VOLUMEN 48
XXXI Verano De la Ciencia
ISSN 2395-9797
www.jovenesenlaciencia.ugto.mx

Automatic analysis of diabetic retinopathy in fundus images using computer vision methods and deep learning

Análisis automático de retinopatía diabética en imágenes de fondo de ojo mediante métodos de visión por computadora y aprendizaje profundo

Christopher Ignacio Espinosa Razo, Fernando Daniel Hernandez-Gutierrez, Jose Ruiz-Pinales, Juan Gabriel Avina-Cervantes*
División De Ingenierías del Campus Irapuato-Salamanca (DICIS), Universidad de Guanajuato.
ci.espinosarazo@ugto.mx, fd.hernandezgutierrez@ugto.mx, pinales@ugto.mx, avina@ugto.mx

*Correspondence author

Abstract

Diabetic retinopathy (DR) is one of the leading causes of preventable vision loss worldwide. Due to the asymptomatic early stages of disease, early detection is crucial to reduce the risk of progression to severe retinal damage. In this work, we propose a computer-vision-based approach for the automatic analysis of DR in fundus images using deep learning techniques. The method follows a three-stage pipeline: image preprocessing, feature extraction, and multiclass classification. Preprocessing involves cropping black borders, enhancing images through green-channel CLAHE, and applying augmentation to improve robustness. Feature extraction is performed with a modified EfficientNetB0 backbone, where Group Normalization and GELU were employed. Finally, classification assigns each image to a DR severity grade using a class-balanced focal loss, thereby supporting early diagnosis and reducing the risk of vision loss. The system was evaluated on the APTOS 2019 dataset, which contains fundus images categorized into five classes (0–4) according to the International Clinical Diabetic Retinopathy (ICDR) scale. The proposed model achieved an accuracy of 0.79, precision of 0.81, recall of 0.79, F1-score of 0.79, and MCC of 0.68. Future work may include the use of larger and more diverse datasets and the integration of advanced preprocessing and feature extraction techniques.

Keywords: Diabetic retinopathy; Deep learning; Convolutional neural networks; EfficientNetB0; Multi-class eye disease classification.

Introduction

Approximately one-third of patients with type 2 diabetes screened in primary care settings in Mexico have some degree of diabetic retinopathy. A study reported a prevalence of 33.6%, comprising 24.2% non-sight-threatening and 9.4% sight-threatening retinopathy (Silva-Tinoco *et al.*, 2024). Screening refers to the systematic examination of at-risk individuals to detect disease at an early stage, before symptoms appear. Diabetic retinopathy often remains asymptomatic for many years, even at advanced stages, which delays detection and increases the risk of irreversible visual damage (Vaughan *et al.,* 2025). The effectiveness of screening is constrained not only by the limited availability of ophthalmologists but also by variability in their clinical expertise and the level of patient awareness, frequently resulting in delayed and inadequate follow-up care.

Diabetic retinopathy (DR) is clinically categorized into non-proliferative and proliferative forms that reflect disease severity. Non-proliferative DR (NPDR) includes mild, moderate, and severe stages: mild NPDR is characterized mainly by microaneurysms; moderate NPDR shows increased retinal hemorrhages, exudates, and possible cotton-wool spots; severe NPDR meets the “4-2-1” rule (severe intraretinal hemorrhages in four quadrants, venous beading in two or more quadrants, or prominent intraretinal microvascular abnormalities) and carries a high risk of progression to proliferation. Proliferative DR (PDR) is defined by neovascularization

of the retina or optic disc and/or vitreous/preretinal hemorrhage and represents the stage with the greatest risk for vision-threatening complications such as tractional retinal detachment (Flaxel *et al.,* 2020).

Figure 1 illustrates the severity stage grading of diabetic retinopathy. Automated grading of diabetic retinopathy offers clinical advantages, including enhanced screening efficiency, improved diagnostic reproducibility, expanded population coverage, and the mitigation of geographic and economic access barriers; deep learning has transformed DR screening and diagnosis in recent years (Akhtar *et al*., 2025). Gulshan *et al*. (2016) showed that a deep learning algorithm could identify diabetic retinopathy in fundus images with performance comparable to ophthalmologists, a finding that motivated subsequent work on automated DR grading.

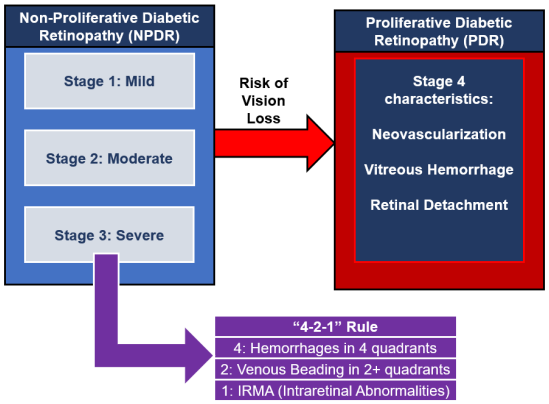


Figure 1. *Stage grading of diabetic retinopathy.*

Several studies have explored the simultaneous classification of multiple eye diseases from fundus images, which is clinically relevant given that different retinal pathologies can coexist or present similar visual features. For instance, Sah *et al.* (2026) proposed a dual-backbone deep learning framework that pairs EfficientNetB0 with three complementary architectures (ResNet50, InceptionV3 and AlexNet) to classify fundus images into four categories: normal, cataract, glaucoma and diabetic retinopathy. This multiclass method helps separate diabetic retinopathy (DR) from other eye diseases, but it doesn't solve the harder task of grading how severe DR is. That requires telling apart stages of the same disease that often look very similar. Among the CNN architectures used specifically for this task, EfficientNetB0 stands out for its computational efficiency, achieved through a compound scaling technique that balances network depth, width, and resolution (Vijayan & Ventakakrishnan, 2023).

This study proposes an EfficientNetB0-based deep learning model for the multiclass classification of diabetic retinopathy severity, evaluating not only overall accuracy but also complementary metrics that better capture model performance, such as recall, precision and F1-score. These measures provide deeper insight into how well the model identifies each severity stage, especially in cases where class imbalance may affect accuracy alone.

Literature Review

The automated analysis of Diabetic Retinopathy (DR) using deep learning techniques is a critical area of research in medical imaging. Given the progressive nature of the disease, accurate and early grading of fundus images is essential to prevent irreversible vision loss. To benchmark and evaluate the performance of neural network architectures, public datasets such as the Asia Pacific Tele-Ophthalmology Society Blindness Detection dataset (APTOS) 2019 have become the gold standard in literature. Recent studies have utilized this dataset to address DR classification as we do, setting a baseline for the methodology proposed in this work.

