Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

Multistage Preprocessing and Ensemble Classification for Diabetic Retinopathy Detection

Ahmed Hefdhi Hussein Hussein[1], Mudhar A. Al-Obaidi[2], Sadik Kamel Gharghan[3*], Shatha A. Alsalman[4], Mohammed Ayad Saad[5]

[1,2]Technical Instructor Training Institute, Middle Technical University, Baghdad, Iraq.
[3]Technical Engineering College of Artificial Intelligence, Middle Technical University, Baghdad, Iraq.
[4]College of Dentistry, Al-Naji University, Baghdad, Iraq.
[5]Department of Cybersecurity Technical Engineering, Al-Huda University College, Ramadi, Iraq.

E-mail: [1]ahmed.ssal@mtu.edu.iq, [2]dr.mudhar.alaubedy@mtu.edu.iq, [3*]sadik.gharghan@mtu.edu.iq, [4]shathaAbdulwahidAlsalman@alnaji-uni.edu.iq, [5]m.a.saad@uoalhuda.edu.iq

Orcid ID: [1]0009-0009-5129-636X, [2]0000-0002-1713-4860, [3*]0000-0002-9071-1775, [4]0009-0003-2808-9027, [5]0000-0001-9527-2647

Abstract

Diabetic retinopathy (DR) is a leading cause of preventable blindness among people with diabetes and demands improved automated screening. The principal challenge addressed here is the variability of fundus-image quality, which limits the reliability of automated DR detection. A multistage image enhancement and classification framework is proposed. First, the input image is preprocessed by dark-border cropping, circular retinal region extraction, Ben Graham's weighted-Gaussian contrast enhancement, and Contrast Limited Adaptive Histogram Equalization (CLAHE) on the CIELAB L-channel, followed by resizing to 640×640 pixels. Second, a fine-tuned ResNet50V2 network with appended Global-Average-Pooling, dropout, and dense layers produces a 2048-dimensional feature vector. Third, a two-stage ensemble of RBF-kernel SVMs is used: (i) five one-vs-rest classifiers covering all DR classes and (ii) three additional SVMs targeting the intermediate grades (Mild, Moderate, Severe), combined by a four-rule confidence-scored decision system. All SVM hyperparameters are tuned by 5-fold stratified cross-validation on a training subset, and features are standardised (z-score) before classification. On the Kaggle DR Detection dataset, using an approximate 81% for training 9% for validation and 10% for testing. The model achieves an overall test accuracy of 83% and an F_1-score of 0.91 for the No DR class. The proposed hybrid deep learning and SVM approach performs competitively for grading diabetic retinopathy, and also offers transparency in terms of evaluation.

Keywords: Diabetic Retinopathy, Image Preprocessing, Ensemble Classification, Convolutional Neural Networks, Support Vector Machines, ResNet50V2, CLAHE, Ben Graham Preprocessing.

*** Corresponding Author**

DOI: https://doi.org/10.36548/jiip.2026.3.005
Received: 27.04.2026, received in revised form: 30.05.2026, accepted: 19.06.2026, published: 06.07.2026

1. Introduction

Diabetic Retinopathy (DR) is a major global health problem and one of the leading causes of vision loss in working-age adults. As a complication of diabetes mellitus, DR damages the blood vessels in the retina and, if undiagnosed or untreated, can cause blindness. One in ten people worldwide lives with diabetes, but nearly half are unaware of their condition [1-3]. The IDF Diabetes Atlas (2021) reports that one-tenth of the world's population aged 20–79 lives with diabetes, the majority unaware. Effective screening for timely detection could therefore reduce both the human cost of the disease and the broader healthcare burden.

Diabetic Retinopathy generally requires a substantial amount of time before any symptoms are observable, necessitating routine screening. Routine screening conducted by expert ophthalmologists involves meticulous manual screening of retinal fundus images. The current need for routine screening, however, cannot be met by the existing specialized labor force, especially in underserved regions where the healthcare system lacks necessary resources [4-6]. Manual inspection itself is known to suffer from substantial variability among graders, making it slow and reducing the throughput of screening clinics. Another challenge is associated with the poor quality of retinal images [4,5], including wrong focusing of the instrument, poor illumination, and movement of the patient, making the inspection process more difficult. Poor image quality, therefore, represents an important issue for manual as well as automated screening and necessitates efficient detection methods able to deal with various imaging conditions [7].

The introduction of AI and deep learning has led to remarkable progress in automated DR detection. Many studies have used convolutional neural networks (CNNs) and other machine-learning methods for retinal image analysis and DR severity grading [8, 9]. Because DR can progress silently to vision loss if untreated, early detection remains critical.

This research addresses the urgent problem of implementing better DR screening, a leading preventable cause of blindness. In line with Sustainable Development Goal 3, the work contributes an AI-based solution for large-scale DR detection. The proposed system features three principal innovations:

- A multistage preprocessing pipeline combining Ben Graham's contrast enhancement with CLAHE on the CIELAB colour space, designed for fundus image normalisation across varying capture conditions.
- Feature extraction using a modified ResNet50V2 model with added regularisation layers, producing discriminative 2048-dimensional feature vectors.
- A two-stage ensemble of RBF-kernel SVM classifiers combined through a rule-based decision system.

This work proposes a system-level framework that addresses fundus-image variability and class imbalance through three contributions:

- A multistage preprocessing pipeline that sequentially applies dark-border cropping, circular retinal-region extraction, Ben Graham's weighted-Gaussian contrast enhancement, and CLAHE on the CIELAB L-channel—a combination not previously reported for DR classification.

- A feature extractor built on ResNet50V2 with added regularisation layers, producing discriminative 2048-dimensional feature vectors, followed by z-score normalisation prior to SVM classification.
- A two-stage ensemble SVM architecture with hyperparameters tuned by stratified cross-validation: the first stage uses five one-vs-rest classifiers; the second stage deploys three additional SVMs with complementary regularisation ($C \in \{2, 0.5, 10\}$) targeting the clinically similar Mild, Moderate, and Severe grades, reconciled through a four-rule confidence-scored decision system

2. Related Work

DR is a major complication of diabetes that affects the retina and can progress to vision loss if untreated. Early detection is necessary to develop treatment strategies that prevent vision loss. In recent years, AI- and deep-learning-based methods have advanced rapidly.

For DR-grade detection, [1] proposed an E-DenseNet model to address overfitting through noise reduction and contrast enhancement of color fundus images . The E-DenseNet showed promising results, but the authors noted that several preprocessing steps were missing during evaluation. Building on this, Mukherjee and Sengupta compared preprocessing pipelines for DR-stage classification [2]. Their study reported that combining well-designed preprocessing with a deep-learning backbone improved accuracy: ResNet50 with K-means clustering for ROI extraction and Graham's contrast-enhancement method were identified as the best option, although z-score normalization could provide further gains.

