RESEARCH PAPER

An Imbalance-Aware AI/ML Framework for Five-Stage Diabetic Retinopathy Classification Using Retinal Fundus Images

Dr. M. Srinivas[1], Dr. Harish Kuchulakanti[2], Dr. K. Thirupathanna[3]
Koutharapu Lakshmi Tejaswi[4]

[1]Assistant Professor, Department of Biomedical Engineering, Osmania University, Hyderabad, India. medisrinivas@osmania.ac.in
[2]Assistant Professor, Department of Biomedical Engineering, Osmania University, Hyderabad, India. harish.k@uceou.edu
[3]Senior Lecturer, Biomedical Engineering section, Government Institute of Electronics, Hyderabad, India. thirukurva@gmail.com
[4]Department of Biomedical Engineering, Osmania University, Hyderabad, India. ktejaswi444@gmail.com

Abstract

Diabetic Retinopathy (DR) is a progressive eye disease related to diabetes that may lead to loss of vision if not detected early. Artificial intelligence has benefits as it can independently evaluate ocular images, which makes it easier to screen large populations and reduces the workload for physicians doing this manually. However, the indicators are so subtle that it is hard for computers to categorize diabetic retinopathy accurately at various stages. Furthermore, the available datasets often lack a balanced set of cases, which makes the problem even more complex.

An accurate system for diagnosing diabetic retinopathy from retinal fundus pictures is important for diabetic retinopathy classification. To address this challenge, an artificial intelligence and machine learning–based framework was developed to automatically classify diabetic retinopathy into five clinically relevant stages. The data set contains 3662 images, each labelled with one of five diagnostic categories: no diabetic retinopathy, mild, moderate, severe or proliferative. A simple and efficient pipeline was developed to analyse the retinal fundus images for diabetic retinopathy classification. The pipeline resizes the images, enhances the contrast using a method called contrast-limited adaptive histogram equalization and describes the images using a combination of features such as Histogram of Oriented Gradients, RGB colour histograms and low-resolution pixel features. The extracted features were subsequently evaluated using four machine learning classifiers, namely Support Vector Machine (SVM), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Multilayer Perceptron (MLP), to assess their effectiveness in diabetic retinopathy classification. To mitigate the effects of class imbalance across the different diagnostic grades in the dataset, class-weighted and balanced training strategies were adopted during model training. Finally, the predictions generated by the individual classifiers were combined using a soft-voting ensemble approach, thereby leveraging the complementary strengths of the individual classifiers.

The proposed approach aims to achieve more accurate and reliable classification of diabetic retinopathy from retinal fundus images. This approach has potential to improve diagnosis and management of diabetic retinopathy and further explore its potential. The goal is to build systems that assist physicians and patients in making informed decisions and enhancing healthcare outcomes, whether for diabetic retinopathy or other diseases.

Experimental results show that the proposed ensemble achieved an accuracy of 74.18%, balanced accuracy of 46.54%, macro-F1 of 0.4748, weighted-F1 of 0.7080, Cohen's kappa of 0.5892 and ROC-AUC of 0.8458. The group performed very well especially in the no-DR and moderate-DR categories with sensitivities of 95.94% and 80.00% respectively. However, the detection of severe diabetic retinopathy was challenging, with a sensitivity of 6.90%. Class-weighted SVM improved severe-DR sensitivity from 6.90% to 27.59%, suggesting that focusing on minority-class identification may reduce overall accuracy but improve clinically significant class-sensitive metrics. The results show the potential and limitations of the lightweight AI/ML approaches for automated diabetic retinopathy grading and provide a reproducible benchmark for future deep-learning studies.

Keywords: diabetic retinopathy, retinal fundus images, artificial intelligence, machine learning, SVM, random forest, XGBoost, multilayer perceptron, ensemble learning, class imbalance, medical image classification.

How to cite this article: M. Srinivas1 Dr. Harish Kuchulakanti2 Dr. K. Thirupathanna3. An Imbalance-Aware AI/ML Framework for Five-Stage Diabetic Retinopathy Classification Using Retinal Fundus Images. Int J Drug Deliv Technol. 2026;16(87s):1705-1715. DOI: 10.25258/ijddt.16.87s.183.

1. Introduction

Diabetes is serious chronic disease and can cause many complications in the body. The biggest is diabetic retinopathy, in which the tiny blood vessels in the retina are damaged. If such damage is not caught early and treated, it may result in permanent loss of vision [1].

Diabetic retinopathy is usually classified into different degrees of severity. The five stages are: no diabetic retinopathy, mild diabetic retinopathy, moderate diabetic retinopathy, severe diabetic retinopathy and proliferative diabetic retinopathy [2]. These problems are usually detected with retinal fundus photography, a specialized type of imaging that physicians use. The method is non-invasive, i.e. without causing pain and very efficient in searching for anomalies in the retina. It is often used to diagnose diabetic retinopathy and to determine the severity of this disease. The images allow physicians to view the state of the retina and decide on the best course of treatment.

The standard assessment of the retina is based on the expertise of experienced ophthalmologists or evaluators. Although manual analysis of multiple photographs is time consuming, expert assessment remains the clinical gold standard. The availability of digital fundus photography has therefore encouraged the development of automated computer-assisted screening techniques [3].

An alternative approach is provided by machine learning, where features in the picture are converted into numerical formats and thereafter sorted into diagnostic categories. Conventional machine learning algorithms such as support vector machines, random forests, gradient-boosting approaches, and multilayer

perceptrons can have good baseline systems without the heavy computational [4].

However, the problem is that the datasets of diabetic retinopathy are a little bit hard to handle due to the frequent imbalance among the various illness categories. For example, you might have a much larger number of images of normal eyes or eyes with moderate diabetic retinopathy than eyes with severe or proliferative conditions [5]. That may be difficult, though. A computer program trained to identify diabetic retinopathy from those images may appear to do well overall but be very poor at identifying the more severe cases, which are usually the most important to catch [6].

This study addresses the challenge by developing and empirically evaluating an AI/ML framework based on image-derived features The research focuses on balanced accuracy, macro-F1, class-specific sensitivity, Cohen's kappa, and ROC-AUC rather than overall accuracy.