Aiche *et al*. (2026) proposed a deep ensemble transfer learning framework for the automated binary classification of diabetic retinopathy (DR) using CLAHE-enhanced APTOS 2019 fundus images. Their pipeline integrates transfer learning with ResNet50, VGG16, VGG19 and DenseNet121, combined with seven ensemble fusion strategies such as: hard voting, soft voting, weighted soft voting, rank-based fusion, Choquet-like integral, Sugeno integral, and average logits fusion. The best performance was achieved with the first three strategies, while weighted soft was assigned to similar weights due to comparable performance of individual classifiers. A major limitation of the proposed model is that it only performs binary classification, meaning it can only distinguish between a healthy eye and one with the disease.

Recently, Das & Majumder (2026) presented a customized DenseNet-121 optimized for multi-class classification of retinal fundus images. Most of the layers of the DenseNet were frozen to retain pre-trained weights and a custom classifier is added to the flatten layer. This classifier converted 2D feature maps into 1D vectors, followed by a dense layer with 128 neurons and ReLU activation, a dropout layer (rate = 0.4) to mitigate overfitting, and a final dense layer with five neurons and softmax activation for five-class classification. The most significant limitation of the proposed system is its difficulty in accurately grading intermediate stages of the disease.

Kobat et al. (2022) introduced a patch-based feature extraction approach for retinal fundus image classification. The images were subdivided into horizontal and vertical patches. A pre-trained DenseNet201 was employed to extract features from the main image, as well as from the horizontal and vertical patches. Neighborhood Component Analysis (NCA) was applied to select the top 500 discriminative features, which were subsequently fed into a Support Vector Machine (SVM) classifier for final prediction. The primary limitation of the proposed system is its dependence on high-quality retinal imaging. The researchers noted that when clinical symptoms are not clearly visible or the image quality is low, the model struggles to provide an accurate classification.

Yasmin *et al.* (2025) employed DenseNet201 as the backbone architecture for diabetic retinopathy detection. Most layers were frozen and deeper layers were fine-tuned to capture task-specific features. On top of the DenseNet backbone, a lightweight classification head was added, consisting of a dense layer with 128 units and ReLU activation, a dropout layer (rate = 0.4), and a final dense layer with five units and softmax activation for five-class classification. The limitation of the proposed DR-RetinaNet is the observed performance gap between its training accuracy (99%) and validation accuracy (94%), which indicates a degree of overfitting.

The last year Dib (2025), proposed a custom convolutional neural network (CNN) architecture featuring eight convolutional layers with 3x3 filters. Each of these layers integrates ReLU activation, 2x2 max-pooling, and batch normalization to ensure training stability and effective feature extraction. The classification head consists of two fully connected layers with 512 and 128 units, respectively, followed by a five-unit SoftMax layer for severity grading. Custom CNN's main limitation is its inability to achieve high performance without the aid of SMOTE, as it struggles to identify features in minority classes due to its lightweight architecture and lack of pre-trained knowledge.

Mohanty *et al.* (2023) developed a hybrid model for diabetic retinopathy detection by combining VGG16 with an XGBoost classifier. In this framework, deep features extracted from the last fully connected layer of VGG16 are used as input to the XGBoost classifier. The integration aims to exploit the strong feature extraction capabilities of CNNs together with the ensemble learning power of the gradient boosting. Given the original class imbalance of the dataset, the reported accuracy alone can be misleading, as it does not reveal how well the models perform on minority classes.

In a recent work, Datta & Bhandari (2026) introduced a modified ResNet-50 architecture that was combined with preprocessing techniques such as contrast-limited adaptative histogram equalization (CLAHE) and enhanced super-resolution generative adversarial networks (ERSGAN) to improve feature representation from retinal fundus images. The primary constraint of this framework is its significant reliance on advanced image enhancement techniques to achieve reliable results. While the integration of ESRGAN and CLAHE effectively highlights clinical features, the model struggles to maintain accuracy when processing raw or minimally enhanced images

Shakibania *et al.* (2024) proposed a dual-branch deep learning network for the detection and stage grading of diabetic retinopathy. The architecture utilizes pre-trained ResNet-50 and EfficientNetB0 models as parallel

feature extraction backbones. The classification head was designed as a fully connected neural network comprising four linear layers with 1364, 512, and 32 units, respectively. To make the model more stable and avoid overfitting, batch normalization was applied, dropout regularization was added with a rate of 0.25, and Tanh activation was used to introduce non-linearity. The main limitation of the proposed system is its difficulty in accurately grading advanced stages of the disease, in this approach is its notably low recall in detecting the 'Severe' stage.

Youldash *et al.* (2024) employed the EfficientNetB3 architecture as backbone for multi-class diabetic retinopathy classification. A preprocessing pipeline was followed that included green channel extraction, contrast-limited adaptative histogram equalization (CLAHE) and Gaussian filtering. To adapt the architecture, a GlobalAveragePooling2D layer was added to the base network, followed by a dense output layer with Softmax activation for five-stage severity grading. A key technical limitation of the EfficientNetB3 model in multi-class tasks is the difficulty to separate, due to a notable overlap between the Moderate, Severe and Proliferative categories.

A lightweight framework is developed by Dejene *et al.* (2026) for diabetic retinopathy detection by knowledge distillation (KD). In this study, a larger, pre-trained MobileNet (with an alpha value of 0.25) served as the teacher model, which provided “soft labels” to guide the learning process. The student model consisted of a significantly reduced MobileNet-like architecture featuring five depthwise convolutional blocks helping to maintain important feature extraction ability while balancing computational efficiency and performance. A significant limitation is its performance decline as the diagnostics complexity increases. When moving from binary to ternary classification, the model struggles to clearly separate non-proliferative from proliferative stages.

Zafat *et al.* (2025) developed a two-stage lightweight framework based on a custom 37-layer convolutional neural network (CNN) architecture with a parallel branch. In the first stage, the model performs binary detection to distinguish between healthy and diseased fundus images; the second stage then employs transfer learning by reusing the pre-trained weights to subclassify the four degrees of disease severity. The primary limitation of this lightweight system is the difficulty in maintaining high accuracy during the multi-level severity identification stage. While it is extremely effective at binary detection, its ability to distinguish between specific grades is slightly reduced, especially in Moderate cases.

A proposed methodology by Awasthi *et al.* (2024) combined a Vision Transformer (ViT) model with Harris Hawk Optimization (HHO) to optimize hyperparameters in diabetic retinopathy detection. The approach enhances ViT’s feature extraction by leveraging its global context modeling capabilities. The main constraint of the ViT-HHO model is the significant computational power required by the Vision Transformer architecture, posing a challenge for real-world implementation in resource-constrained medical settings.

