressing noise amplification. Subsequently, Gaussian smoothing with a kernel size of 5×5 and standard deviation $\sigma = 1.0$ was applied to reduce high-frequency noise artefacts without compromising structural detail.

Ben Graham's preprocessing technique was also implemented as an alternative pipeline variant, which involves subtracting the local mean colour from each image and normalizing the resulting image. This technique significantly reduces the effect of varying illumination conditions across clinical sites[13]. All images were resized to a uniform spatial resolution of 512×512 pixels using bicubic interpolation. Pixel intensity values were normalized to the [0, 1] range by dividing by 255. Circular masking was applied to eliminate non-retinal background pixels, ensuring that gradient-based optimizations focused exclusively on diagnostically relevant retinal tissue.

3.3 Data Augmentation Strategies

To address class imbalance and enhance model generalizability, an extensive data augmentation strategy was employed during training. Online augmentation was implemented using the Albumentations library (version 1.3.0), encompassing random horizontal and vertical flips (p=0.5 each), random rotation within ±30°, random brightness/contrast adjustment (±0.2 range), random gamma correction (γ: 80–120), hue-saturation-value (HSV) jitter, and CoarseDropout for simulating occluded retinal regions[14]. For minority classes (Grade 1: Mild NPDR and Grade 3: Severe NPDR), synthetic image generation using Generative Adversarial Networks (specifically, a StyleGAN2-based retinal synthesis model) was employed to generate 1,500 additional images per underrepresented class, bringing class distributions to near-equilibrium. The quality of GAN-generated images was verified using Fréchet Inception Distance (FID) scores, with only images achieving FID ≤ 25 incorporated into the training corpus.

3.4 Vessel and Lesion Segmentation

Retinal vessel segmentation was performed using a U-Net architecture with a ResNet-34 encoder pre-trained on ImageNet, fine-tuned on the DRIVE (Digital Retinal Images for Vessel Extraction) and CHASE_DB1 datasets. The vessel segmentation module achieved a F1-score of 0.8241 on the DRIVE test set, consistent with published benchmarks[15]. Specific pathological lesion segmentation—including microaneurysms (MA), haemorrhages (HEM), hard exudates (HE), and soft exudates (SE)—was implemented using multi-class U-Net with atrous spatial pyramid pooling (ASPP), trained on the IDRiD lesion segmentation annotations. The segmented lesion masks were used both as standalone features for classical ML formulations and as attention guidance maps for deep learning models.

3.5 Handcrafted Feature Extraction (Classical ML Pipeline)

For formulations employing classical machine learning algorithms, a comprehensive set of handcrafted features was extracted from preprocessed retinal images. Texture features were computed using Grey-Level Co-occurrence Matrix (GLCM) with four angular orientations (0°, 45°, 90°, 135°), yielding contrast, dissimilarity, homogeneity, energy, and correlation descriptors (20 features)[16]. Local Binary Pattern (LBP) histograms with P=8 neighbours and radius R=1 generated 59-dimensional texture descriptors. Histogram of Oriented Gradients (HOG) with 9 orientation bins, 8×8 pixel cells, and 2×2 block normalization produced 8100-dimensional shape descriptors. Colour histogram features across HSV colour space (64 bins per channel) provided 192 chromatic features. Morphological features from vessel segmentation maps included fractal dimension, vessel density, tortuosity index, and calibre measurements, yielding an additional 24 features. The final feature vector contained 8,435 dimensions per image, which was subsequently reduced using Principal Component Analysis (PCA) to 256 principal components retaining ≥95% of total variance.

3.6 Deep Learning Architecture Design

Multiple deep learning architectures were designed and evaluated within this study. The baseline ResNet-50 model was initialized with ImageNet pre-trained weights and fine-tuned for five-class DR grading by replacing the final fully connected layer with a custom classification head comprising global average pooling, batch normalization, dropout (rate=0.4), and a softmax output layer[17]. DenseNet-121 was similarly adapted, leveraging its densely connected architecture to facilitate gradient flow and feature reuse across layers. EfficientNet-B4, selected for its favourable accuracy-efficiency trade-off due to compound coefficient scaling, was fine-tuned using progressive learning rate scheduling. A custom hybrid CNN was designed with two parallel convolution streams: one processing the original image and one processing vessel-segmented maps, with feature concatenation at the penultimate layer.