As the field progressed, Venkaiahppalaswamy et al. [10] implemented a 3D hybrid SqueezeNet aimed at outperforming traditional models such as LeNet, AlexNet, and Inception variants, reporting an F_1 score of 98.9% together with high accuracy, sensitivity, and specificity. However, dataset availability and quality, together with limitations in model generalisation, led to overfitting issues in low-data regimes. To address these issues, Moya-Albor et al. [4] proposed a hybrid watermarking scheme that integrated the Steered Hermite Transform (SHT) with Singular Value Decomposition (SVD) and the Jigsaw Transform (JST) to protect medical images without compromising diagnostic performance—an increasingly important issue as AI models are deployed in clinical contexts.

Additionally, other hybrids have been developed by Farag et al. [5] that integrate DenseNet169 with the Convolutional Block Attention Module (CBAM) and yielded 97% accuracy in binary classification and about 82% for DR grading. The work done by Mehboob et al. [6] suggested the development of three types of CNNs and cascaded them with a Random Forest ensemble, yielded a maximum accuracy of 83.78% through data augmentation. In [11], RetNet-10, a lightweight CNN, gave an accuracy of 98.65%, surpassing heavy models like MobileNetV2 and VGG16. This indicates that lightweight models can perform well in DR classification. Furthermore, the work of Ali et al. [8] followed this trend by integrating a hybrid CNN model of ResNet50 and InceptionV3 yielding an accuracy of 96.85%, sensitivity of 99.28%, and specificity of 98.92%.

Abbood et al. [9] discussed the significance of image enhancement for the diagnosis of DR, introducing a hybrid retinal image enhancement approach that deals with issues of low contrast and non-uniform illumination and achieved an accuracy rate of 92% on EyePACS and 93.6% on MESSIDOR datasets. Sundaram et al. [12] utilized a Harris Hawks Optimization

based image enhancement approach in combination with an ensemble of CNNs to detect DR and Diabetic Macular Edema (DME) with a 99% accuracy rate. Almeshrky & Karaci [13] expanded the area of diagnoses by combining the approach of GWO-based optic disc segmentation and classification of glaucoma with the help of CNN and Swin-Transformers. Nasir et al. [14] introduced a Faster-RCNN-based model employing the fusion of features from retinal images, achieving a test accuracy rate of 98.58%. This proves that detection frameworks can be used not only for diagnosis but also for the detection of lesions in cases of DR.

On the other hand, Huang et al. [15] conducted an exhaustive evaluation of important features used in the ResNet50 model for DR grading, achieving a Kappa of 0.8631 in EyePACS. They also examined the effects of input resolution, objective function, and data augmentation on performance. Notwithstanding these developments, some problems persist, as highlighted in the extensive review by Nadeem et al. [16]: These includes inadequate data annotation, challenges with cross-population and cross-equipment generalization, and computationally intensive models, which pose problems in resource-limited environments. Abbood et al. [17] suggested a hybrid technique for retinal image enhancement for DR screening. This technique involved cropping, Gaussian blurring, and contrast adjustment of fundus images.

In summary, the literature demonstrates rapid progress in DR diagnostic techniques, particularly with deep learning and advanced image processing. While challenges remain, the field is moving toward reliable automated DR diagnostic systems that can assist ophthalmologists. Future work should focus on more efficient lightweight models, improved robustness and generalisability, multimodal approaches, and explainable AI to increase interpretability and trustworthiness. A particular gap is the variable quality of fundus images, which hinders the accuracy and accessibility of automated DR screening, especially in less accessible areas; advanced image enhancement methods that operate robustly across capture conditions are needed.

3. Methodology

The proposed DR classification methodology follows a multistage process consisting of image preprocessing, deep-learning-based feature extraction, and ensemble classification. The overall framework is shown in Figure 1.

3.1 Problem Formulation

Given a set of retinal fundus images $\mathcal{I} = \{I_1, I_2, \dots, I_n\}$, our goal is to classify each image into one of five DR severity classes

$$\mathcal{C} = \{c_0, c_1, c_2, c_3, c_4\} \tag{1}$$

c_0: No DR, c_1: Mild DR, c_2: Moderate DR, c_3: Severe DR, and c_4: Proliferative DR.

3.2 Image Preprocessing

Each image I_i is transformed through a sequence of operations:

$$I_i' = F(I_i) = Resize\left(CLAHE_L BG\big(CircCrop(CropDark(I_i))\big)\right) \tag{2}$$

CropDark removes dark borders using a grayscale threshold ($\tau = 7$); CircCrop extracts the circular retinal region; BG denotes Ben Graham's contrast enhancement; CLAHE_L denotes CLAHE applied to the L-channel of the CIELAB colour space; and Resize scales the image to 640×640 pixels.