Verma *et al.* (2026) introduced a novel hybrid feature fusion framework for the binary classification of diabetic retinopathy using the APTOS 2019 dataset. The architecture integrates two parallel backbones: a customized AlexNet (CNN) to capture localized textures and a Swin Transformer (ViT) to model global spatial relationships. Both networks were adapted to generate 1024-dimensional feature vectors, which were subsequently concatenated into a unified 2048-dimensional retinal representation. The fused representation was employed for classification through a Random Forest classifier with 150 estimators. The model is currently limited to binary classification, failing to distinguish between the five clinical stages of severity.

Badar *et al.* (2025) presented MSCAS-Net (Multi-Scale Cross and Self-Attention Network) a methodology for fine-grained diabetic retinopathy classification. The architecture employs a Swin Transformer backbone to extract hierarchical features at three resolutions (12 × 12, 24 × 24, and 48 × 48), enabling the capture of both global retinal structures and local lesions. A dual attention mechanism is integrated, where self-attention refines dependencies within each scale, while cross-attention fuses global context with local details to construct a comprehensive representation. The unified feature vector is subsequently processed through Global Average Pooling (GAP) and finalized with a Softmax classification layer. The primary limitation of MSCAS-Net lies in the complex task of balancing the weights of features extracted across multiple scales; the researchers noted that the optimization of these mechanisms during model construction still needs refinement

Ainapur & Virupakshappa (2025) provided a methodology that combines advanced techniques such as lighting adjustment, gray scaling, and Gaussian blur. They introduced a novel binary Archimedes optimization with a

local search strategy (LS-NBA) and texture features were extracted using the Gray Level Co-occurrence Matrix (GLCM). The classification was performed by a Stacked Ensemble-based Logistic Adaboost Light Gradient (SE-LALG) model that includes include AdaBoost, LightGBM, Gradient Boosting, and Logistic Model Tree (LMT); A Logistic Regression model served as meta-learner, learning the optimal weights to combine the predictions from the four base models. Due to the use of multiple base models and a meta-learner, the system requires significantly longer training time. This can slow the development process and the ability to update the model quickly.

The methodology of Sah, *et al*. (2026) is based on a dual-backbone deep learning framework that integrates EfficientNetB0 with ResNet50 through weighted fusion strategy. In this approach, features are extracted concurrently from both models and compressed via a global average pooling into enhanced representation vectors; these are then combined using trainable weights that allow the system to learn the relative importance of features from each architecture. Since the system was built for multi-class eye disease detection such as glaucoma, cataract, normal and DR, it does not classify the specific severity levels or grades of diabetic retinopathy.

Table 1. *Associated methods for automated diabetic retinopathy analysis.*

Article	Dataset	Method	Classes	Limitation
(Aiche *et al*., 2026)	APTOS 2019	Ensemble (ResNet50, VGG16, VGG19, DenseNet121)	DR, No DR	The model cannot determine how advanced the disease is.
(Das & Majumder, 2026)	APTOS 2019	DenseNet-121	No DR, Mild, Moderate, Severe, Proliferative	The model's performance is limited, particularly for the Moderate class, resulting in lower precision and recall.
(Kobat *et al*., 2022)	APTOS 2019	DenseNet201	Normal, Mild DR, Moderate DR, Severe DR, Proliferative DR	The model's accuracy is significantly affected by fundus-image quality and symptom visibility.
(Yasmin *et al.*, 2025)	APTOS 2019	DR_RetinaNet	No DR, Mild, Moderate, Severe, Proliferative DR	The system struggles with the Moderate stage because its clinical features often overlap with Mild or Severe cases.
(Dib, 2025)	APTOS 2019	Custom CNN	No DR, Mild, Moderate, Severe, Proliferative	The model is highly sensitive to class imbalance.
(Mohanty *et al*., 2023)	APTOS 2019	Hybrid network (VGG16 + XGBoost)	DR, No DR	The model is evaluated using accuracy alone.
(Datta & Bhandari, 2026)	APTOS 2019, DDR, DRIVE, local dataset	ResNet50	Normal, Mild, Moderate, Severe, Proliferative	The model's high performance depends heavily on a complex, multistage enhancement pipeline involving ESRGAN and CLAHE.
(Shakibania *et al.*, 2024)	APTOS 2019	Dual-branch deep learning network (ResNet50 and EfficientNetB0)	No DR, Mild, Moderate, Severe, PDR	The model struggles to distinguish between Severe and Proliferative DR stages.

Article	Dataset	Method	Classes	Limitation
(Youldash *et al*., 2024)	APTOS 2019	EfficientNetB3	No DR, Mild, Moderate, Severe, Proliferative	The model suffers from substantial misclassification between adjacent stages in the five-class system.
(Dejene *et al*., 2026)	APTOS 2019	MobileNet	No DR, NPDR, PDR	Performance declines when moving from binary to ternary classification.
(Zafar *et al.,* 2025)	APTOS 2019	Custom CNNs	No DR, DR (Mild, Moderate, Severe, PDR)	The model has difficulty maintaining high accuracy for intermediate lesion stages.
(Awasthi *et al*., 2024)	APTOS 2019	Vision Transformer + Harris Hawks Optimization	No DR, Mild, Moderate, Severe, Proliferative	The method requires substantial computational resources, which may limit deployment in clinics with restricted hardware.
(Verma *et al.,* 2026)	APTOS 2019	AlexNet + Swin Transformer + Random Forest	Benign (No DR), Malignant (DR)	Classification is limited to two classes.
(Badar *et al*., 2025)	APTOS 2019	MSCAS-Net	No DR, Mild, Moderate, Severe, Proliferative	The model requires improved optimization to balance feature weights across different scales.
(Ainapur & Virupakshappa, 2025)	APTOS 2019	Binary Archimedes Optimization + stacked ensemble learning model (SE-LALG), including AdaBoost, Logistic Regression, and LightGBM	No DR, Mild, Moderate, Severe, Proliferative	The system requires a long training time.
(Sah *et al.,* 2026)	IDRiD, Ocular Recognition Database, Messidor-2, ODIR	EfficientNetB0 + ResNet50	Cataract, DR, Glaucoma, Normal	The system does not classify the specific severity levels of diabetic retinopathy.

Objective of the Study

The main objective of this work is to design and evaluate an EfficientNetB0-based deep learning model for multiclass classification of diabetic retinopathy severity, aiming to improve diagnostic reliability under class imbalance and provide clinically meaningful insights through comprehensive performance analysis.

Specific Objectives

- Develop a custom EfficientNetB0 model adapted for multiclass severity classification of diabetic retinopathy.
- Compare its performance against established baseline architectures such as ResNet, DenseNet, and ConvNeXtLike.
- Evaluate the model using diverse metrics (accuracy, macro-precision, macro-recall, macro-F1-score, and MCC) to capture different aspects of performance.
- Analyze confusion matrices to identify misclassification patterns across severity stages.
- Highlight the clinical relevance of metrics (recall, F-score, MCC) beyond overall accuracy.