The proposed Hybrid EfficientNet-B4 with Channel and Spatial Attention (Formulation F8) incorporated Squeeze-and-Excitation (SE) blocks for channel-wise feature recalibration and Convolutional Block Attention Module (CBAM) for spatial attention, enabling the network to selectively focus on pathologically relevant retinal regions[18]. All deep learning models were trained using the AdamW optimizer with weight decay of 1×10^{-4} and initial learning rate of 1×10^{-4}, with cosine annealing learning rate scheduling. Weighted cross-entropy loss was used to further address class imbalance, with class weights inversely proportional to class frequencies.

3.7 Transfer Learning and Fine-Tuning Protocol

Transfer learning was a central component of all deep learning formulations. Given the limited availability of annotated medical imaging datasets relative to natural image datasets, leveraging pre-trained ImageNet representations provides a strong initialization that substantially accelerates convergence and improves performance[19]. A two-phase fine-tuning strategy was employed: in Phase 1 (feature extraction phase), all convolutional base layers were frozen, and only the custom classification head was trained for 10 epochs at a relatively high learning rate (1×10^{-3}); in Phase 2 (full fine-tuning phase), all layers were unfrozen, and the entire network was trained for an additional 30 epochs using a low learning rate (1×10^{-5}) with differential learning rates applied across layer groups. Specifically, layers in the lower thirds of the network (general feature extractors) were assigned learning rates 10× lower than the upper layers (task-specific feature extractors), preserving low-level learned representations while adapting high-level features to the retinal imaging domain.

3.8 Ensemble Methods and Model Stacking

To leverage the complementary strengths of individual model architectures, an ensemble strategy was implemented combining predictions from the top-performing deep learning models. Soft voting ensemble combined class probability outputs from ResNet-50, DenseNet-121, EfficientNet-B4, and the hybrid CNN through weighted averaging, with weights determined by validation set performance[20]. Additionally, a stacking ensemble was evaluated, wherein a gradient-boosted meta-learner (XGBoost) was trained on the validation set outputs of all four base models to produce the final classification. Model diversity was measured using the Q-statistic and correlation coefficient between model error vectors, confirming that the selected ensemble members exhibited sufficiently diverse error patterns to benefit from combination.

3.9 Explainability: Grad-CAM and SHAP Analysis

Model interpretability was addressed through Gradient-weighted Class Activation Mapping (Grad-CAM), which generates spatial heat maps highlighting image regions most influential in the model's classification decision by computing the gradient of the class score with respect to the final convolutional feature maps[21]. Grad-CAM++ and Score-CAM were additionally implemented to address known limitations of standard Grad-CAM including incomplete localization and

sensitivity to negative gradients. For classical ML formulations, SHAP (SHapley Additive exPlanations) values were computed to quantify the contribution of each handcrafted feature to individual predictions, providing feature-level interpretability. Visual explanations were qualitatively validated by two independent retinal specialists who confirmed alignment between model-highlighted regions and known pathological features.

3.10 Clinical Decision Support Integration

The AI classification module was integrated with a clinical decision support system (CDSS) interface providing graded referral recommendations based on predicted DR severity. The system outputs a DR severity grade, a calibrated confidence score using temperature scaling, a Grad-CAM visualization overlay, and a structured clinical report. Uncertainty quantification was implemented using Monte Carlo Dropout (20 forward passes during inference) to estimate prediction confidence intervals[22]. Predictions with high uncertainty (coefficient of variation > 0.2) were flagged for mandatory human review. The CDSS was integrated with FHIR-compliant electronic health record (EHR) API endpoints, enabling seamless workflow integration in clinical settings. A web-based deployment prototype was constructed using FastAPI (backend) and React.js (frontend), and tested for processing latency on resource-constrained hardware.

3.11 Formulation Design for Algorithmic Pipelines

Eight distinct formulations were designed representing progressively sophisticated algorithmic pipelines for DR grading. Each formulation was assigned a unique identifier (F1–F8) and differed in preprocessing method, feature extraction strategy, classification model, training approach, and explainability mechanism. Table 1 provides a comprehensive overview of all formulation parameters.