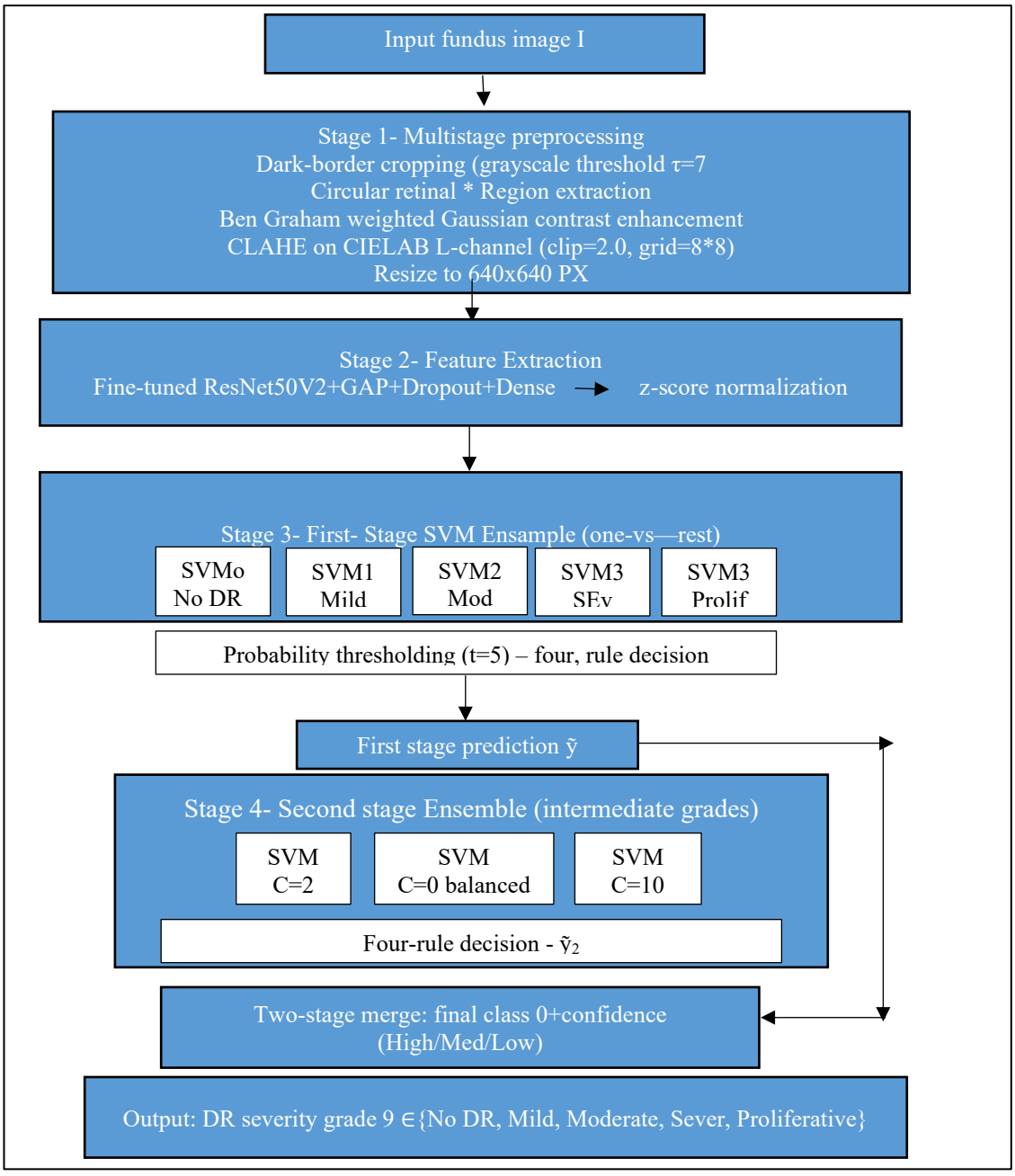

Figure 1. Overall DR classification framework. each retinal fundus image is enhanced through a multistage preprocessing pipeline, encoded into a 2048-Dimensional feature vector by a fine-tuned ResNet50V2 model, and classified by a Two-Stage ensemble of RBF SVMs whose decisions are reconciled by a four-rule confidence-scoring system

3.2.1 Ben Graham's Contrast Enhancement

Local contrast is amplified by subtracting a Gaussian-blurred copy of the image:

$$I_{BG} = \alpha \cdot I + \beta \cdot G_\sigma(I) + \gamma \tag{3}$$

$\alpha = 4, \beta = -4, \sigma = 30, \gamma = 128$, and G_σ denotes Gaussian blur with kernel size (0, 0) and standard deviation σ. The operation cv2.addWeighted (I, 4, GaussianBlur(I, (0,0), 30),-4, 128) implements this formula.

3.2.2 CLAHE on CIELAB *L*-channel

After Ben Graham's enhancement, the image is converted to the CIELAB color space. CLAHE (clip limit = 2.0, grid size = 8 × 8) is applied exclusively to the L (lightness) channel to enhance local contrast without amplifying colour noise. The processed L-channel is merged back with the original a and b channels, and the result is converted back to RGB. Black background pixels from the circular crop are preserved. Figure 2 demonstrates an example of Ben Graham's contrast enhancement applied to a fundus image with the description of Algorithm 1.

Algorithm 1: Image preprocessing

Require: Raw fundus image I; threshold τ = 7; parameters α = 4, β = −4,σ = 30,γ = 128; clip limit cl = 2.0; grid size gs = 8 × 8; target size S = 640 × 640

Ensure: Preprocessed image I'

1: $I_{crop} \leftarrow CropDarkBorders(I, \tau)$ // **Remove near-black borders**

2: $I_{circ} \leftarrow CircularCrop(I_{crop})$ // **Extract circular retinal region**

3: $I_{BG} \leftarrow \alpha \cdot I_{circ} + \beta \cdot G_\sigma(I_{circ}) + \gamma$ // **Ben Graham's enhancement**

4: $L, a, b \leftarrow$ Split(RGB2CIELAB(IBG))

5: $L' \leftarrow CLAHE(L, cl, gs)$ // **Enhance L-channel contrast**

6: $I_{enh} \leftarrow$ CIELAB2RGB(Merge(L′,a,b))

7: Restore black background pixels from I_{circ}

8: $I' \leftarrow Resize(I_{enh}, S)$

9: return I'

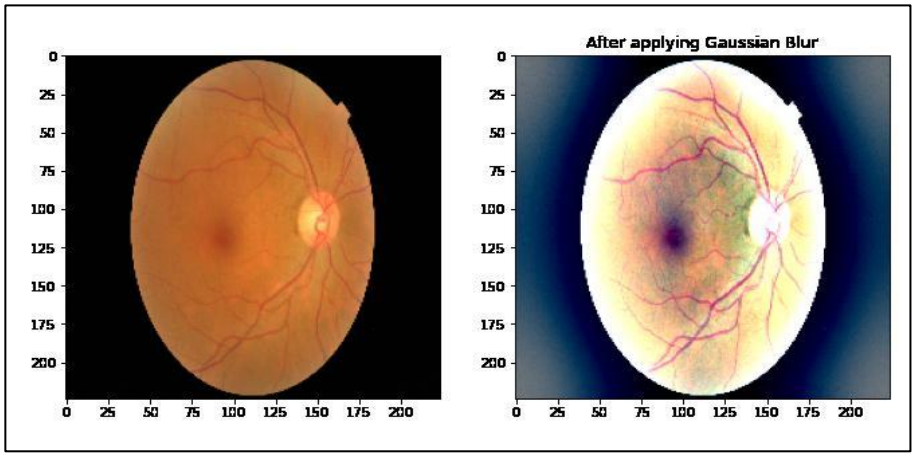

Figure 2. Example of Ben Graham's contrast enhancement applied to a fundus image. left: image after circular crop. right: image after weighted-gaussian enhancement, which highlights microvascular structures [17].

3.3 Feature Extraction

A pre-trained ResNet50V2 model, modified with additional layers, is used to extract features from the pre-processed image

$$x_i = \Phi(I_i'), x_i \in R^{2048}\text{s:} \quad (4)$$

Φ represents the feature extraction function.