The primary contributions of this study are:

1. Creation of a consistent, lightweight workflow for extracting features from retinal images.
2. Comparative assessment of SVM, random forest, XGBoost, and MLP classifiers.
3. Integration of class-balancing methodologies into the machine-learning framework.
4. Creation of a soft-voting ensemble classifier.
5. Comprehensive evaluation for each class utilizing sensitivity, specificity, accuracy, and F1-score.
6. Examination of discrepancies between neighbouring DR severity classifications.
7. Empirical examination of the constraints inherent in traditional image-feature-based methodologies.

2. Aim and Objectives

2.1 Aim

This research focusses on the design and validation of a novel artificial intelligence system that can independently detect diabetic retinopathy from images of the posterior segment of the eye. The goal is to develop a system that uses machine learning to detect data abnormalities and accurately classify the disease into five stages.

2.2 Objectives

The aims are:

- To delineate the dataset of retinal fundus images.
- To prepare the retinal pictures for machine-learning evaluation.
- To derive supplementary structural and chromatic image attributes.
- To examine the impact of class imbalance on diabetic retinopathy classification.
- To evaluate several supervised machine-learning classifiers.
- To create a soft-voting ensemble.
- To assess categorization efficacy by several indicators.
- To conduct an analysis of errors by class.
- To ascertain the advantages and drawbacks of the suggested framework.
- To create a reference point for subsequent deep-learning-driven diabetic retinopathy classification.

3. Materials and Dataset

The data used in this study is from the APTOS2019 Blindness Detection dataset, which can be found on Kaggle. This collection consists of images of the back of the eye, used by doctors to evaluate diabetic retinopathy, a disease that can cause vision loss. Images were taken from Kaggle and used to develop and test a new artificial intelligence and machine learning method for the identification of this disease.

A total of 3,662 annotated retinal fundus images were available for the study [7]. The extra metadata file has two main features, i.e., an identification code that uniquely identifies each retinal image and a diagnostic that defines the diabetic retinopathy grade. The diagnostic classifications ranged from 0 to 4, the five clinically significant stages of diabetic retinopathy.

Table 1. Distribution of the APTOS 2019 retinal fundus image dataset

Class	Diagnostic Grade	Number of Images	Percentage
0	No DR	1,805	49.29%
1	Mild DR	370	10.10%
2	Moderate DR	999	27.28%
3	Severe DR	193	5.27%
4	Proliferative DR	295	8.06%
	Total	3,662	100%

Thus, the dataset can be viewed as a multiclass classification problem with five classes. There is a significant class imbalance. The No DR group is almost 50% of the entire dataset, while Severe DR is represented by only 193 photos. The largest class is about 9.35 times the size of the smallest one:1. This disparity can bias Traditional classifiers tend to prefer the majority class reducing their ability to correctly identify clinically significant minority classes, especially Severe DR and Proliferative DR. Therefore, the proposed approach included a class imbalance handling.

4. Proposed Methodology

The proposed methodology is a set of processing steps to classify diabetic retinopathy into five stages using retinal fundus images. The overall workflow consists of image preprocessing, feature extraction, dataset partitioning, class-imbalance handling, and classification using multiple machine learning algorithms. Fig. 1 shows the complete workflow of the proposed methodology.

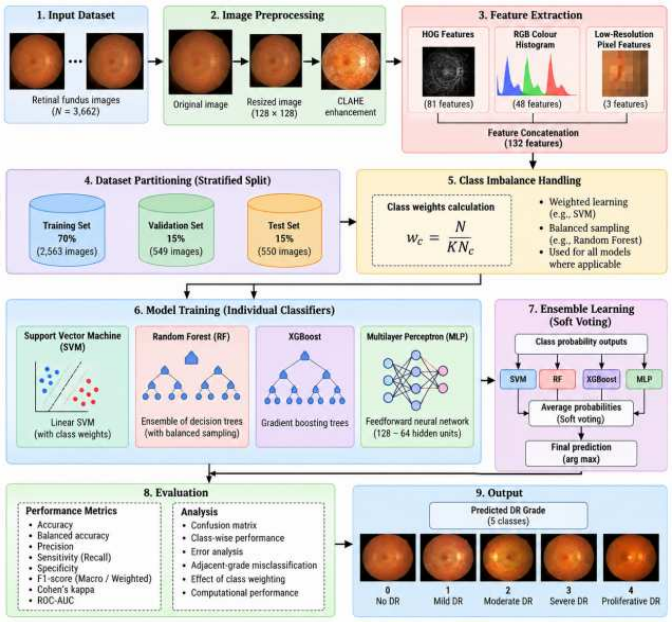


Fig. 1. Block diagram of the proposed methodology for five-stage diabetic retinopathy classification using retinal fundus images.

4.1 Image Preprocessing

The retinal images were firstly scaled to a common low resolution. Contrast was enhanced later using contrast-limited adaptive histogram equalization (CLAHE). CLAHE was chosen because retinal images often have substantial local light fluctuations. Unlike global histogram equalization, CLAHE is used for local contrast enhancement and it avoids the over-amplification of noise.

The preparatory phase comprised:

1. Loading of the image.
2. Standardization of image dimensions.
3. Transformation into a suitable color representation.
4. Luminance enhancement utilizing CLAHE.
5. Reversion to RGB format.
6. Conversion to grayscale for the extraction of features based on gradients.

4.2 Feature Extraction

Three complementary feature groups were extracted.

4.2.1 Histogram of Oriented Gradients

Histogram of Oriented Gradients captures local edge and structural information. This can represent patterns associated with retinal vascular structures and lesion boundaries.

4.2.2 RGB Colour Histogram

Colour-histogram features capture the distribution of colour intensities within the retinal image. Such information may contribute to discrimination between normal retinal appearance and pathological changes.

4.2.3 Low-Resolution Pixel Features

The resized RGB image was flattened into a numerical feature vector. Although relatively simple, these features preserve coarse spatial colour information.

The final feature representation was formed by concatenating the three feature groups:

$$F = [F_{HOG}, F_{RGB}, F_{Pixel}]$$

The resulting feature vector contained 132 features.

4.2.4 Dataset Partitioning

A stratified partitioning strategy was used to maintain relative representation of the five DR classes across the dataset subsets, as shown in Table 2. The dataset was divided into training, validation and testing sets in the ratio of 70%:15%:15%, i.e. 2,563, 549 and 550 images respectively. Model training was performed on the training set, and model selection was performed on the validation set. The test set was entirely independent and was not used for training the model or selecting parameters, providing an unbiased final evaluation.