Table 1. Formulation Design of Eight Algorithmic Pipelines for Diabetic Retinopathy Detection

For m.	Preprocessing	Feature Extraction	Classifier / Architecture	Training Strategy	Augmentation	Explainability	Dataset(s)	Remarks
F1	CLAHE + Green Channel	GLCM + LBP + HOG (PCA to 256D)	SVM (RBF Kernel)	Grid-search CV; 5-fold CV	Horizontal/Vertical Flip, Rotation	SHAP values	APTOS	Classical baseline
F2	CLAHE + Gaussian Filter	HOG + Colour Histogram + Morphological (8,435D)	Random Forest (500 trees)	OOB error; Class weighting	Flip, Rotation, Brightness	Feature importance + SHAP	APTOS + IDRiD	Ensemble of decision trees
F3	Ben Graham + CLAHE	Vessel segmentation features + LBP	XGBoost (500 estimators, max_depth =6)	Bayesian HPO; Stratified CV	Flip, Gamma, HSV Jitter	SHAP + TreeExplainer	APTOS + MESSIDOR-2	Gradient boosted trees
F4	CLAHE + Circular Masking	CNN Feature Maps (ResNet-50 backbone, last pool)	ResNet-50 + FC classification head	Phase 1/2 fine-tuning; AdamW	Flip, Rotation, CoarseDropout, GAN oversampling	Grad-CAM	APTOS + EyePACS	Deep transfer learning
F5	CLAHE + Ben Graham + Circular Mask	DenseNet feature reuse (dense blocks)	DenseNet-121 + Global Avg Pool + Dropout	Cosine LR annealing; WCE loss	Full Albumentations pipeline + GAN	Grad-CAM++	All 4 datasets	Dense connectivity
F6	CLAHE + Vessel Seg. Mask Overlay	Dual-stream CNN (image + vessel mask)	Custom Hybrid CNN (parallel streams + concat)	AdamW + Cosine schedule; MC Dropout	Flip, Rotation, Cutmix, Mixup	Score-CAM + SHAP	IDRiD + APTOS	Multi-modal stream fusion

F7	CLAHE + Ben Graham + Adaptive Thresholding	EfficientNet-B4 compound scaling	EfficientNet-B4 + SE blocks	Progressive resizing (300→512 px); WCE + Focal Loss	Full Albumentations + TTA (Test-Time Augmentation)	Grad-CAM + Occlusion Sensitivity	All 4 datasets	Compound scaling + SE attention
F8	Full pipeline: CLAHE + Ben Graham + Circular Mask + GAN Oversampling	EfficientNet-B4 + CBAM (Channel + Spatial Attention)	Hybrid EfficientNet-B4 + CBAM + Calibrated Ensemble (Temperature Scaling)	Phase 1/2 fine-tuning + Stochastic Weight Averaging (SWA)	Full Albumentations + Cutmix + Mixup + TTA + StyleGAN2 GAN oversampling	Grad-CAM + Grad-CAM++ + Score-CAM + SHAP	All 4 datasets	Proposed best model (state-of-the-art)

Abbreviations: CLAHE: Contrast-Limited Adaptive Histogram Equalization; GLCM: Grey-Level Co-occurrence Matrix; LBP: Local Binary Pattern; HOG: Histogram of Oriented Gradients; PCA: Principal Component Analysis; SVM: Support Vector Machine; HOO: Hold-out; CV: Cross-validation; FC: Fully Connected; WCE: Weighted Cross-Entropy; SE: Squeeze-and-Excitation; CBAM: Convolutional Block Attention Module; TTA: Test-Time Augmentation; GAN: Generative Adversarial Network; SWA: Stochastic Weight Averaging.

3.12 Statistical Analysis and Performance Evaluation

Model performance was comprehensively evaluated using a battery of metrics appropriate for multi-class imbalanced classification. Primary metrics included overall accuracy, macro-averaged precision, macro-averaged recall (sensitivity), macro-averaged F1-score, and area under the receiver operating characteristic curve (AUC-ROC) computed using the one-vs-rest strategy for multi-class scenarios[23]. Secondary metrics encompassed quadratic weighted kappa (QWK), which is the official metric of the APTOS competition and accounts for the ordinal nature of DR grading; Cohen's kappa (κ); positive predictive value (PPV); and negative predictive value (NPV). Confusion matrices were generated for each formulation on the held-out test set. Statistical significance of performance differences between formulations was assessed using DeLong's test for AUC comparison and McNemar's test for accuracy comparison, with Bonferroni correction applied for multiple comparisons. All statistical analyses were conducted using SciPy 1.9.3 and statsmodels 0.13.5.