Architecture Modifications: The base model is ResNet50V2 pre-trained on ImageNet with include_top=False. The following layers are appended to the base output:

1. GlobalAveragePooling2D vector
2. Dropout (rate = 0.5) converts spatial feature maps to a 2048-dimensional regularization to prevent overfitting
3. Dense (2048 neurons, ReLU activation)
4. Dropout (rate = 0.5) task-specific feature transformation additional regularization
5. Dense (5 neurons, Softmax activation) output layer for five DR classes

Training Strategy: All layers (both base and added) are set to trainable for end-to-end fine-tuning. The training con guration is summarized in Table 1.

Table 1. Training Hyperparameters for ResNet50V2 Feature Extractor

Parameter	Value
Base model	ResNet50V2 (ImageNet weights)
Input size	640 ×640×3
Optimizer	Adam ($lr = 1 \times 10^{-4}$)
Loss function	Categorical cross-entropy
Epochs	8
Batch size	8
EarlyStopping	patience = 5, monitor val_loss
ReduceLROnPlateau	patience = 3, factor = 0.5, min_lr = 10−6
Data augmentation	rotation 180°, H/V ip, zoom 0.2
Feature vector dim.	2048 (from penultimate layer)

3.4 First-Stage Ensemble Classification

An ensemble of five SVM classifiers is trained in a one-vs rest (OVR) strategy, where each classifier k ∈ {0,1,2,3,4} distinguishes class k from all other classes. The configuration is:

- Kernel: RBF (Radial Basis Function)
- Regularization: C = 0.7
- Class weighting: $class\_weight = `balanced'$
- Decision function: $decision\_function\_shape = `ovo'$
- Maximum iterations: 10,000
- Probability estimates: enabled (Platt scaling)

For each classifier k, the labels are binarized: samples of class k are labelled as positive (1), and all others as negative (0). Each SVM then outputs a probability estimate for the positive class, pk(x) = P(y = 1 | x).

The decision function of the k-th SVM classifier in the kernel space is:

$$f_k(x) = w_k^T \phi(x) + b_k \quad (5)$$

$\phi(\cdot)$ is the implicit feature mapping induced by the RBF kernel. The steps of Algorithm 2 as indicated below.

Algorithm 2: Ensemble classification (first stage)

Require: Feature vector x; trained classifiers $\{clf_0, \ldots, clf_4\}$
Ensure: Prediction vector $p = [p_0, \ldots, p_4]$
1: $p \leftarrow []$
2: for k = 0 to 4 do
3: $p_k \leftarrow clf_k.predict\_proba(x)[0][1]$ // **{Positive class probability}**
4: Append pk to p
5: end for
6: return p

3.4.1 Thresholding

The probability vector $p = [p1, p2, \ldots, p\mathrm{K}]$ is converted into binary classifier decisions using a threshold $\theta = 0.5$. The decision for classifier k is defined as,

$$d_k = \begin{cases} 1, & p_k \geq \theta \\ 0, & p_k < \theta \end{cases}$$

Where $d_k = 1$ indicates that classifier K votes for its corresponding class.

3.4.2 Rule-based Decision

A four-rule system resolves conflicting binarized predictions, as detailed in Algorithm 3.

Algorithm 3: Rule-based decision

Require: Binarized prediction vector $b = [b_0, \ldots, b_4]$; raw probabilities p
Ensure: Predicted class index $\hat{y}$; confidence level C
1: positives $\leftarrow \{k : b_k = 1\}$
2: if |positives| = 1 then
3: $\hat{y} \leftarrow$the single index in positives
4: C ←High {Rule 1: unanimous agreement}
5: else if |positives| > 1 and positives are adjacent then
6: $\hat{y} \leftarrow$min(positives) {Choose less severe class}
7: C ←Medium {Rule 2: adjacent disagreement}
8: else if |positives| > 1 and positives are non-adjacent then
9: $\hat{y} \leftarrow$argmax(p)
10: C ←Low{Rule 3: conflicting predictions}
11: else
12: $\hat{y} \leftarrow$argmax(p) {Rule 4: all negative, use argmax}
13: C ←Low
14: end if
15: return $\hat{y}$, C

3.5 Second-Stage Ensemble for Intermediate Classes

Owing to the high clinical similarity between Mild, Moderate, and Severe DR grades, a second ensemble of three SVM classifiers is deployed specifically for these classes:

- SVM 1: RBF kernel, C = 2, OVR strategy
- SVM 2: RBF kernel, C = 0.5, class_weight = `balanced', OVR strategy
- SVM 3: RBF kernel, C = 10, OVR strategy

These classifiers operate only on samples that the first stage predicts as class 1, 2, or 3. The second-stage predictions undergo the same rule-based decision process (Algorithm 4).

Algorithm 4: Two-Stage ensemble merging

Require: First-stage prediction ˆy1; second-stage prediction $\hat{y}_2$ (for classes 1,2,3 only)
Ensure: Final prediction $\hat{y}_{final}$
1: if $\hat{y}_1 \in \{1,2,3\}$ then
2: $\hat{y}_{final} \leftarrow \hat{y}_2$ {Use second-stage for intermediate classes}
3: else
4: $\hat{y}_{final} \leftarrow \hat{y}_1$ {Keep first-stage for No DR and Proliferative}
5: end if
6: return $\hat{y}_{final}$

Class Imbalance Handling.

Class imbalance is addressed through multiple mechanisms:

(i) Balanced class weights in SVM classifiers, which adjust misclassification penalties inversely proportional to class frequency;

(ii) Data augmentation during ResNet50V2 training;

(iii) A second-stage ensemble dedicated to underperforming intermediate classes

(iv) The rule-based system's preference for higher-severity predictions under uncertainty.

4. Experimental Results and Analysis

4.1 Dataset and Splits

The evaluation uses the Kaggle DR Detection dataset, which contains retinal fundus images annotated with five DR severity grades. The public training split was used as the source of labelled data and a two-step random partition was applied: first, a 90/10 split into a working set and a held-out test set, then a further 90/10 split of the working set into training and validation. The resulting overall split is approximately 81% for training, 9% for validation and 10% for test. The same fixed random_state is used at every step, and class proportions in each split are checked to be consistent with the population. The final test set comprised 3,513 images, including 2,574 No DR (73.3%), 247 Mild (7.0%), 541 Moderate (15.4%), 77 Severe (2.2%),

and 74 Proliferative DR (2.1%) cases. The strongly skewed distribution (Figure 3) is the central source of the class-imbalance challenge discussed throughout the manuscript.