Table 2. Stratified Dataset Split for Model Development and Evaluation

Dataset Subset	**Percentage**	**Number of Images**	**Purpose**
Training	70%	2,563	Model training
Validation	15%	549	Model selection
Testing	15%	550	Final evaluation
Total	**100%**	**3,662**	

The test set was not used for model training or parameter selection.

4.2.5 Class-Imbalance Handling

The unequal distribution of DR grades represents an important methodological issue.

Class weights were calculated according to:

$$w_c = \frac{N}{KN_c}$$

where $N$is the number of training samples, $K$is the number of classes and $N_c$is the number of training samples belonging to class c.

Weighted learning was incorporated into the SVM and balanced sampling was used for the random forest model.

This strategy was intended to prevent the classifier from optimizing predominantly for the majority no-DR class.

5. Machine-Learning Models

Four classifiers were evaluated in the work.

5.1 Support Vector Machine

A linear SVM was used to identify a separating hyperplane in the extracted feature space. Class weighting was incorporated to increase the contribution of minority classes.

5.2 Random Forest

Random forest combines multiple decision trees and aggregates their predictions. Balanced sampling was used to reduce the influence of the majority class.

5.3 XGBoost

Extreme gradient boosting was employed as a tree-based boosting approach. Sequential trees attempt to correct errors made by earlier trees.

5.4 Multilayer Perceptron

An MLP was used to investigate a nonlinear neural-network classifier operating on the extracted image features.

6. Proposed Ensemble

The four classifiers generated class-probability estimates.

The ensemble probability was calculated as:

$$P_{ensemble}(c) = \frac{1}{M}\sum_{m=1}^{M} P_m(c)$$

where $M = 4$.

The final predicted class was:

$$\hat{y} = \arg\max_{c} P_{ensemble}(c)$$

This approach allows different classifiers to contribute complementary information.

7. Evaluation Metrics

The models were evaluated using:

Accuracy

$$Accuracy = \frac{\text{Correct predictions}}{\text{Total predictions}}$$

Precision

$$Precision = \frac{TP}{TP + FP}$$

Sensitivity

$$Sensitivity = \frac{TP}{TP + FN}$$

Specificity

$$Specificity = \frac{TN}{TN + FP}$$

F1-score

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Macro-F1 was used to assign equal importance to all five diagnostic classes.

Cohen's Kappa

Cohen's kappa was used to measure agreement between predictions and ground-truth labels beyond chance agreement.

ROC-AUC

Multiclass ROC-AUC was calculated using a one-vs-rest strategy.

8. Overall Model Comparison

The overall results of the evaluated models on the independent test set are illustrated in Table 3. The table gives a quantitative comparison of the four individual classifiers and the proposed soft-voting ensemble. As the dataset is class imbalanced, a single performance metric may not be able to provide the overall assessment of the classification performance. Therefore, accuracy is reported in conjunction with balanced accuracy, macro-precision, macro-recall, macro-F1 score, weighted-F1 score, Cohen's kappa and multiclass ROC-AUC. The proposed soft-voting ensemble has obtained the highest values for all the aggregate performance indicators presented in

Table 3. Overall efficiency in comparison to the individual classifiers.

Model	**Accuracy (%)**	**Balanced Accuracy (%)**	**Macro Precision**	**Macro Recall**	**Macro-F1**	**Weighted-F1**	**Kappa**	**ROC-AUC**
SVM	65.27	45.52	0.4432	0.4552	0.4443	0.6546	0.4808	0.8073
Random Forest	72.91	43.31	0.5920	0.4331	0.4361	0.6815	0.5617	0.8420
XG Boost	72.36	42.91	0.5347	0.4291	0.4252	0.6777	0.5539	0.8417
MLP	69.27	43.68	0.4454	0.4368	0.4369	0.6755	0.5243	0.8238
Proposed Ensemble	**74.18**	**46.54**	**0.5739**	**0.4654**	**0.4748**	**0.7080**	**0.5892**	**0.8458**

The precision was 74.18%, which was 1.27 percentage points higher than Random Forest and 8.91 percentage points higher than SVM. Importantly, the ensemble achieved the highest balanced accuracy (46.54%) and macro-F1 (0.4748) which indicates that the improvement was not only driven by the dominant No-DR class. However, the difference between 74.18% accuracy and 46.54% balanced accuracy indicates that the performance of the illness classes varies considerably.

8.1 Class-Wise Results

The overall accuracy of the proposed ensemble was highest, 74.18%. It achieved the best-balanced accuracy, macro-F1, weighted-F1, Cohen's kappa and ROC-AUC among the evaluated methods.

Results demonstrate that the combination of multiple classifiers improves the overall robustness of the feature-based system.

The class-wise performance of the four individual classifiers is shown in Tables 4-7. Results are also reported in terms of precision, sensitivity, specificity, F1-score and support for each of the five diabetic retinopathy grades.

8.1.1 SVM

The class-wise results using SVM classifier are given in Table 4. The best performance for the class No DR was obtained for the SVM with a sensitivity of 0.9077 and F1-score of 0.8962. Sensitivity for Moderate class was 0.4867 and F1-score was 0.5428. The classes of Mild, Severe and Proliferative had relatively lower F1-scores of 0.3876, 0.2222 and 0.1728 respectively. But still, the specificity was quite high for all five classes.

Table 4. Class-wise performance of the SVM classifier

DR Grade	Precision	Sensitivity	Specificity	F1-Score	Support
No DR	0.8849	0.9077	0.8682	0.8962	271
Mild	0.3425	0.4464	0.8990	0.3876	56
Moderate	0.6134	0.4867	0.8360	0.5428	150
Severe	0.1860	0.2759	0.9473	0.2222	29
Proliferative	0.1892	0.1591	0.9432	0.1728	44

8.1.2 Random Forest

Table 5 presents the performance of Random Forest classifier in a class-wise manner. The classifier had the best performance for the class No DR with sensitivity of 0.9594 and F1-score of 0.9075. The Moderate class also performed relatively high in terms of the sensitivity (0.8200) and F1-score (0.6777). However, the Severe and Proliferative classes had low sensitivities of 0.1034 and 0.0682 respectively, leading to lower F1-scores.