Methodology

Dataset

The experiments were conducted using the APTOS 2019 dataset, which comprises a total of 3,662 retinal fundus images categorized into five classes (No DR, Mild DR, Moderate DR, Severe DR and Proliferative DR). The dataset was divided into training, validation, and test subsets to ensure robust model evaluation. The training set contained 2,930 images, the validation set included 366 images, while the test set comprised 366 images. This distribution corresponds to an approximately 80/10/10 split, consistent with the data-partitioning strategy commonly adopted in the literature for evaluating deep learning models on the APTOS 2019. This dataset is widely used in automated retinal image analysis because it provides a realistic benchmark for training and evaluating deep learning models under different imaging conditions.

Table 2 presents the dataset distribution of proposed splitting for each class and subset.The fundus images were preprocessed by removing the black peripheral borders to prevent non-retinal regions from influencing feature extraction. CLAHE was then applied to the luminance channel of the CIELAB color space using a clip limit of 3.0 and an 8 × 8 tile grid to enhance local contrast while limiting noise amplification. The images were resized to 512 × 512 pixels and normalized using ImageNet statistics. During training, data augmentation included random horizontal flipping with a probability of 0.5, random vertical flipping with a probability of 0.2, rotations within ±15°, and random adjustments in brightness and contrast of up to 15%, saturation of up to 10%, and hue of up to 0.02. These transformations were applied to increase image variability and reduce overfitting, whereas validation and test images underwent only deterministic preprocessing.

Table 2. *Dataset distribution through subsets.*

Subset	Count	Class 0	Class 1	Class 2	Class 3	Class 4
Train	2,930	1,434	300	808	154	234
Validation	366	172	40	104	22	28
Test	366	199	30	87	17	33
Total	3,662	1,805	370	999	193	295



Figure 2. *illustrates representative retinal fundus images corresponding to each stage of diabetic retinopathy severity in the APTOS 2019 dataset. This visual reference provides a clear overview of the clinical spectrum of the disease. Sample image for each class in APTOS 2019 dataset.*

Function Loss

To address the severe class imbalance across DR severity grades, the model was trained using a class-balanced focal loss (Soni *et al.,* 2025). This loss integrates two complementary components: first, the focal loss modulating factor $(1-p_t)^{\gamma}$, which acts as a dynamic weight that reduces the influence of examples already classified correctly (γ = 1.5); and second the class-balanced weighting scheme proposed by Cui *et al.*, (2019) where the weight assigned to each class is derived form the effective number of samples. The effective number of samples denoted as E_n is defined as:

$$E_n = \frac{1-\beta}{1-\beta^{n_y}}$$

Where β is a value that controls the importance given to the number of samples ($\beta = 0.9999$) and y is the ground- truth class. The full Class-Balanced Focal Loss formula is:

$$CB_{focal}(p_t, y) = -\frac{1-\beta}{1-\beta^{n_y}}(1-pt)^{\gamma} log(p_t)$$

Where the focal component $(1-pt)^{\gamma} log(p_t)$ let the model to focus on "hard" examples by reducing the weight of easy, well-classified samples. This approach allows the model to achieve better performance where a few classes dominate the dataset.

Preprocessing

Prior to being fed into the network, all fundus images were processed. First, black borders surrounding the retinal field of view were automatically removed. This was achieved by thresholding pixel intensity to detect the illuminated region and cropping it within a bounding box, while retaining a small margin to avoid cutting peripheral retinal structures. Because this preprocessing step relies solely on pixel intensity rather than class labels, it does not introduce any risk of label leakage. Second, contrast enhancement was applied on the green channel of the RGB image using Contrast Limited Adaptive Histogram Equalization (CLAHE). The green channel was selected because it provides the highest contrast between retinal structures such as microaneurysms, hemorrhages, and blood vessels against the background (Datta & Bhandari, 2026).

During training, a standard augmentation strategy was applied to improve generalization and mitigate overfitting: random horizontal flipping (p = 0.5), random vertical flipping (p = 0.2), random rotation (±15°), and mild color jittering (brightness, contrast, and saturation (± 0.10–0.15, hue (± 0.02). These augmentations increased the diversity of the training samples while preserving the clinical relevance of retinal features. These

augmentations increased the diversity of the training samples while preserving the clinical relevance of retinal features. *Figure 3* illustrates the fundus image preprocessing pipeline.

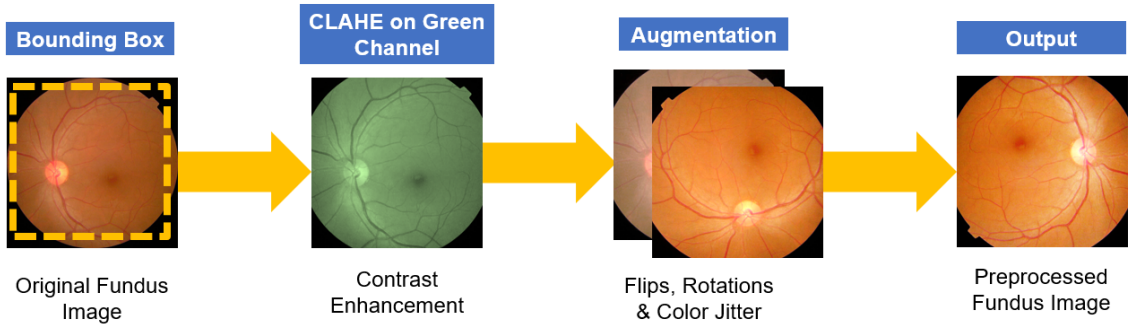


Figure 3. Fundus image preprocessing pipeline.

Proposed model

Figure 4 illustrates the proposed model, which is based on a modified EfficientNet-B0 architecture. The network preserves the main structure of the original EfficientNet-B0 backbone, including an initial stem block, seven mobile inverted bottleneck convolution (MBConv) stages with progressively increasing depth and channel width, a final convolutional block, and a classification head. As shown in the upper panel of *Figure 4*, the input image, with dimensions of 224 × 224 × 3, is first processed by the stem block. This block consists of a 3 × 3 convolution with a stride of 2, followed by Group Normalization with eight groups, producing a feature map of 112 × 112 × 32. Unlike the original EfficientNet-B0 architecture, which employs Batch Normalization, the proposed model consistently uses Group Normalization throughout the network.

The middle panel of *Figure 4* illustrates the internal structure of the MBConv blocks. When required, an expansion convolution first increases the channel dimension of the input feature map, followed by Group Normalization and a Gaussian Error Linear Unit (GELU) activation. The expanded representation is then processed by a depthwise convolution, followed by an additional Group Normalization layer and GELU activation. In contrast to the original EfficientNet formulation, which employs the SiLU/Swish activation function, the proposed architecture consistently uses GELU. This smooth activation function weights each input according to the cumulative distribution function of a standard Gaussian distribution. The resulting feature map is subsequently passed through a squeeze-and-excitation block, which also employs GELU internally, to adaptively recalibrate the channel-wise feature responses. A projection convolution then reduces the representation to the required number of output channels, after which Group Normalization is applied. Finally, a DropPath layer provides stochastic-depth regularization, and a residual connection adds the original block input to its output when the spatial dimensions and number of channels are compatible.