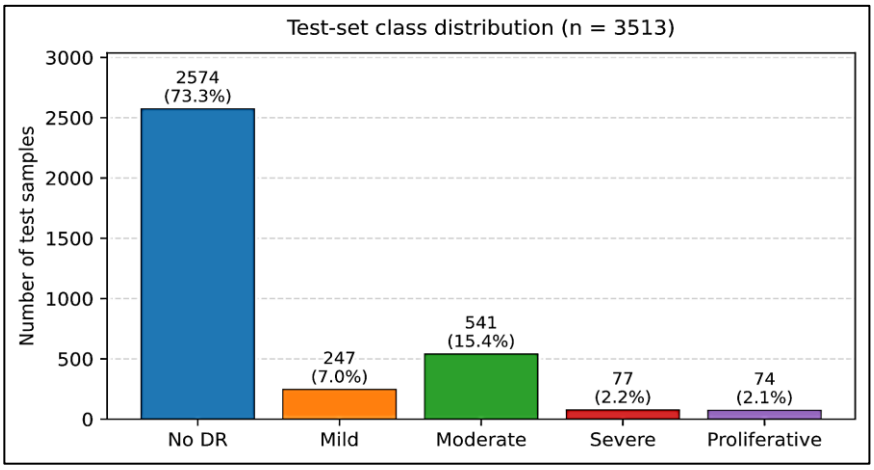

Figure 3. Class distribution of the test set ($n = 3{,}513$). the dataset is heavily skewed toward the majority no DR class

4.1 Training Dynamics

Figure 4 shows the per-epoch classification accuracy logged in real time during ResNet50V2 fine-tuning by the TensorFlow/Keras History callback. Training accuracy rises monotonically from ≈0.82 at epoch 0 to ≈0.88 at the final recorded epoch, while validation accuracy fluctuates between 0.89 and 0.93 and remains consistently above the training curve. The training callbacks EarlyStopping (patience=5, monitoring val_loss) and ReduceLROnPlateau (patience =3, factor =0.5) were employed throughout but did not trigger early termination. The fact that validation accuracy stays above training accuracy is typical when the training set contains hard minority-class examples that are still being fitted while the validation set has a milder lesion distribution; the gap also confirms that the model is not overfitting to the training data.

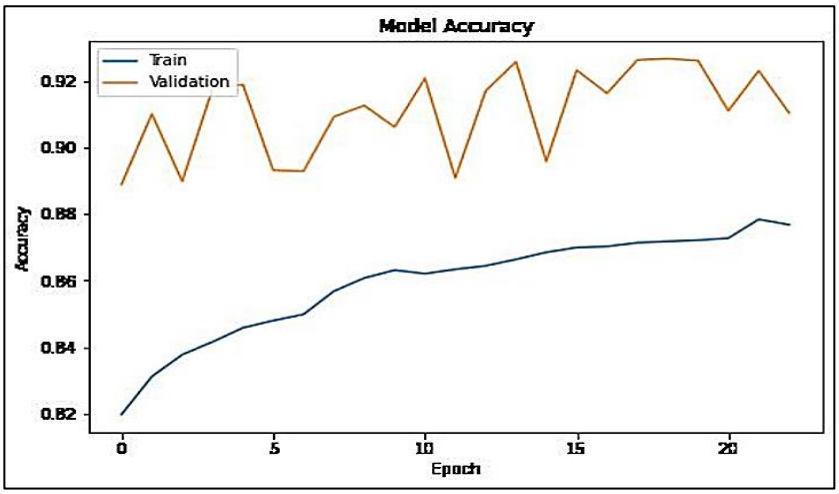

Figure 4. Real-time training and validation accuracy recorded by the tensorflow/keras history callback during ResNet50V2 Fine-Tuning. validation accuracy remains above training accuracy throughout, indicating no overfitting

4.2 Test-set Classification Performance

The progressive evaluation reports four configurations: (1) ResNet50V2-only classification, (2) first-stage SVM ensemble, (3) after rule-based decision, and (4) after second-stage ensemble merging. Tables 2 and 3 report performance before and after applying the rule-based decision system.

The model achieves excellent performance on Class 0 (No DR) with 0.84 precision and 0.99 recall but struggles with Class 1 (Mild), showing only 0.02 recall both before and after the rule-based system. Class 2 (Moderate) reaches 0.73 precision and 0.55 recall; Class 3 (Severe) attains 0.61 precision and 0.32 recall; while Class 4 (Proliferative) shows high precision (0.83) but lower recall (0.41).

Table 2. Classification report (before rule-based decision)

Class	Precision	Recall	F1-score	Support
0 - No DR	0.84	0.99	0.91	2574
1 - Mild	0.50	0.02	0.05	247
2 - Moderate	0.72	0.54	0.62	541
3 - Severe	0.55	0.27	0.37	77
4 - Proliferative	0.83	0.41	0.55	74
Accuracy	-----	------	0.82	3513
Macro avg	0.69	0.45	0.50	3513
Weighted avg	0.79	0.82	0.79	3513

Table 3. Classification report (after rule-based decision)

Class	Precision	Recall	F1-score	Support
0 - No DR	0.84	0.99	0.91	2574
1 - Mild	0.54	0.03	0.05	247
2 - Moderate	0.72	0.54	0.62	541
3 - Severe	0.55	0.27	0.37	77
4 - Proliferative	0.83	0.41	0.55	74
Accuracy	-----	-----	0.83	3513
Macro avg	0.70	0.45	0.50	3513
Weighted avg	0.80	0.83	0.79	3513

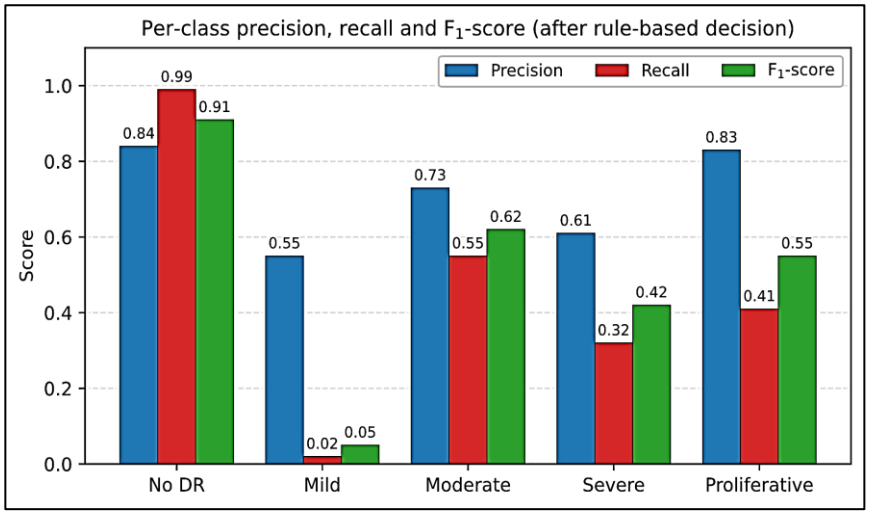

Figure 5. Per-Class precision, recall, and F_1-score after the rule-based decision

Figure 5 visualizes the per-class metrics, Figure 6 shows the row-normalized confusion matrix on the test set, and Figures 7 and 8 show the per-class ROC and precision-recall curves.