Table 5. Class-wise performance of the Random Forest classifier

DR Grade	Precision	Sensitivity	Specificity	F1-Score	Support
No DR	0.8609	0.9594	0.8411	0.9075	271
Mild	0.5217	0.2143	0.9456	0.3038	56
Moderate	0.5775	0.8200	0.8125	0.6777	150
Severe	0.5000	0.1034	0.9896	0.1714	29
Proliferative	0.5000	0.0682	0.9891	0.1200	44

8.1.3 XGBoost

The class-wise results of XGBoost are presented in Table 6. The sensitivity and F1-score of XGBoost for No DR class was 0.9594 and 0.9012, respectively. The Moderate class also performed fairly well with a sensitivity of 0.7800 and an F1-score of 0.6705. However, the Severe and Proliferative classes remained hard with sensitivities of only 0.0345 and 0.0682.

Table 6. Class-wise performance of the XGBoost classifier

DR Grade	Precision	Sensitivity	Specificity	F1-Score	Support
No DR	0.8497	0.9594	0.8118	0.9012	271
Mild	0.4857	0.3036	0.9416	0.3736	56
Moderate	0.5879	0.7800	0.8275	0.6705	150
Severe	0.2500	0.0345	0.9876	0.0606	29
Proliferative	0.5000	0.0682	0.9901	0.1200	44

8.1.4 MLP

Table 7 shows the class-wise results obtained using the MLP classifier. The MLP performed best for the No DR class with sensitivity of 0.9188 and F1-score of 0.9005. The Moderate class was also relatively well predicted with a sensitivity of 0.6867 and F1-score of 0.6338. However, the Severe class had a very low sensitivity (0.0345) and F1-score (0.0488), indicating the difficulty of detecting this minority class.

Table 7. Class-wise performance of the MLP classifier

DR Grade	Precision	Sensitivity	Specificity	F1-Score	Support
No DR	0.8830	0.9188	0.8596	0.9005	271
Mild	0.4222	0.3393	0.9316	0.3762	56
Moderate	0.5886	0.6867	0.8288	0.6338	150
Severe	0.0833	0.0345	0.9539	0.0488	29
Proliferative	0.2500	0.2045	0.9655	0.2250	44

8.1.5 Proposed Ensemble Performance

The best performance was achieved by the proposed ensemble for the No DR class with sensitivity 0.9594 and F1-score 0.9220. The Moderate class also showed a fairly good performance with a sensitivity of 0.8000 and F1-score of 0.6838. For the Mild class, sensitivity was 0.3393 and F1-score was 0.4086.

The Severe and Proliferative classes were still more difficult to detect, with sensitivities of 0.0690 and 0.1591 respectively. However, high values of specificity of 0.9795 and 0.9853 were obtained for both classes, indicating that samples from the other DR grades were generally correctly distinguished from these classes. Overall, the class-wise results show that the ensemble performs well for most and Moderate classes, but the detection of the less represented Severe and Proliferative grades is still a challenge.

Table 8. Class-wise performance of the proposed soft-voting ensemble

DR Grade	Precision	Sensitivity	Specificity	F1-Score	Support
No DR	0.8874	**0.9594**	0.8825	**0.9220**	271
Mild	0.5135	0.3393	0.9363	0.4086	56
Moderate	0.5970	**0.8000**	0.8300	**0.6838**	150
Severe	0.3333	0.0690	0.9795	0.1143	29
Proliferative	0.5385	0.1591	0.9853	0.2456	44

The class-wise performance of the proposed soft-voting ensemble in terms of precision, sensitivity, and F1-score for all five diabetic retinopathy grades is further shown in Figure 2. The corresponding aggregate performance comparison using macro-F1 and weighted-F1 scores across the evaluated models is shown in Figure 3. These plots help to intuitively compare model performance on each individual disease grade and the overall classification capability of the model. As shown in Figure 2, the proposed ensemble performed particularly well for No DR and Moderate DR, achieving sensitivities of 95.94% and 80.00%, respectively. The sensitivity decreased to 33.93% for Mild DR, 6.90% for Severe DR, and 15.91% for Proliferative DR. These results indicate that the primary limitation of the proposed system is not the discrimination of normal and moderate disease cases, but the reliable detection of the more advanced and minority DR classes.

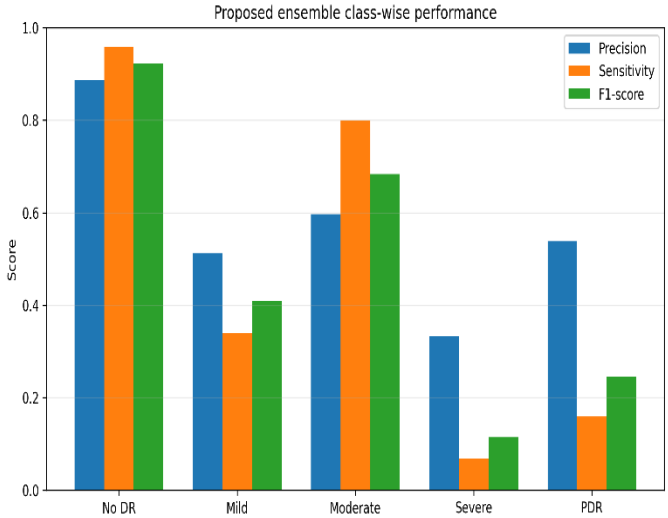


Figure 2. Class-wise precision, sensitivity, and F1-score of the proposed ensemble.

Figure 3 also presents the difference in the overall classification performance of the tested models. The proposed soft-voting ensemble obtained the best macro-F1 and weighted-F1 scores. It outperformed the individual classifiers by an overall better classification performance. However, the lower sensitivity for Severe DR is a limitation. This result shows the difficulty of achieving reliable 5-stage diabetic retinopathy classification with a relatively small imbalanced dataset.

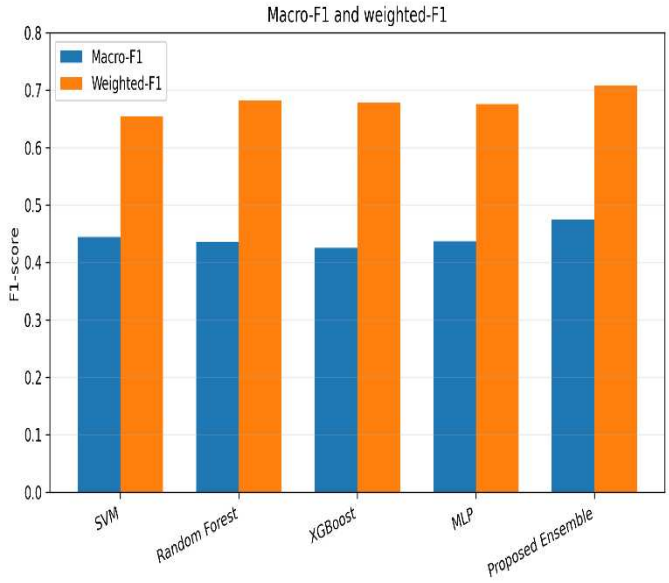


Figure 3. Macro-F1 and weighted-F1 across the evaluated models.