The bottom panel of *Figure 4* shows how the MBConv blocks are assembled to form the complete network. Following the stem block, the seven MBConv stages progressively transform the feature representation from 112 × 112 × 16 at the output of MBConv1 to 7 × 7 × 320 at the output of MBConv7. The numbers of block repetitions across the seven stages are 1, 2, 2, 3, 3, 4, and 1, respectively, consistent with the original EfficientNet-B0 specification. A final 1 × 1 convolution expands the channel dimension, producing a 7 × 7 × 1280 feature map. Global average pooling then aggregates the spatial information into a 1 × 1 × 1280 feature vector, which is passed to a linear classification layer with five outputs corresponding to the No DR, Mild DR, Moderate DR, Severe DR, and Proliferative DR classes.

VOLUMEN 48
XXXI Verano De la Ciencia
ISSN 2395-9797
www.jovenesenlaciencia.ugto.mx

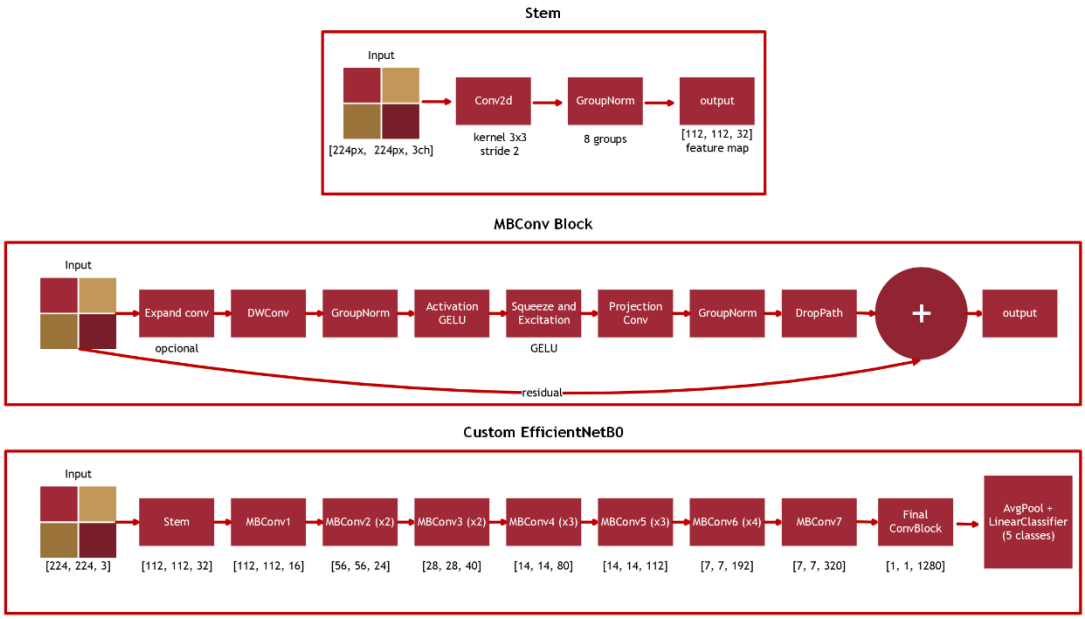


Figure 4. Proposed model of custom EfficientNetB0 architecture.

Results

Evaluation metrics

To evaluate the five-class classification of diabetic retinopathy severity, accuracy, macro-averaged precision, recall, F1-score, and multiclass Matthew's correlation coefficient (MCC) were considered. Because the dataset is imbalanced, accuracy alone may be dominated by the majority classes and conceal poor recognition of Severe and Proliferative DR. The macro-averaged metrics assign equal importance to each class, whereas MCC provides a global measure of agreement based on the complete confusion matrix. Since MCC does not indicate which classes contribute to the errors, it was interpreted together with the macro-averaged and class-specific metrics.

Let $\mathbf{C} = [C_{ij}] \in \mathbb{N}^{K\times K}$ denote the multiclass confusion matrix, where $K = 5$, the row index i represents the true class, and the column index j represents the predicted class. Thus, C_{ij} is the number of samples from true class i assigned to predicted class j. The total number of evaluated samples is

$$N = \sum_{i=1}^{K}\sum_{j=1}^{K} C_{ij}.$$

For each class k, the multiclass problem can be decomposed into a one-versus-rest comparison. The corresponding confusion-matrix components are

$$\mathrm{TP}_k = C_{kk}.$$

VOLUMEN 48
XXXI Verano De la Ciencia
ISSN 2395-9797
www.jovenesenlaciencia.ugto.mx

$$\mathrm{FP}_k = \sum_{\substack{i=1 \\ i \neq k}}^{K} C_{ik}.$$

$$\mathrm{FN}_k = \sum_{\substack{j=1 \\ j \neq k}}^{K} C_{kj}.$$

$$\mathrm{TN}_k = N - \mathrm{TP}_k - \mathrm{FP}_k - \mathrm{FN}_k,$$

where TP_k denotes the number of samples from class k that are correctly classified; FP_k denotes samples from the remaining classes incorrectly assigned to class k; FN_k denotes samples from class k assigned to another class; and TN_k denotes samples that neither belong to nor are predicted as class k.

Accuracy

Accuracy quantifies the proportion of samples assigned to their correct class and is obtained from the diagonal of the confusion matrix as

$$\text{Accuracy} = \frac{\sum_{k=1}^{K} C_{kk}}{N}.$$

Values closer to 1 indicate a greater overall proportion of correct predictions. However, because every sample contributes equally to this calculation, the metric can be dominated by the most prevalent DR grades and should not be interpreted independently when the class distribution is imbalanced.

Precision

The precision of class k measures the reliability of predictions assigned to that class and is defined as

$$\text{Precision}_k = \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FP}_k} = \frac{C_{kk}}{\sum_{i=1}^{K} C_{ik}}.$$

Values closer to 1 indicate that few samples from other DR grades are incorrectly assigned to class k, whereas lower values reflect a larger proportion of false-positive predictions.

Recall

Recall, also referred to as sensitivity, measures the proportion of samples from class k that are correctly identified by the model:

$$\text{Recall}_k = \frac{\text{TP}_k}{\text{TP}_k + \text{FN}_k} = \frac{C_{kk}}{\sum_{j=1}^{K} C_{kj}}.$$

Values closer to 1 indicate that most samples belonging to class k are detected, whereas lower values indicate that a larger proportion is assigned to another severity grade.