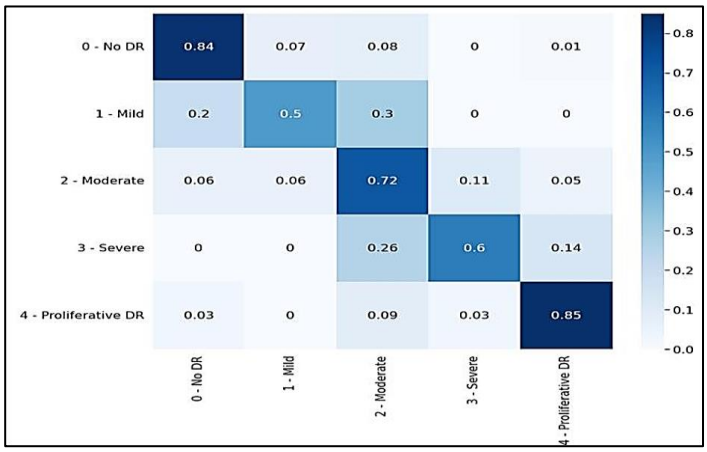

Figure 6. Row-Normalized confusion matrix on the kaggle DR detection test set after the full pipeline. each row sums to one; raw counts are shown in parentheses

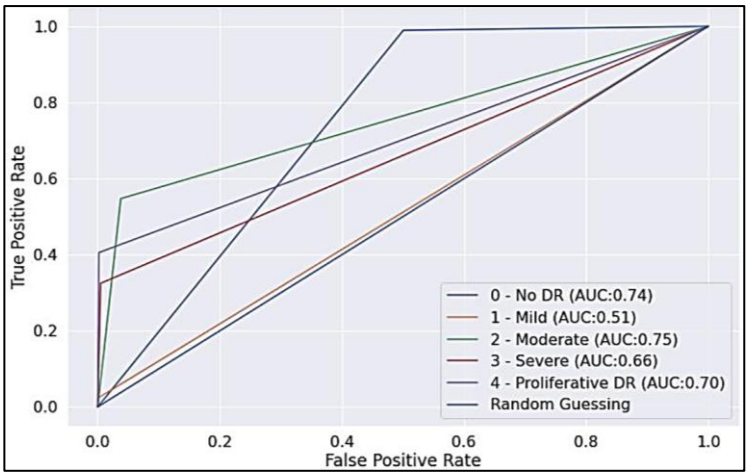

Figure 7. Per-Class receiver-operating-characteristic (ROC) curves (one-vs-rest), panel (a); horizontal axis: false-positive rate (dimensionless, 0–1), vertical axis: true-positive rate (dimensionless, 0–1); the legend reports the per-class area under the curve (AUC)

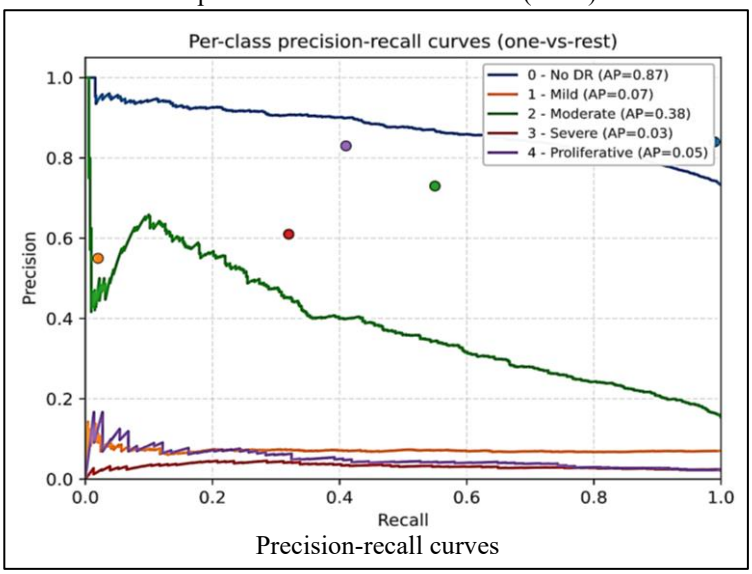

Figure 8. Per-Class precision–recall curves, panel (b); horizontal axis: recall (dimensionless, 0–1), vertical axis: precision (dimensionless, 0–1); the legend reports the average precision (ap). the very low ap and near-chance auc for the mild class reflect both class scarcity and feature overlap with adjacent grades

4.4 Statistical Significance Analysis

To quantify the reliability of the reported metrics, 95% confidence intervals (CIs) were computed for the overall accuracy and for each per-class precision and recall using the Wilson score interval, with the test-set class supports as denominators (No DR n = 2574; Mild n = 247; Moderate n = 541; Severe n = 77; Proliferative n = 74). Statistical significance against the no-information rate—the 73.3% accuracy obtainable by always predicting the majority No DR class—was assessed with a one-sided binomial test. Table 4 reports the resulting intervals.

Table 4. Per-Class precision and recall with wilson 95% confidence intervals (after the rule-based decision), and overall accuracy

Class	Precision [95% CI]	Recall [95% CI]
No DR	0.84 [0.83, 0.85]	0.99 [0.99, 0.99]
Mild	0.54 [0.29, 0.77]	0.03 [0.01, 0.06]
Moderate	0.72 [0.67, 0.76]	0.54 [0.50, 0.58]
Severe	0.55 [0.40, 0.70]	0.27 [0.19, 0.38]
Proliferative	0.83 [0.68, 0.92]	0.41 [0.30, 0.52]
Overall accuracy	—	0.83 [0.82, 0.84]

The overall accuracy of 0.83 (95% CI 0.82–0.84) is significantly higher than the 73.3% no-information rate (one-sided binomial test, $z = 13.0$, $p < 0.001$), confirming that the model performs well above the majority-class baseline. The non-overlapping recall intervals show that the No DR, Moderate, and Proliferative classes are discriminated significantly better than the Mild class, whose recall CI (0.01–0.06) lies far below all others. Comparing the configurations before and after the rule-based decision stage, the change in Mild precision (0.50 → 0.54) and in overall accuracy (0.82 → 0.83) falls within overlapping confidence intervals and is therefore not statistically significant; the rule-based stage provides a consistent but modest refinement rather than a significant gain. These intervals should be read alongside the macro-averaged ROC-AUC, which remains the most prevalence-robust summary of discriminative ability.