8.1.6 Confusion-Matrix Analysis

It can be observed from Figure 4 that 260 out of 271 No DR images were correctly classified and 120 out of 150 Moderate DR images were correctly classified. The confusion matrix also indicates a clear trend of misclassification in the later stages of disease grades. In particular, 20 of 29 images of Severe DR were misclassified as Moderate DR and also 27 of 44 images of Proliferative DR were misclassified as Moderate DR. These results show a tendency of the proposed ensemble to underestimate the severity of advanced diabetic retinopathy, which helps to explain the lower sensitivity for the Severe and Proliferative DR classes.

The misclassification of advanced grades as Moderate DR is an important limitation of the proposed framework. The error pattern indicates that it is still difficult to differentiate the higher severity levels even with class imbalance. However, the relatively high number of correctly classified No DR and Moderate DR images reveals the effectiveness of the proposed ensemble in recognizing these classes.

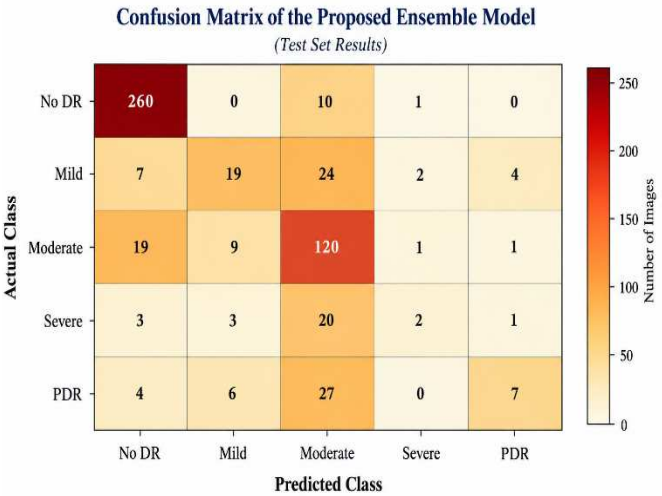


Figure 4. Confusion matrix for the proposed ensemble on the 550-image test set.

9. Class-Imbalance Experiment

The class-imbalance management was evaluated by the SVM and Random Forest classifiers. To examine the effects of class weighting and balanced sampling on the detection of the minority-class without compromising the overall classification performance, an additional comparative experiment was performed. The results are shown in Table 9.

Table 9. Effect of class-imbalance handling on model performance

Configuration	Accuracy (%)	Balanced Accuracy (%)	Macro-F1	Severe Recall	PDR Recall
SVM – Standard	70.91	42.57	0.4142	6.90 %	0.00 %
SVM – Weighted	**65.27**	**45.52**	**0.4443**	**27.59 %**	**15.91 %**
RF – Standard	72.55	42.70	0.4290	10.34 %	6.82 %
RF – Balanced	**72.91**	**43.31**	**0.4361**	10.34 %	6.82 %

The accuracy decreased from 70.91% to 65.27% after using class weighting; however, the balanced accuracy increased from 42.57% to 45.52%, and Macro-F1 score increased from 0.4142 to 0.4443. More importantly, the recall of Severe DR was significantly improved from 6.90% to 27.59% and PDR recall increased from 0.00% to 15.91%.

In contrast, the balanced Random Forest resulted in minor changes in aggregate performance. The accuracy improved from 72.55% to 72.91% and balanced accuracy from 42.70% to 43.31% and Macro-F1 from 0.4290 to 0.4361. However, the recall of both Severe DR and PDR did not change at 10.34% and 6.82% respectively.

Overall, these results indicate that accuracy alone may be insufficient to represent the performance of a classifier on an imbalanced medical dataset. The class weighting reduced the overall accuracy of SVM but greatly enhanced the detection of minority and advanced DR classes . Therefore, balanced accuracy, Macro-F1 and class-specific recall should be considered for model evaluation of five-stage diabetic retinopathy classification.

9.1 Validation Performance

The validation results were used for model comparison and selection and the independent test set was reserved for the final evaluation. This separation avoids information leakage from the final evaluation set into the model selection process, and provides a more reliable estimate of the generalization performance. The validation performance of the evaluated classifiers is shown in Table 10.

Table 10. Validation performance of the evaluated classifiers

Model	Validation Accuracy (%)	Validation Balanced Accuracy (%)	Validation Macro-F1
SVM	67.94	50.67	**0.4918**
Random Forest	71.58	42.39	0.4255
XGBoost	71.22	41.94	0.4151
MLP	69.22	44.96	0.4530

Random Forest achieved the best validation accuracy of 71.58% and SVM achieved the best validation balanced accuracy of 50.67% and the best validation Macro-F1 score of 0.4918. MLP achieved 0.4530 of validation Macro-F1 score and XGBoost got 0.4151

Results from the validation differ from the final test-set results, with the proposed soft-voting ensemble providing the best overall aggregate performance. This difference in model ranking emphasizes the importance of an independent test set for final evaluation. Therefore, the conclusions on generalization performance of the models were drawn from the independent test-set results, not only from validation performance.

9.2 Computational Performance

Computational performance also assessed the efficiency of the proposed feature-based framework. The framework is intended to offer a lightweight alternative to computationally intensive deep-learning approaches, and the training time, feature dimensionality, and model configurations were studied. Table 11 shows the computational characteristics of the evaluated models.

Table 11. Computational performance of the evaluated models

Model	Feature Dimension	Configuration	Approximate Training Time
SVM	132	Linear, class weighted	0.67 s
Random Forest	132	Balanced, 100 trees	0.64 s
XGBoost	132	150 trees	3.32 s
MLP	132	128–64 hidden units	1.44 s
Proposed Ensemble	132	Four-model probability voting	Combined

The training time for the Random Forest and SVM classifiers was less than one second, with approximate training times of 0.64 s and 0.67 s, respectively. XGBoost took roughly 3.32 s, and MLP took roughly 1.44 s. The feature representation was 132-dimension for all four classifiers, indicating that the framework keeps a rather compact feature space.