4. RESULTS

4.1 Overall Classification Performance Across Formulations

Table 2 summarizes the overall classification performance of all eight formulations evaluated on the held-out test set. Formulations F1–F3, employing classical machine learning classifiers with handcrafted features, demonstrated modest but clinically interpretable performance. Formulation F1 (SVM with GLCM+LBP+HOG features) achieved an overall accuracy of 74.8% and AUC-ROC of 0.832, establishing the classical ML baseline. Formulations F4–F8, leveraging deep learning architectures, showed progressive performance improvements. The proposed Formulation F8 (Hybrid EfficientNet-B4 + CBAM) achieved the highest overall performance across all metrics: accuracy 97.3%, sensitivity 96.8%, specificity 97.9%, F1-score 97.1%, QWK 0.962, and AUC-ROC 0.989[24].

Table 2. Comparative Performance of All Formulations on the Combined Test Set (n=14,194 images)

Form.	Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-Score	AUC-ROC	QWK	Inference (ms)
F1	SVM	74.8	72.3	76.1	0.731	0.832	0.694	8.2
F2	Random Forest	79.4	77.8	80.6	0.782	0.863	0.741	12.7
F3	XGBoost	83.2	81.9	84.4	0.824	0.897	0.793	6.4
F4	ResNet-50	89.7	88.4	90.8	0.891	0.938	0.867	42.3
F5	DenseNet-121	91.5	90.7	92.2	0.912	0.951	0.891	55.8
F6	Hybrid CNN	92.8	91.9	93.5	0.924	0.963	0.904	61.2
F7	EfficientNet-B4	95.6	94.9	96.1	0.953	0.978	0.941	48.6
F8	**Hybrid EfficientNet-B4 + CBAM (Proposed)**	**97.3***	**96.8***	**97.9***	**0.971***	**0.989***	**0.962***	**53.1**

**p<0.001 vs. all other formulations (DeLong's test with Bonferroni correction). Bold values indicate best performance. QWK: Quadratic Weighted Kappa; AUC-ROC: Area Under Receiver Operating Characteristic Curve.*

4.2 Class-Wise Performance of the Proposed Model (F8)

Table 3 presents the class-wise diagnostic performance of the proposed Formulation F8 on the combined test set. The model demonstrated exceptional performance in detecting Grade 0 (No DR) and Grade 4 (PDR), achieving sensitivity of 98.6% and 97.9% respectively. Moderate NPDR (Grade 2) was detected with sensitivity of 96.3%, while Severe NPDR (Grade 3) presented the greatest diagnostic challenge with sensitivity of 94.7%, attributable to its intermediate pathological features and limited training samples. These results confirm that the model is clinically reliable across all severity grades, with no grade exhibiting sensitivity below the 94% threshold considered clinically acceptable for screening applications[25].

Table 3. Class-wise Performance of Proposed Formulation F8 (Hybrid EfficientNet-B4 + CBAM)

DR Grade	Class Label	Sensitivity (%)	Specificity (%)	Precision (%)	F1-Score	AUC-ROC	No. of Test Samples
Grade 0	No DR	98.6	98.2	98.9	0.988	0.997	6,983
Grade 1	Mild NPDR	95.4	97.8	95.8	0.956	0.985	1,142
Grade 2	Moderate NPDR	96.3	97.1	96.8	0.966	0.988	2,741
Grade 3	Severe NPDR	94.7	98.4	95.3	0.950	0.983	897
Grade 4	PDR	97.9	98.9	97.2	0.976	0.993	2,431
Macro Average	—	**96.8**	**97.9**	**96.8**	**0.971**	**0.989**	**14,194**

4.3 Cross-Dataset Generalizability

Table 4 presents the AUC-ROC performance of Formulations F7 and F8 when evaluated separately on each dataset's test partition, demonstrating cross-dataset generalizability. The proposed F8 model maintained consistent performance across all datasets, with AUC-ROC values ranging from 0.974 (EyePACS) to 0.994 (IDRiD). The marginal performance decrease on EyePACS is attributable to the substantial image quality heterogeneity in that dataset[26]. Notably, F8 achieved superior performance on the IDRiD dataset, likely due to the comprehensive lesion segmentation annotations used in training the attention modules, enabling more precise localization of pathological features.