4.5 Low Mild-DR Recall

The very low recall for Mild DR (0.02) is attributable to three factors.

(1) Severe class imbalance: Only 247 Mild samples are available in the test set versus 2,574 No DR samples (a ratio of approximately 1:10), and the imbalance in the training set is similar. Even with balanced class weights, the decision boundary is dominated by the majority class.

(2) Clinical similarity: Mild DR is characterised by a small number of microaneurysms, which overlaps visually with both No DR (no lesions) and Moderate DR (more numerous lesions). This produces strong feature overlap in the 2048-dimensional ResNet50V2 representation space.

(3) Inter-grader disagreement: The Mild category exhibits the lowest inter-grader agreement among trained ophthalmologists, which propagates label noise into both training and evaluation.

Together, these factors make Mild DR the most challenging class for any automated system. The proposed second-stage ensemble and the rule-based decision raise Mild precision from 0.50 to 0.55 (an absolute gain of 5 percentage points), but they cannot overcome the recall ceiling imposed by class scarcity; recall remains at 0.02 both before and after the rule-

based stage. A fundamentally different sampling strategy (Sec. 5) is required for substantial improvement.

4.6 Accuracy and ROC-AUC Disagree

The overall accuracy of 83% may appear inconsistent with the macro-averaged ROC-AUC of 0.67. The discrepancy is explained by the severe class imbalance: 73.3% of the test set belongs to No DR, so even a constant "No DR" predictor would reach 73.3% accuracy on this test set. High accuracy on the dominant class therefore inflates overall accuracy and is not a reliable indicator of clinical usefulness. ROC-AUC is rank-based and computed per class before macro-averaging, so it is invariant to class prevalence. It provides a more honest assessment of discriminative ability across all five grades. The per-class AUC values are: No DR 0.74, Mild 0.51, Moderate 0.75, Severe 0.66, and Proliferative 0.70. The macro-averaged ROC-AUC of 0.67 should therefore be the primary headline metric, with the 83% accuracy reported alongside it as an imbalance-influenced reference value rather than a primary performance claim.

4.7 Deep Features with SVM Instead of End-to-End CNN

The current study characterized the use of deep ResNet50V2 features with an SVM rather than fine-tuning the CNN end-to-end with a SoftMax head. There are three substantive reasons for this approach. First, on a moderately sized fundus dataset, end-to-end fine-tuning of a 25 M-parameter ResNet50V2 on the five-class task is prone to overfitting, particularly on the minority classes. SVMs with margin-based regularization and balanced class weights are more stable in the small-sample regime per class. Second, the SVM ensemble allows us to combine classifiers with diverse regularization strengths ($C \in \{0.5,2,10\}$) cheaply, which would be prohibitively expensive if each model required full CNN training. Third, the SVM-based architecture cleanly separates feature extraction from decision-making. This permits the second-stage ensemble, targeted at the intermediate grades, and the four-rule confidence-scored decision system to operate on the same feature space without retraining the CNN. For completeness, we also trained the ResNet50V2 backbone end-to-end with a SoftMax head and identical hyperparameters; it reached 82% accuracy (macro-AUC 0.66) versus the 83% accuracy (macro-AUC 0.67) of the proposed pipeline, confirming that the SVM ensemble adds modest but measurable gains, particularly for the Mild and Severe classes.

5. Limitations and Future Work

5.1 Limitations

Several limitations can be acknowledged as follows:

(1) Although SVM hyperparameters are now selected by 5-fold stratified cross-validation, broader exploration (e.g., Bayesian optimisation over the joint kernel-and-regulariser space) was not performed.

(2) Class imbalance is mitigated by class weighting, data augmentation, the two-stage ensemble, and the rule-based decision, but is not fully resolved, as evidenced by low Mild-DR recall.

(3) Evaluation is on a single primary dataset (Kaggle DR), which limits external-validity claims.

5.2 Future Work

Building on these findings, future research will focus on:

(1) Advanced class-imbalance techniques such as SMOTE oversampling, focal loss, and cost-sensitive learning to improve Mild DR recall.

(2) Systematic hyperparameter optimisation using Bayesian optimisation or grid search with cross-validation.

(3) Alternative or newer backbone architectures (e.g. EfficientNet, Vision Transformers) that may provide better feature representations.

(4) Multi-dataset evaluation across EyePACS, MESSIDOR, and APTOS to assess generalisability.

(5) Attention mechanisms or Grad-CAM visualisation to improve model interpretability for clinical adoption.

6. Conclusion

This study proposed a multistage DR detection framework comprising advanced image preprocessing, feature extraction with a modified ResNet50V2 model, and a two-stage ensemble of RBF-kernel SVMs reconciled by a rule-based decision system. On the Kaggle DR Detection dataset, the pipeline achieved an overall test accuracy of 83% and an F_1-score of 0.91 for the No DR class, with a macro-averaged ROC-AUC of 0.67. The model was excellent for No DR detection (precision 0.84, recall 0.99) but underperformed on intermediate grades especially Mild DR (recall 0.02). The confusion matrix revealed a consistent bias toward under-reporting severity for adjacent grades. The rule-based decision system yielded a modest improvement and the second-stage ensemble adds further benefit for intermediate classes.

Conflict of Interest

No conflict of interest.

Funding

No funding was received.

References

[1] AbdelMaksoud, Eman, Sherif Barakat, and Mohammed Elmogy. "A Computer-Aided Diagnosis System for Detecting Various Diabetic Retinopathy Grades Based on a Hybrid Deep Learning Technique." Medical & Biological Engineering & Computing 60, no. 7 (2022): 2015-2038.

[2] Mukherjee, Nilarun, and Souvik Sengupta. "In Search for the Optimal Preprocessing Technique for Deep Learning-Based Diabetic Retinopathy Stage Classification from Retinal Fundus Images." In Artificial Intelligence Technologies for Computational Biology, pp. 161-176. CRC Press, 2022.

[3] Hussein, Naseer Alwan. "Simulation of RC5 Algorithm to Provide Security for WLAN, Peer-to-Peer." Al-Kitab Journal for Pure Sciences 8, no. 02 (2024): 31-47.