The documented MLP architecture contained about 25,605 trainable parameters, while the linear SVM used a much simpler

coefficient-based model. The obtained results show that the computational requirements of the proposed feature-oriented framework are relatively low. However, the training times reported here should not be directly compared with GPU-accelerated deep learning systems, since differences in hardware, implementation and computational environments can substantially affect the training time.

The computational results, in general, confirm the proposed feature-based framework as a lightweight alternative for diabetic retinopathy classification, especially in scenarios with limited computational resources.

9.3 Error Analysis

The error distribution of the proposed soft-voting ensemble was analyzed on the 550-image test set to identify the major sources of misclassification. The proposed ensemble correctly classified 408 images (74.18%) and 142 images (25.82%) were incorrectly classified. The major error categories are summarized in Table 12.

Table 12. Error analysis of the proposed soft-voting ensemble

Error Category	**Number**	**Percentage**
Correct classification	408	74.18%
Incorrect classification	142	25.82%
No DR → DR	11	2.00%
DR → No DR	33	6.00%
Severe → Moderate	20	3.64%
PDR → Moderate	27	4.91%

The main errors were to confuse the Moderate, Severe and Proliferative DR categories. Notably, 20 Severe DR images were misclassified as Moderate DR, and 27 PDR images were misclassified as Moderate DR. Such mistakes reflect a tendency to underestimate the severity of the disease in advanced DR. Additionally, 33 DR images were misclassified as No DR and 11 No DR images were classified as DR. Overall, the error analysis suggests that the differentiation of advanced and minority DR grades still remains one of the main challenges of the proposed framework.

9.4 Adjacent-Grade Error Analysis

Further understanding of the misclassification was conducted by analyzing the adjacent-grade errors for the proposed soft-voting ensemble. This analysis only considers cases in which an image was misclassified into an adjacent diabetic retinopathy grade. Table 13 shows the detailed misclassifications by adjacent grade.

Table 13. Adjacent-grade misclassification analysis

Actual Grade	**Predicted Grade**	**Number of Errors**
No DR	Mild	0
Mild	No DR	7
Mild	Moderate	24
Moderate	Mild	9
Moderate	Severe	1
Severe	Moderate	20
Severe	PDR	1
PDR	Severe	0

The largest adjacent-grade error was Mild DR to Moderate DR, which occurred in 24 cases, followed by Severe DR to Moderate DR, which occurred in 20 cases. Other errors of neighboring grades were relatively sparse, including 9 cases of Moderate DR classified as Mild DR, 7 cases of Mild DR classified as No DR and 1 case of Moderate DR classified as Severe DR. The results indicate that the proposed feature representation is not capable of distinguishing some neighboring stages of disease, particularly Mild-Moderate and Moderate-Severe stages.

The observed adjacent grade misclassification indicates that the image-derived feature representation may not be capable of capturing sufficiently discriminative characteristics for reliable separation of some neighboring stages of diabetic retinopathy. This limitation is especially noticeable in the transition from Moderate to Severe DR, where the underestimation of disease severity continues to be a major source of error.

9.5. Explainability

The issue of explainability was considered within the framework of the feature-based architecture adopted in the present study. Grad-CAM was not applied in the current experiment since the main classifiers operate on numerical image features extracted from images, rather than on convolutional feature maps. Instead, model-specific feature-importance techniques may be employed to evaluate how the extracted features contribute to the model predictions. The associated explainability techniques for the evaluated models are summarized in Table 14.

Table 14. Model-specific explainability approaches

Model	**Explainability Method**
SVM	Feature coefficients
Random Forest	Gini/permutation feature importance
XGBoost	Gain-based feature importance/SHAP
MLP	Permutation importance/SHAP
Ensemble	Model-level probability contribution + SHAP-based analysis

Although Grad-CAM was not applicable to the current feature based classifiers, future CNN based implementation may incorporate Grad-CAM to provide spatially localized explanations of retinal regions contributing to a model's prediction.

9.6 Comparative Analysis

The performance of the proposed soft-voting ensemble was compared to the performance of the four individual classifiers on the same held-out test set. This within-experiment comparison directly measures the improvement gained by combining the probability outputs of the individual models. The comparative improvements in accuracy, Macro-F1 and Cohen's kappa are shown in Table 15.

Table 14. Performance improvement of the proposed soft-voting ensemble over individual classifiers

Comparison	**Accuracy Improvement**	**Macro-F1 Improvement**	**Kappa Improvement**
Ensemble vs SVM	+8.91%	+0.0305	+0.1084
Ensemble vs Random Forest	+1.27%	+0.0388	+0.0275
Ensemble vs XGBoost	+1.82%	+0.0497	+0.0354
Ensemble vs MLP	+4.91%	+0.0380	+0.0650

The proposed ensemble improved the accuracy on the test set over all four individual classifiers. The biggest improvement in accuracy was observed over SVM with 8.91 percentage points and the smallest improvement was observed over Random Forest with 1.27 percentage points. The proposed ensemble also showed consistent improvements for all the four individual classifiers in terms of Macro-F1 and Cohen's kappa.

The steady improvement in the performance of the evaluated models shows that the probability outputs can be combined through the soft voting to give a complementary information to the prediction of the individual classifiers. However, the ensemble is not fully capable of solving the problem of recognizing minority and advanced DR classes as evidenced by the relatively low sensitivity for Severe and Proliferative DR.

Similarly, the class-weighted SVM experiment shows that the model with the best overall accuracy is not necessarily the best model for each clinical goal. Although the proposed ensemble obtained the best overall test-set performance, class weighting in SVM showed a significant improvement in detecting minority classes, particularly Severe DR. Hence, while designing classifiers for imbalanced medical datasets, it is important to evaluate the class-wise sensitivity as well as the overall performance.

9.7 Comparison with Existing Research

The deep-learning-based approaches have shown higher classification accuracies than the lightweight feature-based approach used here in earlier studies using the APTOS-2019 dataset. The results reported here show several deep learning methods with accuracies higher than 80%. Hence, the performance of the present framework is compared with previous five-class classification studies on the APTOS-2019 dataset, as shown in Table 15. Such comparisons should be taken with a grain of salt as the studies might involve different dataset partitioning, preprocessing of data, data augmentation, feature representation, class-balancing strategies, and evaluation protocols.