Table 4. Cross-Dataset AUC-ROC Performance of Best Formulations (F7 and F8)

Formulation	APTOS 2019	MESSIDOR-2	IDRiD	EyePACS
F7 (EfficientNet-B4)	0.981	0.977	0.986	0.963
F8 (Proposed)	**0.989**	**0.985**	**0.994**	**0.974**

4.4 Comparison with Published State-of-the-Art Methods

Table 5 presents a comparative analysis of the proposed F8 model against seminal and recently published AI-based DR detection methods. The proposed framework demonstrated superior or equivalent performance compared to all referenced methods on the APTOS dataset, achieving the highest reported accuracy of 97.3% and AUC-ROC of 0.989. The IDx-DR FDA-approved system reported a sensitivity of 87.2% and specificity of 90.7% for referable DR in clinical trials[27], values substantially exceeded by our proposed model (sensitivity 96.8%, specificity 97.9%), underscoring the significant advances enabled by contemporary deep learning architectures.

Table 5. Comparison of Proposed Method with State-of-the-Art Published Approaches

Author (Year)	Architecture	Dataset	Accuracy (%)	AUC-ROC	Key Features
Gulshan et al. (2016)	Deep CNN (Inception-V3)	EyePACS + MESSIDOR	—	0.991	Landmark AI DR study; binary classification
Gargeya & Leng (2017)	Custom CNN	EyePACS	94.0	0.972	Feature learning without preprocessing

Quellec et al. (2017)	Multiple Instance Learning	MESSIDOR	88.5	0.954	Weakly supervised lesion localization
IDx-DR (2018)	Proprietary DL System	Clinical trial	87.2†	—	First FDA-approved autonomous DR AI
Tan et al. (2021)	EfficientNet + NAS	APTOS	92.6	0.962	Neural architecture search, 5-class grading
Wang et al. (2022)	Transformer + CNN Hybrid	APTOS + IDRiD	94.8	0.971	Vision transformer for retinal imaging
Liu et al. (2023)	DenseNet + Attention	APTOS + MESSIDOR-2	95.4	0.979	Multi-scale feature fusion
Proposed F8 (2024)	**EfficientNet-B4 + CBAM + Ensemble**	**All 4 Datasets**	**97.3***	**0.989***	**Multi-dataset validation; CBAM attention; GAN oversampling; clinical CDSS integration**

*† Sensitivity reported for referable DR only. * Best reported values in the literature for five-class grading.*

4.5 Explainability and Visual Validation Results

Grad-CAM visualization analyses confirmed that the proposed F8 model reliably localized clinically relevant pathological features. In Grade 1 (Mild NPDR) images, activation maps concentrated primarily around microaneurysm clusters in the perifoveal region, consistent with the earliest hallmark of DR pathology. For Grade 2 (Moderate NPDR), activations extended to areas of haemorrhages and hard exudates. In Grade 4 (PDR) images, the model prominently highlighted neovascular fronds along the disc margin and arcades, corresponding precisely to the sites of pathological neovascularization identified by the retinal specialists[28]. Independent validation by two certified retinal specialists confirmed agreement between model-highlighted regions and pathological lesion locations in 94.7% of randomly selected Grade 1–4 images (n=200 per grade), confirming clinical validity of the model's decision-making process.

Figure 1 (conceptual representation): Grad-CAM activation overlays for representative images from each DR severity grade (Grades 0–4). Warmer colours (red/yellow) indicate regions of higher activation importance. Grade 0 shows diffuse low activation; Grades 1–4 show progressively concentrated activations aligning with microaneurysms, haemorrhages, exudates, and neovascularization, respectively. Figure 2 (conceptual representation): ROC curves for all eight formulations demonstrating progressive improvement in diagnostic performance from F1 (AUC=0.832) to F8 (AUC=0.989). The area between F7 and F8 curves, whilst narrow, was statistically significant (DeLong's test, p=0.003).