F1-score

The F1-score is the harmonic means of precision and recall and provides a single class-specific measure that accounts for both false-positive and false-negative predictions:

$$\text{F1}_k = 2\frac{\text{Precision}_k\,\text{Recall}_k}{\text{Precision}_k + \text{Recall}_k} = \frac{2\text{TP}_k}{2\text{TP}_k + \text{FP}_k + \text{FN}_k}.$$

The score approaches 1 only when both precision and recall are high. It is therefore informative when false-positive and false-negative predictions are both relevant to the evaluation.

For the overall multiclass evaluation, class-specific precision, recall, and F1-score were summarized through macro-averaging. For a generic class-specific metric M_k, the macro-average is defined as

$$M_{\text{macro}} = \frac{1}{K}\sum_{k=1}^{K} M_k\,, \qquad M_k \in \{\text{Precision}_k, \text{Recall}_k, \text{F1}_k\}.$$

Macro-averaging assigns the same contribution to every DR grade, independently of its number of samples. Consequently, poor performance in underrepresented grades cannot be concealed by strong performance in majority classes. Macro-F1 corresponds to the arithmetic mean of the class-specific F1-scores and should not be confused with an F1-score calculated from macro-precision and macro-recall. The averaging strategy reported in the manuscript must coincide with that used in the implementation because macro-, weighted-, and micro-averaging are not interchangeable.

Mathews Correlation Coefficient

MCC quantifies the overall association between the true and predicted labels while incorporating all entries of the multiclass confusion matrix. For a K-class problem, the generalized coefficient is expressed as (Gorodkin, 2004)

$$\text{MCC} = \frac{cs - \sum_{k=1}^{K} p_k\, t_k}{\sqrt{(s^2 - \sum_{k=1}^{K} p_k^2)(s^2 - \sum_{k=1}^{K} t_k^2)}},$$

VOLUMEN 48
XXXI Verano De la Ciencia
ISSN 2395-9797
www.jovenesenlaciencia.ugto.mx

where

$$c = \sum_{k=1}^{K} C_{kk}.$$

$$s = \sum_{i=1}^{K} \sum_{j=1}^{K} C_{ij} = N.$$

$$p_k = \sum_{i=1}^{K} C_{ik}.$$

$$t_k = \sum_{j=1}^{K} C_{kj}.$$

In this formulation, c is the number of correctly classified samples, s is the total number of evaluated samples, p_k is the number of samples predicted as class k (the sum of column k), and t_k is the number of samples whose true label is class k (the sum of row k). The numerator contrasts the observed agreement with that associated with the true and predicted marginal distributions, whereas the denominator normalizes this quantity according to their variability.

MCC approaches 1 for perfect agreement, remains near 0 when the association between predicted and true labels is weak, and may become negative under systematic disagreement. Its use is particularly appropriate for imbalanced multiclass classification because it considers correct and incorrect predictions across all classes and is less influenced than accuracy by majority-class dominance. However, as a global measure, MCC should be interpreted together with macro-averaged and class-specific metrics. The multiclass formulation follows Gorodkin (2004), while its suitability for imbalanced classification has been highlighted by Boughorbel *et al*. (2017) and Chicco & Jurman (2020).

Before evaluating the proposed model, several widely used convolutional neural network architectures were trained and assessed under identical experimental conditions. This baseline comparison provides a consistent reference for DR severity classification and examines their behavior under class imbalance. *Table 3* reports accuracy, macro-precision, macro-recall, macro-F1-score, and multiclass MCC for each model.

Table 3. *Performance analysis based on diverse models.*

Model	Recall	Precision	F1-score	MCC
ResNet50 baseline	0.5696	0.7112	0.5995	0.7272
DenseNet121 baseline	0.5067	0.5944	0.5107	0.7046
ResNet18 baseline	0.5826	0.6534	0.5919	0.6986
EfficientNetB0 baseline	0.3867	0.2930	0.37	0.6274
ConvNeXtLike baseline	0.3984	0.3661	0.3645	0.5968

Figure 5 presents the confusion matrices corresponding to the baseline models. These visualizations provide valuable insights into the distribution of correct and incorrect predictions across the different severity stages of diabetic retinopathy. By highlighting specific misclassification patterns, the confusion matrices serve as an essential reference point for the subsequent comparison with the proposed model, enabling a more detailed discussion of improvements in sensitivity, balance, and clinical relevance.

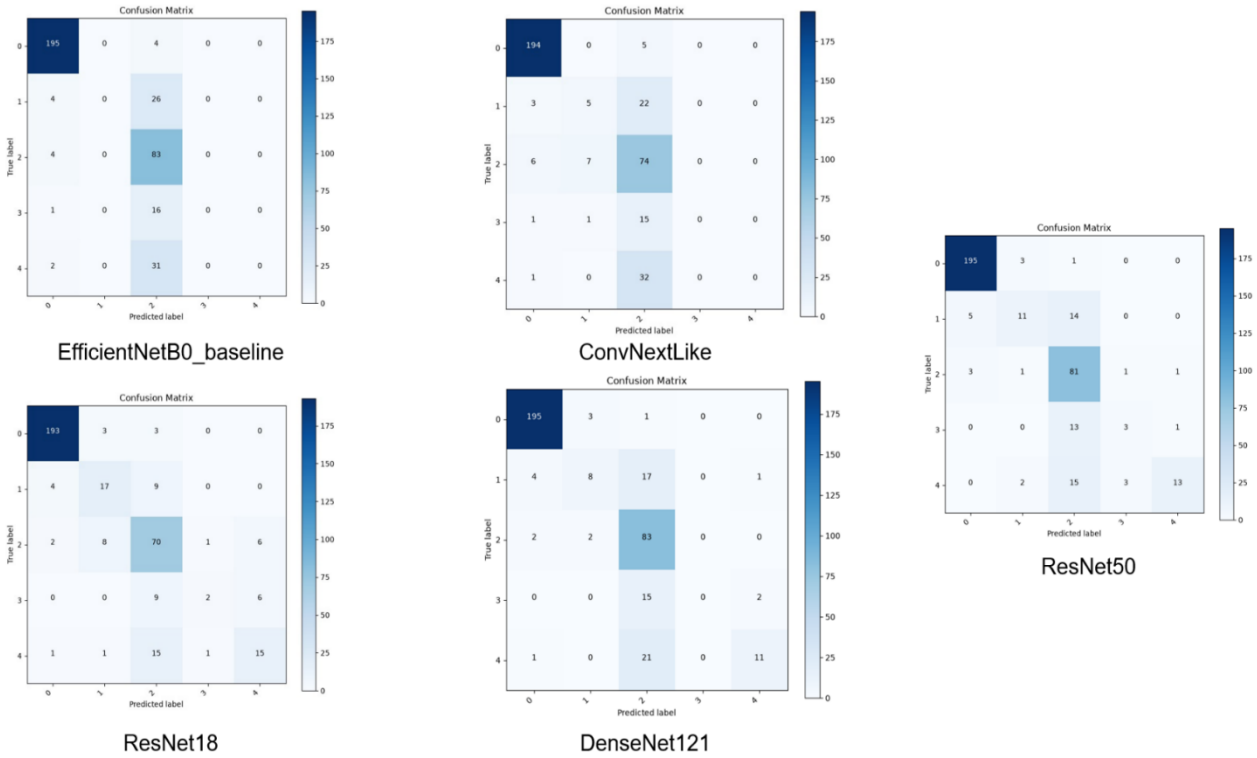


Figure 5. Confusion matrices for baseline models.