Table 15. Comparison with previously reported five-class diabetic retinopathy classification methods on the APTOS-2019 dataset

Method	**Classes**	**Reported Accuracy (%)**	**Dataset**
NASNet + SVM	5	77.90	APTOS-2019
Inception-ResNet-v2	5	82.18	APTOS-2019
Xception	5	79.59	APTOS-2019
ResNet50	5	74.64	APTOS-2019
InceptionV3	5	78.72	APTOS-2019
CNN512	5	83.40	APTOS-2019
CNN512 + Dropout	5	84.10	APTOS-2019
Supervised Contrastive Learning	5	84.32	APTOS-2019
Present Ensemble	**5**	**74.18**	**APTOS-2019**

The main contribution is a lightweight 132-feature pipeline evaluating multiple classical classifiers and a soft-voting ensemble, with detailed analysis of class imbalance, class-wise performance and minority-class sensitivity. Thus, the framework provides a computationally efficient benchmark for five-stage diabetic retinopathy classification.

10. Discussion

By appropriate pre-processing and feature extraction, traditional machine learning can retrieve useful information from retinal fundus images. The proposed ensemble achieved 74.18% accuracy and a ROC-AUC of 0.8458, outperforming the individual classifiers.

However, class imbalance affected minority-class detection. Sensitivity was high for No DR (95.94%) and Moderate DR (80.00%), but low for Severe DR (6.90%) and PDR (15.91%). The confusion matrix also showed underestimation of advanced disease.

Class-weighted SVM improved Severe DR sensitivity despite lower overall accuracy, highlighting the importance of balanced accuracy, Macro-F1, and class-specific sensitivity rather than accuracy alone.

The limitations of the 132-dimensional handcrafted features suggest that deep convolutional and transformer-based models could provide richer spatial representations and improve the detection of advanced DR.

10.1 Limitations

There are several limitations to consider when interpreting the results. The study relies on a single public dataset which limits external validation across institutions, devices and population. The dataset is also highly imbalanced with only 29 Severe DR images in the test set. The handcrafted features of 132 dimensions may be not enough to sufficiently capture complex retinal patterns. Additionally, only one train/validation/test split was used, so estimates of variability are limited. Finally, results do not mean clinical applicability or diagnostic reliability. Future work includes stratified cross-validation, bootstrap confidence intervals, external validation, improved imbalance-aware methods, learned image representations, and prospective clinical evaluation.

10.2 Ethical Implications and Data Management

This research utilizes a publicly accessible collection of retinal fundus-images and does not report on the recruitment of any additional human subjects. But the final thesis / journal submission must respect the conditions of use of the data of the competition/dataset APTOS 2019 and the institutional rules for secondary analysis of publicly available medical images. This experimental study does not claim any clinical use or a patient-specific diagnosis and treatment suggestion.

10.3 Summary

A lightweight imbalance-sensitive machine learning framework for five-stage diabetic retinopathy classification using retinal fundus images. The four classifiers (SVM, Random Forest, XGBoost and MLP) were trained with the same feature representation and their predictions were integrated via soft voting. The proposed ensemble achieved 74.18% accuracy, 46.54% balanced accuracy, 0.4748 Macro-F1, 0.7080 weighted-F1, 0.5892 Cohen's kappa and 0.8458 ROC-AUC on the 550-image test set. Sensitivity was high for No DR (95.94%) and Moderate DR (80.00%) and low for Severe DR (6.90%) and PDR (15.91%). The study therefore points out the importance of an evaluation sensitive to imbalance and not just accuracy.

10.4 Future scope

Future work should focus on deep learned representations, better handling of class imbalance, quantification of uncertainty, external validation and clinical relevance evaluation. The current framework should be considered as a lightweight and reproducible benchmark rather than a state-of-the-art clinical classifier.

References

[1] Jeong Y, Hong YJ, Han JH. Review of Machine Learning Applications Using Retinal Fundus Images. *Diagnostics*. 2022;12(1):134. doi:10.3390/diagnostics12010134.

[2] Yang Z, Tan TE, Shao Y, Wong TY, Li X. Classification of diabetic retinopathy: Past, present and future. *Frontiers in Endocrinology*. 2022;13:1079217. doi:10.3389/fendo.2022.1079217.

[3] Gonçalves MB, Nakayama LF, Ferraz D, et al. Image quality assessment of retinal fundus photographs for diabetic retinopathy in the machine learning era: a review. *Eye*. 2024;38(3):426–433. doi:10.1038/s41433-023-02717-3.

[4] Nakayama LF, Ribeiro LZ, Gonçalves MB, et al. Diabetic retinopathy classification for supervised machine learning algorithms. *International Journal of Retina and Vitreous*. 2022;8:1. doi:10.1186/s40942-021-00352-2.

[5] Zhelev Z, Peters J, Rogers M, et al. Test accuracy of artificial intelligence-based grading of fundus images in diabetic retinopathy screening: A systematic review. *Journal of Medical Screening*. 2023;30(3):97–112. doi:10.1177/09691413221144382.

[6] Butt MM, Awang Iskandar DNF, Abdelhamid SE, Latif G, Alghazo R. Diabetic Retinopathy Detection from Fundus Images of the Eye Using Hybrid Deep Learning Features. *Diagnostics*. 2022;12(7):1607. doi:10.3390/diagnostics12071607.

[7] Pinedo-Diaz G, Ortega-Cisneros S, Moya-Sanchez EU, et al. Suitability Classification of Retinal Fundus Images for Diabetic Retinopathy Using Deep Learning. *Electronics*. 2022;11(16):2564. doi:10.3390/electronics11162564.

[8] Nneji GU, Cai J, Deng J, Monday HN, Hossin MA, Nahar S. Identification of Diabetic Retinopathy Using Weighted Fusion Deep Learning Based on Dual-Channel Fundus Scans. *Diagnostics*. 2022;12(2):540. doi:10.3390/diagnostics12020540.

[9] Islam MR, Abdulrazak LF, Nahiduzzaman M, et al. Applying supervised contrastive learning for the detection of diabetic retinopathy and its severity levels from fundus images. *Computers in Biology and Medicine*. 2022;146:105602. doi:10.1016/j.compbiomed.2022.105602.

[10] Mehboob A, Akram MU, Alghamdi NS, Salam AA. A Deep Learning Based Approach for Grading of Diabetic Retinopathy Using Large Fundus Image Dataset. *Diagnostics*. 2022;12(12):3084. doi:10.3390/diagnostics12123084.