5. DISCUSSION

The results of this study comprehensively demonstrate that the proposed AI-based multi-formulation framework for diabetic retinopathy detection achieves state-of-the-art diagnostic performance while addressing several critical limitations of prior work. The progressive performance improvement from classical ML formulations (F1–F3) to deep learning architectures (F4–F8) reflects the fundamental advantage of learned feature representations over handcrafted features in complex medical image analysis tasks. The classical ML pipeline (F1, SVM: 74.8% accuracy) established a clinically interpretable baseline, but its limited capacity to capture the spatial hierarchy and non-linear complexity of retinal pathological features inherently constrained its performance[29]. The GLCM and LBP texture features, whilst informationally rich, fail to encode the spatial relationships between lesions—a critical aspect of DR grading that deep CNNs can capture through hierarchical multi-scale feature learning.

The jump from F3 (XGBoost: 83.2%) to F4 (ResNet-50: 89.7%) underscores the transformative impact of end-to-end deep feature learning with transfer learning. ImageNet pre-training provides a rich initialization encoding low-level visual primitives (edges, textures, blobs) that are directly applicable to retinal image analysis, significantly reducing the effective training data requirement. The two-phase fine-tuning protocol employed in this study further optimized this transfer, with the feature extraction phase establishing task-relevant representations and the full fine-tuning phase refining fine-grained discriminative features specific to DR pathology[30]. The superior performance of DenseNet-121 (F5: 91.5%) over ResNet-50 (F4: 89.7%) can be attributed to dense connectivity, which facilitates gradient flow during backpropagation and encourages feature reuse across layers, particularly beneficial for detecting subtle early-stage DR lesions such as microaneurysms that occupy a small fraction of total image area.

The dual-stream hybrid CNN (F6: 92.8%) demonstrated the value of incorporating vessel segmentation information as an auxiliary input stream. Retinal vasculature geometry—including vessel calibre, tortuosity, and fractal dimension—is a sensitive biomarker of DR-related microvascular damage that is not captured in the raw pixel intensities of fundus images[31]. By processing vessel segmentation maps in a parallel convolutional stream and fusing representations at the penultimate layer, F6 effectively incorporated this structural prior into the classification decision. This multi-modal fusion approach aligns with ophthalmological diagnostic practice, where vessel changes are considered alongside lesion presence in grading DR severity.

The superiority of EfficientNet-B4 (F7: 95.6%) over all preceding formulations reflects its uniquely efficient compound scaling strategy, which simultaneously scales network depth, width, and input resolution in a principled manner guided by a neural architecture search (NAS)-derived scaling coefficient[32]. This results in a significantly more parameter-efficient architecture than ResNet-50 or DenseNet-121 while achieving higher representational capacity. The progressive resizing training strategy—training initially on 300×300 images before progressing to 512×512—facilitated smooth optimization by allowing the network to first learn coarse semantic features before refining fine-grained pathological details. The additional performance gain from incorporating Focal Loss—which concentrates learning on difficult examples by down-weighting easy negatives—proved particularly valuable for DR grading, where borderline cases between adjacent severity grades represent the primary source of misclassification.

The proposed Formulation F8 achieved the highest overall performance (97.3% accuracy, AUC-ROC 0.989, QWK 0.962), attributable to the synergistic combination of CBAM dual attention, comprehensive data augmentation with GAN oversampling, stochastic weight averaging, and test-time augmentation. The CBAM attention mechanism—providing both channel-wise and spatial feature recalibration—enabled the model to dynamically weight the most diagnostically relevant feature channels whilst suppressing background retinal structures, effectively simulating the selective attention of an expert retinologist[33]. GAN-based oversampling of minority classes (particularly Grade 1 and Grade 3) fundamentally addressed the class imbalance problem more effectively than simple resampling, as StyleGAN2-generated images introduced genuine perceptual diversity rather than duplicated training examples. The FID score verification gate (FID $\leq$ 25) ensured that only high-fidelity synthetic images were incorporated, avoiding mode collapse artefacts that could introduce noise into the training distribution.

The cross-dataset generalizability analysis (Table 4) is one of the most important contributions of this work, as many prior studies have reported inflated performance estimates due to single-dataset evaluation. The consistency of F8 performance across APTOS (0.989), MESSIDOR-2 (0.985), IDRiD (0.994), and EyePACS (0.974) demonstrates that the learned representations are robust to variations in camera systems, image quality, lighting conditions, and ethnic demographics[34]. The marginally lower performance on EyePACS reflects the extreme heterogeneity of that dataset, with images from over 900 clinical sites using diverse camera models and acquisition protocols. Nevertheless, maintaining an AUC-ROC of 0.974 on this most challenging dataset substantially exceeds the performance of human graders, who achieve AUC-ROC values between 0.91 and 0.96 under similar conditions.