After establishing the baseline performance of ResNet50, DenseNet121, ResNet18, EfficientNetB0, and ConvNeXtLike (Table 3), the evaluation focuses on the proposed EfficientNetB0-based model. *Figure 6* presents the confusion matrix, which displays the distribution of predictions across the five diabetic retinopathy

severity grades and reveals the principal misclassification patterns. *Figure 7* depicts the corresponding normalized confusion matrix, enabling a clearer comparison of class-wise performance under the imbalanced class distribution. The quantitative assessment further includes accuracy, macro-precision, macro-recall, macro-F1-score, and MCC.

The proposed model was assessed on the APTOS 2019 dataset according to the ICDR severity scale. It achieved an accuracy of 0.79, precision of 0.81, recall of 0.79, F1-score of 0.79, and an MCC of 0.68. Although this overall MCC is slightly lower than that of the best-performing baseline, ResNet50 (0.7272), aggregate metrics alone do not capture the clinically relevant differences between models. As detailed in the per-class analysis below, the proposed model's macro-averaged scores reflect a more balanced distribution of performance across all five severity grades, rather than the concentration of correct predictions on the majority classes observed in the baseline models.

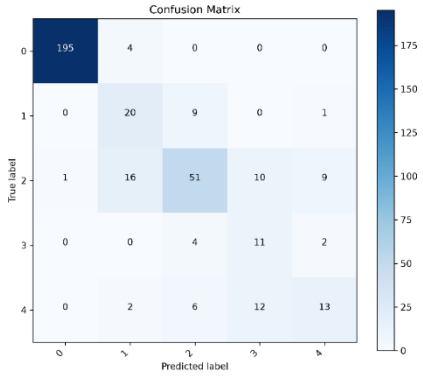


Figure 6. Confusion matrix for Custom EfficientNetB0.

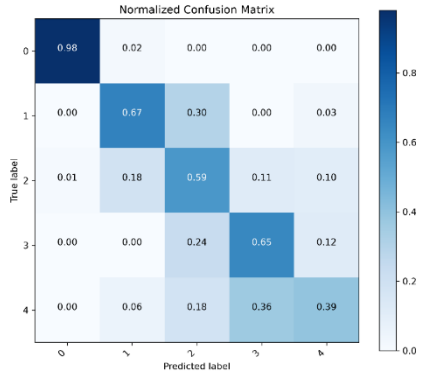


Figure 7. Normalized confusion matrix for Custom EfficientNetB0.

Discussion

While *Table 3* shows that all five baseline architectures achieved moderate-to-high overall MCC values (ranging from 0.5968 for ConvNeXtLike to 0.7272 for ResNet50), these aggregate metrics become critical failure pattern that only becomes apparent through per-class analysis: the baseline models were not able to reliably identified the more advanced, vision-threatening stages of diabetic retinopathy. The EfficientNetB0 baseline, despite its architectural efficiency, was unable to correctly classify a single image belonging to classes 1 (Mild), 3 (Severe), or 4 (Proliferative), collapsing its predictions on the majority classes (No DR and Moderate) and making it clinically unusable for the task it is intended to support. The ConvNeXtLike baseline showed a similar, failing entirely on classes 3 and 4 (the two stages associated with the highest risk of

irreversible vision loss). DenseNet121 improved marginally over these two models but still failed to correctly identify any Proliferative DR case (class 4), the single most sight-threatening grade in the diagnostic spectrum. ResNet18 and ResNet50 achieved the strongest overall baseline performance and showed some capacity to detect the more advanced classes; however, both remained highly inefficient at classifying class 3 (Severe DR), the transitional stage between non-proliferative and proliferative disease that is clinically urgent for timely intervention.

In contrast, the proposed model improved recognition of these underrepresented, severe classes. The proposed model correctly identified 65% of Severe (class 3) cases, compared to near-zero for this class across the baseline models. This improvement did not come at the cost of the model's ability to distinguish between milder stages, indicating that the class-balanced re-weighting scheme successfully redirected the model's learning capacity toward the harder, minority-class examples without simply trading off overall performance for minority-class sensitivity.

Despite this improvement, the proposed model still exhibits reduced reliability when classifying between classes 2 to 4 (Moderate, Severe, and Proliferative). In these stages, the visual differences become more delicate, which increases the chance of confusion between adjacent grades. This indicates that using class-balanced focal loss helps prevent the model from ignoring minority classes, but it is not enough to fully separate advanced stages. To address this limitation, further improvements may be needed, such as preprocessing methods that enhance specific lesions (e.g., vessels or exudates) instead of global contrast adjustments, or architectural modules that strengthen multi-scale feature extraction, such as Atrous Spatial Pyramid Pooling (ASPP). Furthermore, the system requires validation and testing on larger and more diverse datasets to better assess its generalizability and robustness across different acquisition conditions and patient populations.

Conclusions

This work proposed a modified EfficientNetB0 model for multi-class diabetic retinopathy severity grading in fundus images, trained with a class-balanced focal loss to address severe class imbalance. Compared to five baseline architectures, the proposed model showed a marked improvement in detecting Severe and Proliferative DR, classes that EfficientNetB0, ConvNeXtLike, and DenseNet121 baselines almost entirely failed to identify, and that ResNet18/ResNet50 classified inefficiently. This highlights the importance of per-class evaluation over aggregate metrics in clinical screening tasks.