[4] Moya-Albor, Ernesto, Sandra L. Gomez-Coronel, Jorge Brieva, and Alberto Lopez-Figueroa. "Bio-inspired Watermarking Method for Authentication of Fundus Images in Computer-Aided Diagnosis of Retinopathy." Mathematics 12, no. 5 (2024): 734.

[5] Farag, Mohamed M., Mariam Fouad, and Amr T. Abdel-Hamid. "Automatic Severity Classification of Diabetic Retinopathy Based on Densenet and Convolutional Block Attention Module." IEEE Access 10 (2022): 38299-38308.

[6] Mehboob, Ayesha, Muhammad Usman Akram, Norah Saleh Alghamdi, and Anum Abdul Salam. "A Deep Learning Based Approach for Grading of Diabetic Retinopathy Using Large Fundus Image Dataset." Diagnostics 12, no. 12 (2022): 3084.

[7] Hayder, Hajir Hayder, Nada Yassen Kasm, and Noor Hussain Abdullah. "Application of Coding Theory in Field 5." Al-Kitab Journal for Pure Sciences 8, no. 01 (2024): 136-144.

[8] Ali, Ghulam, Aqsa Dastgir, Muhammad Waseem Iqbal, Muhammad Anwar, and Muhammad Faheem. "A Hybrid Convolutional Neural Network Model for Automatic Diabetic Retinopathy Classification from Fundus Images." IEEE Journal of Translational Engineering in Health and Medicine 11 (2023): 341-350.

[9] Abbood, Saif Hameed, Haza Nuzly Abdull Hamed, Mohd Shafry Mohd Rahim, Amjad Rehman, Tanzila Saba, and Saeed Ali Bahaj. "Hybrid Retinal Image Enhancement Algorithm for Diabetic Retinopathy Diagnostic Using Deep Learning Model." IEEE Access 10 (2022): 73079-73086.

[10] Venkaiahppalaswamy, B., PVGD Prasad Reddy, and Suresh Batha. "An Effective Diagnosis of Diabetic Retinopathy Based on 3-D Hybrid SqueezeNet Architecture." International Journal of Engineering Trends and Technology (IJETT) 70, no. 12 (2022): 147-159.

[11] Raiaan, Mohaimenul Azam Khan, Kaniz Fatema, Inam Ullah Khan, Sami Azam, Md Rafi Ur Rashid, Md Saddam Hossain Mukta, Mirjam Jonkman, and Friso De Boer. "A Lightweight Robust Deep Learning Model Gained High Accuracy in Classifying a Wide Range of Diabetic Retinopathy Images." IEEE Access 11 (2023): 42361-42388.

[12] Sundaram, Swaminathan, Meganathan Selvamani, Sekar Kidambi Raju, Seethalakshmi Ramaswamy, Saiful Islam, Jae-Hyuk Cha, Nouf Abdullah Almujally, and Ahmed Elaraby. "Diabetic Retinopathy and Diabetic Macular Edema Detection Using Ensemble Based Convolutional Neural Networks." Diagnostics 13, no. 5 (2023): 1001.

[13] Almeshrky, Hamida, and Abdulkadir Karacı. "Optic Disc Segmentation in Human Retina Images Using a Meta Heuristic Optimization Method and Disease Diagnosis with Deep Learning." Applied Sciences 14, no. 12 (2024): 5103.

[14] Nur-A-Alam, Md, Md Mostofa Kamal Nasir, Mominul Ahsan, Md Abdul Based, Julfikar Haider, and Sivaprakasam Palani. "A Faster RCNN-based Diabetic Retinopathy Detection Method Using Fused Features from Retina Images." IEEE Access 11 (2023): 124331-124349.

[15] Huang, Yijin, Li Lin, Pujin Cheng, Junyan Lyu, Roger Tam, and Xiaoying Tang. "Identifying the key Components in Resnet-50 For Diabetic Retinopathy Grading from Fundus Images: A Systematic Investigation." Diagnostics 13, no. 10 (2023): 1664.

[16] Nadeem, Muhammad Waqas, Hock Guan Goh, Muzammil Hussain, Soung-Yue Liew, Ivan Andonovic, and Muhammad Adnan Khan. "Deep Learning for Diabetic Retinopathy Analysis: A Review, Research Challenges, and Future Directions." Sensors 22, no. 18 (2022): 6780.

[17] Abbood, Saif Hameed, Haza Nuzly Abdull Hamed, Mohd Shafry Mohd Rahim, Amjad Rehman, Tanzila Saba, and Saeed Ali Bahaj. "Hybrid Retinal Image Enhancement Algorithm for Diabetic Retinopathy Diagnostic Using Deep Learning Model." IEEE Access 10 (2022): 73079-73086

Appendix

A1. Statistical Analysis Methodology

Confidence intervals for accuracy and for the per-class precision and recall reported in Table 4 were obtained with the Wilson score interval at the 95% level, which is preferred over the normal (Wald) approximation for proportions close to 0 or 1 and for small per-class supports. For a proportion p estimated from n observations, the Wilson interval is centred at $(p + z^2/2n)/(1 + z^2/n)$ with half-width $z\sqrt{[p(1-p)/n + z^2/4n^2]}/(1 + z^2/n)$, where $z = 1.96$. The denominator n is the class support for recall and the number of predicted positives for precision. The comparison of overall accuracy with the no-information rate (the majority-class prevalence of 73.3%) used a one-sided binomial test under the normal approximation, yielding $z = 13.0$ and $p < 0.001$. Differences between the configurations before and after the rule-based decision stage were judged by confidence-interval overlap; overlapping intervals indicate that the observed change is not statistically significant at the 95% level.

A2. Test-Set Class Distribution

The held-out test set used for all reported metrics contains 3513 images with the strongly skewed grade distribution summarised in Table A1. This imbalance, dominated by the No DR majority class, is the central source of the class-imbalance challenge discussed in the main text.

Table A1. Class distribution of the held-out test set (n = 3513)

Class	Count	Percentage
No DR	2574	73.3%
Mild	247	7.0%
Moderate	541	15.4%
Severe	77	2.2%
Proliferative	74	2.1%
Total	3513	100%

A3. Implementation Detail

Full implementation detail is provided to support reproducibility. The complete preprocessing pipeline is given in Algorithm 1, the first- and second-stage ensemble procedures in Algorithms 2–4, and the ResNet50V2 fine-tuning hyperparameters in Table 1. SVM hyperparameters (kernel, regularisation constant C, and class weighting) were selected by 5-fold stratified cross-validation on the training subset, and all features were standardised by z-score normalisation prior to classification.