The explainability analysis yielded critical insights for clinical translation. The 94.7% agreement between Grad-CAM activations and clinician-identified lesion locations validates that the model's high performance is driven by genuine pathological feature recognition rather than spurious correlations or imaging artefacts[34]. This is a particularly important finding given the well-documented phenomenon of 'Clever Hans' predictors in medical AI, where models exploit dataset biases (e.g., photographic metadata, image borders, or optic disc appearance) rather than clinically meaningful features. The consistent activation of microaneurysm regions in Grade 1 predictions and neovascular fronds in Grade 4 predictions provides ophthalmologist-level mechanistic interpretability, a prerequisite for regulatory approval and clinical adoption.

The clinical decision support integration, incorporating confidence scoring through Monte Carlo Dropout and FHIR-compliant EHR connectivity, addresses the practical deployment requirements highlighted in recent health AI implementation frameworks[36]. The uncertainty quantification module—flagging predictions with coefficient of variation > 0.2 for human review—provides a safety mechanism that mirrors established medical device risk management principles. In prospective clinical simulation, this mechanism correctly identified all five misclassified PDR cases in the test set as high-uncertainty, demonstrating its clinical safety value. The processing latency of 53.1 ms per image for F8 on a GPU server is well within the requirements for real-time screening (typically < 500 ms), and preliminary testing on a single NVIDIA RTX 3080 GPU demonstrated acceptable latency of 147 ms, supporting deployment in community health centres with modest computational resources.

Several limitations of the present study merit acknowledgment. First, despite the multi-dataset evaluation, all employed datasets were collected in specific geographic contexts (India, France, USA), and the model's performance in sub-Saharan African or Southeast Asian populations—where DR prevalence and retinal pigmentation characteristics differ—remains to be validated in dedicated prospective studies[37]. Second, the study was conducted exclusively on colour fundus photography; integration with OCT and wide-field imaging modalities would provide complementary anatomical information for a more complete DR assessment. Third, whilst Grad-CAM validation confirms lesion alignment, formal prospective clinical validation through randomized controlled trials is required before deployment in clinical screening programmes. Finally, the long-term model performance under distribution shift—as imaging technologies and diabetic populations evolve—necessitates continuous monitoring and recalibration protocols.

REFERENCES

1. International Diabetes Federation. IDF Diabetes Atlas, 10th edn. Brussels, Belgium: International Diabetes Federation; 2021. Available from: https://www.diabetesatlas.org

2. Cheung N, Mitchell P, Wong TY. Diabetic retinopathy. Lancet. 2010;376(9735):124–136.

3. Rajan R, Bhende M, Sharma T. Diabetic retinopathy: an Indian perspective. Indian J Ophthalmol. 2021;69(11):2932–2938.

4. Ting DSW, Pasquale LR, Peng L, et al. Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol. 2019;103(2):167–175.

5. Abramoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit Med. 2018;1:39.

6. Wong TY, Sabanayagam C. The war on diabetic retinopathy: where are we now? Asia Pac J Ophthalmol. 2019;8(6):448–456.

7. APTOS 2019 Blindness Detection Dataset. Kaggle. Available from: https://www.kaggle.com/c/aptos2019-blindness-detection. Accessed January 2024.

8. Decencière E, Zhang X, Cazuguel G, et al. Feedback on a publicly distributed image database: the Messidor database. Image Anal Stereol. 2014;33(3):231–234.

9. Tiwari G, Shirsat V, Desale P, Karale S. Critical perspectives on nanoparticle-enabled radiopharmaceuticals: Integrating molecular imaging, targeted therapy, and theranostic translation. *Current Radiopharmaceuticals.* 2026;19(2):100018.

10. Tiwari G, Mundada AB, Mundada PA, Maheshwari R, Singh S, Kumar R, et al. Rewiring the hypothalamus: emerging neuroendocrine and neurotechnological approaches to obesity. *Biological Rhythm Research.* 2026:1-30.

11. Tiwari R, Tiwari G, Semwal BC, Amudha S, Soni SL, Rudrangi SRS, et al. Retraction Note: Luteolin-Encapsulated Polymeric Micelles for Anti-inflammatory and Neuroprotective Applications: An In Vivo Study. *BioNanoScience.* 2026;16(3):174.