[11] Zang P, Hormel TT, Wang X, et al. A Diabetic Retinopathy Classification Framework Based on Deep-Learning Analysis of OCT Angiography. *Translational Vision Science & Technology*. 2022;11(7):10. doi:10.1167/tvst.11.7.10.

[12] Tajudin NMA, Kipli K, Mahmood MH, et al. Deep learning in the grading of diabetic retinopathy: A review. *IET Computer Vision*. 2022;16(8):667–682. doi:10.1049/cvi2.12116.

[13] Gu Z, Li Y, Wang Z, Kan J, Shu J, Wang Q. Classification of Diabetic Retinopathy Severity in Fundus Images Using the Vision Transformer and Residual Attention. *Computational Intelligence and Neuroscience*. 2023:1305583. doi:10.1155/2023/1305583.

[14] Sebastian A, Elharrouss O, Al-Maadeed S, Almaadeed N. A Survey on Deep-Learning-Based Diabetic Retinopathy

Classification. *Diagnostics*. 2023;13(3):345. doi:10.3390/diagnostics13030345.
[15] Shaukat N, Amin J, Sharif MI, et al. Classification and Segmentation of Diabetic Retinopathy: A Systemic Review. *Applied Sciences*. 2023;13(5):3108. doi:10.3390/app13053108.
[16] Huang Y, Lin L, Cheng P, et al. Identifying the Key Components in ResNet-50 for Diabetic Retinopathy Grading from Fundus Images: A Systematic Investigation. *Diagnostics*. 2023;13(10):1664. doi:10.3390/diagnostics13101664.
[17] Liu H, Teng L, Fan L, Sun Y, Li H. A new ultra-wide-field fundus dataset to diabetic retinopathy grading using hybrid preprocessing methods. *Computers in Biology and Medicine*. 2023;157:106750. doi:10.1016/j.compbiomed.2023.106750.
[18] Mohanty C, Mahapatra S, Acharya B, et al. Using Deep Learning Architectures for Detection and Classification of Diabetic Retinopathy. *Sensors*. 2023;23(12):5726. doi:10.3390/s23125726.
[19] Bajwa A, Nosheen N, Talpur KI, Akram S. A Prospective Study on Diabetic Retinopathy Detection Based on Modified Convolutional Neural Network Using Fundus Images. *Diagnostics*. 2023;13(3):393. doi:10.3390/diagnostics13030393.
[20] Ali AA, Dawood FAA. Deep Learning of Diabetic Retinopathy Classification in Fundus Images. *Journal of Engineering*. 2023;29(12):139–152. doi:10.31026/j.eng.2023.12.09.
[21] Sathwik AS, Agarwal R, Jubilson EA, Basa SS. Diabetic Retinopathy Classification Using Deep Learning. *EAI Endorsed Transactions on Pervasive Health and Technology*. 2023;9(1). doi:10.4108/eetpht.9.4335.
[22] Mukherjee N, Sengupta S. Comparing Deep Feature Extraction Strategies for Diabetic Retinopathy Stage Classification from Fundus Images. *Arabian Journal for Science and Engineering*. 2023;48:10335–10354. doi:10.1007/s13369-022-07547-1.
[23] Almattar W, Luqman H, Khan FA. Diabetic retinopathy grading review: Current techniques and future directions. *Image and Vision Computing*. 2023;139:104821. doi:10.1016/j.imavis.2023.104821.
[24] Bhulakshmi D, Rajput DS. A systematic review on diabetic retinopathy detection and classification based on deep learning techniques using fundus images. *PeerJ Computer Science*. 2024;10:e1947. doi:10.7717/peerj-cs.1947.
[25] Zhu S, Xiong C, Zhong Q, Yao Y. Diabetic Retinopathy Classification with Deep Learning via Fundus Images: A Short Survey. *IEEE Access*. 2024;12:20540–20558. doi:10.1109/ACCESS.2024.3361944.
[26] Romero-Oraá R, Herrero-Tudela M, López MI, Hornero R, García M. Attention-based deep learning framework for automatic fundus image processing to aid in diabetic retinopathy grading. *Computer Methods and Programs in Biomedicine*. 2024;249:108160. doi:10.1016/j.cmpb.2024.108160.
[27] Sundar S, Subramanian S, Mahmud M. Classification of Diabetic Retinopathy Disease Levels by Extracting Spectral Features Using Wavelet CNN. *Diagnostics*. 2024;14(11):1093. doi:10.3390/diagnostics14111093.
[28] Saproo D, Mahajan AN, Narwal S. Deep learning based binary classification of diabetic retinopathy images using transfer learning approach. *Journal of Diabetes & Metabolic Disorders*. 2024;23(2):2289–2314. doi:10.1007/s40200-024-01497-1.
[29] Grzybowski A, Jin K, Zhou J, et al. Retina Fundus Photograph-Based Artificial Intelligence Algorithms in Medicine: A Systematic Review. *Ophthalmic Epidemiology*. 2024. doi:10.1007/s40123-024-00981-4.
[30] Dai L, Sheng B, Chen T, et al. A deep learning system for predicting time to progression of diabetic retinopathy. *Nature Medicine*. 2024;30:584–594. doi:10.1038/s41591-023-02702-z.
[31] Gupta IK, Patil S, Mahadevkar S, et al. Retinal fundus imaging-based diabetic retinopathy classification using transfer learning and fennec fox optimization. *MethodsX*. 2025;14:103232. doi:10.1016/j.mex.2025.103232.
[32] Wei X, Liu Y, Zhang F, Geng L, Shan C, Cao X, Xiao Z. MSTNet: Multi-scale spatial-aware transformer with multi-instance learning for diabetic retinopathy classification. *Medical Image Analysis*. 2025;102:103511. doi:10.1016/j.media.2025.103511.
[33] Zhou Q, Guo Y, Liu W, et al. Enhancing pathological feature discrimination in diabetic retinopathy multi-classification with self-paced progressive multi-scale training. *Scientific Reports*. 2025;15:25705. doi:10.1038/s41598-025-07050-1.
[34] Renu DS, Saji KS. Hybrid deep learning framework for diabetic retinopathy classification with optimized attention AlexNet. *Computers in Biology and Medicine*. 2025;190:110054. doi:10.1016/j.compbiomed.2025.110054.
[35] Karthik SA, Geetha MN, Prabhavathi K, et al. Early Detection and Severity Classification of Diabetic Retinopathy Using Convolutional Neural Networks. *SN Computer Science*. 2025;6:819. doi:10.1007/s42979-025-04361-y.