References

Aiche, I., Brik, Y., Attallah, B., Bouguerra, O., Rabehi, A., Habib, M., . . . & El-kenawy, E.-S. (2026). RDE-DR: robust deep ensamble CNNs for automated diabetic retinopathy detection from fundus images. *Scientific Reports, 16*, 15226. https://doi.org/10.1038/s41598-026-48669-y

Ainapur, S., & Virupakshappa. (2025). Automated diabetic retinopathy detection using a multi-step framework with stacked ensemble-based classification model. *Expert Systems with Applications, 284*, 127709. https://doi.org/10.1016/j.eswa.2025.127709

Akhtar, S., Aftab, S., Ahmad , M., Khan, M., Abbas, S., & Ghazal, T. (2025). A deep learning based model for diabetic retinopathy grading. *Scientific Reports, 15*, 3763. https://doi.org/10.1038/s41598-025-87171-9

Awasthi, V., Awasthi, N., Kumar, H., Singh, S., Singh, P., Dixit, P., & Agarwal, R. (2024). ViT-HHO: Optimized vision transformer for diabetic retinopathy detection using Harris Hawk optimization. *MethodsX, 13*, 103018. https://doi.org/10.1016/j.mex.2024.103018

Badar, D., Abbas, J., Alsini, R., Abbas, T., ChengLiang, W., & Daud, A. (2025). Transformer attention fusion for fine grained medical image classification. *Scientific Reports, 15*, 20655. https://doi.org/10.1038/s41598-025-07561-x

Boughorbel, S., Jarray, F., & El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLos ONE, 12*(6). https://doi.org/10.1371/journal.pone.0177678

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics, 21*, 6. https://doi.org/10.1186/s12864-019-6413-7

Cui, Y., Jia, M., Song, Y., & Belongie, S. (2019). Class-Balanced Loss Based on Effective Number of Samples. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9260-9269. https://doi.org/10.1109/CVPR.2019.00949

Das, S., & Majumder, S. (2026). Detection of diabetic retinopathy 5-stages by denseNet, a CNN model. *Franklin Open*, 15, 100577. https://doi.org/10.1016/j.fraope.2026.100577

Datta, H., & Bhandari, A. (2026). A multimodal hybrid retinopathy classification model for diagnosis and grading of diabetic disease from fundus images. *Intelligence-Based Medicine, 14*, 100379. https://doi.org/10.1016/j.ibmed.2026.100379

Dejene, F. M., Ayano, Y., Feyisa, D., Debelee, T., Mekonnen, H., Gessesse, G., . . . & Mulie, D. (2026). Knowledge distillation-based lightweight MobileNet model for diabetic retinopathy classification. *Scientific Reports, 16*, 1181. https://doi.org/10.1038/s41598-025-30893-7

Dib, O. (2025). A decentralized privacy-preserving framework for diabetic retinopathy detection using federated learning and blockchain. *Results in Engineering, 26*, 105456. https://doi.org/10.1016/j.rineng.2025.105456

Flaxel, C., Adelman, R., Bailey, S., Fawzi, A., Lim, J., Vemulakonda, G., & Ying, G.-s. (2020). Diabetic Retinopathy Preferred Practice Pattern. *Ophthalmology, 127*(9), 1279. https://doi.org/10.1016/j.ophtha.2020.06.047

Gorodkin, J. (2004). Comparing two K-category assignments by a K-category correlation coefficient. *Computational Biology and Chemistry, 28*, 367-374. https://doi.org/10.1016/j.compbiolchem.2004.09.006

Gulshan, V., Peng, L., Coram, M., Stumpe, M., Wu, D., Cuadros, J., . . . & Webster, D. (2016). Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA, 316*(22), 2402-2410. https://doi.org/10.1001/jama.2016.17216

Kobat, S., Baygin, N., Yusufoglu, E., Baygin, M., Barua, P., Dogan, S., . . . & Acharya, U. (2022). Automated Diabetic Retinopathy Detection Using Horizontal and Vertical Patch Division-Based Pre-Trained DenseNET with Digital Fundus Images. *Diagnostics, 12*, 1975. https://doi.org/10.3390/diagnostics12081975

Mohanty, C., Mahapatra, S., Acharya, B., Kokkoras, F., Gerogiannis, V., Karamitsos, I., & Kanavos, A. (2023). Using Deep Learning Architectures for Detection and Classification of Diabetic Retinopathy. *Sensors, 23*, 5726. https://doi.org/10.3390/s23125726

Sah, U., Chatterjee, J., & Sujatha, R. (2026). Multi-class eye disease classification using deep learning EfficientNetB0 fusion techniques. *Scientific Reports, 16*, 6368. https://doi.org/10.1038/s41598-026-35357-0

Shakibania, H., Raoofi, S., Pourafkham, B., Khotanlou, H., & Mansoorizadeh, M. (2024). Dual branch deep learning network for detection and stage grading of diabetic retinopathy. *Biomedical Signal Processing and Control, 93*, 106168. https://doi.org/10.1016/j.bspc.2024.106168

Silva-Tinoco, R., Cabrera-Gerardo, D., Castillo-Martínez, L., & Cuatecontzi-Xochitiotzi, T. (2024). Diabetic retinopathy prevalence in Mexico: Results from a primary public access initiative for screening in patients with type 2 diabetes. *Diabetes Epidemiology and Management, 12*, 100182. https://doi.org/10.1016/j.deman.2023.100182

Soni, T., Gupta, S., Bharany, S., Rehman, A., Ghoniem, R., & Taye, B. (2025). Retinal vessel segmentation using multi sclae feature attention with MobileNetV2 encoder. *Scientific Reports, 15*, 43369. https://doi.org/10.1038/s41598-025-28707-x

Vaughan, M., Denmead, P., Tay, N., Rajendram, R., Michaelides, M., & Patterson, E. (2025). How early can we detect diaabetic retinopathy? A narrative review of imaging tools for structural assessment of the retina. *Graefes Arch Clin Exp Ophthalmol, 263*(9), 2413-2425. https://doi.org/10.1007/s00417-025-06828-3

Verma, P., Latif, G., Kaur, J., & Butt, M. (2026). Optimized Deep Learning Framework for Diabetic Retinopathy Detection and Classification Using Fundus Imaging. *Insipire Health, 1*, 1. https://doi.org/10.65718/inspireHealth.2026.2005

Vijayan, M., & Ventakakrishnan, S. (2023). A Regression-Based Approach to Diabetic Retinopathy Diagnosis Using Efficientnet. *diagnostics, 13*(4), 774. https://doi.org/10.3390/diagnostics13040774

Yasmin, S., Hasan, M., Jahan, F., & Mridha, M. (2025). DR-RetinaNet: A Deep Learning Approach for Early Screening of Diabetic Retinopathy. *Human-Centric Intelligent Systems, 5*, 497-513. https://doi.org/10.1007/s44230-025-00114-5

Youldash, M., Rahman, A., Alsayed, M., Sebiany, A., Alyazat, J., Aljishi, N., . . . & Alaqtani, M. (2024). Early Detection and Classification of Diabetic Retinopathy: A Deep Learning Approach. *AI, 5*, 2586-2617. https://doi.org/10.3390/ai5040125

Zafar, A., Kim, K., Ali, M., Byun, J., & Kim, S.-H. (2025). A lightweight multi-deep learning framework for accurate diabetic retinopathy detection and multi-level severity identification. *Frontiers in Medicine, 12*, 15513115. https://doi.org/10.3389/fmed.2025.1551315