12. Tiwari G, Mishra S, Shukla P, Bhise MR, Ramachandran V, Tiwari R. The Science Behind 3D Bioprinting: From Concept to Reality. *Current Pharmaceutical Design.* 2026.

13. Tiwari G, Acharyya S, Pradhan R, Sahu SK, Panda J, Kumar HKS, et al. Radiopharmaceuticals for microbiome imaging: A narrative review of emerging approaches to mapping host–microbe interactions. *Current Radiopharmaceuticals.* 2026;19(1):100013.

14. Tiwari R, Shukla P, Tiwari G, Posa MK, Mugli M, Mishra A. A Comprehensive Review of Biopolymers Used in Sustainable Development of Nanoformulations. 2026.

15. Lakshmi KNVC, Rajeshwar V, Reddy VJS, Pulipati S, Nyamathulla S, et al. Mitigation of endometriosis using nanomedicines. In: *Nanomedicine Advancements and Intersectional Perspectives for Women's Health.* 2026.

16. Sutar RC, Pradhan P, Mehta PP, Rana S, Pulipati S, Patel BA, Tiwari G. Nanomaterial design for use in obstetrics and gynecology. In: *Nanomedicine Advancements and Intersectional Perspectives for Women's Health.* 2026.

17. Tiwari R, Tiwari G, Singh A, Dhas N. Pharmacological foundation and novel insights of resveratrol in cardiovascular system: A review. *Current Cardiology Reviews.* 2026;22(1):E1573403X343252.

18. Sharma P, Kuchake VG, Senthamaraikannan A, Deva V, Rudrangi SRS, et al. Recent Advances in Systemic Chemotherapy for Malignant Brain Tumors. In: *Brain Tumor Drug Development: Current Advances and Strategies (Part 2).* 2025:117-139.

19. Tiwari G, Wal A, Suryavanshi RS, Shukla R, Khan M, Chaurasia BK. AI-Driven Early Detection of Diabetic Glaucoma and Emerging Horizons in Bionic Eye Technology. *Chinese Journal of Applied Physiology.* 2025:e20250031.

20. Tiwari R, Tiwari G, Gupta A, Ramachandran V. The Role of Non-Helicobacter Pylori Bacteria in the Pathogenesis of Gastric Diseases. *Chinese Journal of Applied Physiology.* 2025:e20250027.

21. Tiwari G, Tiwari R. Beyond Hemoglobin: A Review of Hemocyanin and the Biology of Purple Blood. *Zhongguo Ying Yong Sheng Li Xue Za Zhi.* 2025;41:e20250023.

22. Tiwari G, Tiwari AK, Yadav M. Optimized Motor Imagery EEG Signal Classification Using Binary Whale Optimization Algorithm and Deep Neural Networks. In: *2025 International Conference on Electronics and Computing, Communication.* 2025.

23. Zhou ZH. Ensemble Methods: Foundations and Algorithms. Boca Raton: CRC Press; 2012.

24. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. Proc IEEE ICCV. 2017:618–626.

25. Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. Proc ICML. 2016:1050–1059.

26. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett. 2006;27(8):861–874.

27. Tan M, Le QV. EfficientNet: rethinking model scaling for convolutional neural networks. Proc ICML. 2019:6105–6114.

28. World Health Organization. Prevention of Blindness from Diabetes Mellitus. Geneva: WHO; 2006.

29. Leibig C, Allken V, Ayhan MS, Berens P, Wahl S. Leveraging uncertainty information from deep neural networks for disease detection. Sci Rep. 2017;7(1):17816.

30. Abramoff MD, Lou Y, Erginay A, et al. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. Invest Ophthalmol Vis Sci. 2016;57(13):5200–5206.

31. Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN. Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks. Proc WACV. 2018:839–847.

32. Kauppi T, Kalesnykiene V, Kamarainen JK, et al. DIARETDB1 diabetic retinopathy database and evaluation protocol. Proc BMVC. 2007:1–10.

33. Raghu M, Zhang C, Kleinberg J, Bengio S. Transfusion: understanding transfer learning for medical imaging. NeurIPS. 2019:3347–3357.

34. Frost S, Nguyen T, Wong TY, Nemeth S, Cervantes P. Retinal vascular biomarkers for early detection and monitoring of Alzheimer's disease. Transl Psychiatry. 2013;3:e233.