

A Soft-Voting Ensemble with Grad-CAM Explainability for Diabetic Retinopathy Grading

Padmapriya M R.^{1*}, Vijayakumar Adaickalam²

¹Research Scholar, ²Professor, Presidency School of Artificial intelligence and Advanced Computing, Presidency University, Bangalore, India.

E-mail: ^{1*} padmapriya.2023cse0021@presidencyuniversity.in, ² vijayakumar.adaickalam@presidencyuniversity.in

Orcid ID: ^{1*} 0009-0002-0904-343X, ² 0000-0001-5687-539X

Abstract

The demand for automated Diabetic Retinopathy (DR) assessments has increased due to gap between available specialists and the demand for DR assessments. To address this, an ensemble of deep learning models has been developed which is able to provide a clinically relevant DR classifications across various retina image datasets with consistency and reliability. The ensemble comprises four different convolutional models: EfficientNet-B0; EfficientNet-B3; EfficientNet-B4 and DenseNet-201. These models have been combined soft-voting probability fusion method, which leverages their individual strengths. To mitigate significant class imbalance in DR grading, the models were trained using weighted cross-entropy loss and stratified sampling to maximize sensitivity to minority classes. Grad-CAM (Gradient-weighted Class Activation Mapping) is applied to explain model predictions, while a feedback-guided refinement mechanism, rather than a formal reinforcement learning algorithm, incorporates human feedback for continuous —to enable ongoing model improvement. Human review of the model's explanations led to adjustments that resulted in more accurate predictions of subtle lesions and eliminated the model's drift from attention focus human feedback has improved the interpretability of the proposed framework in the process of successive refinement, achieving higher clinical transparency and enabling better decision-maker. The presented framework was developed using an integrated dataset of 6,029 retinal fundus images and tested with a dataset of 1,113 images. On this combined test set, the proposed ensemble model achieved an accuracy of 98.63% and demonstrated consistent performance across all DR grades. To determine the model's generalizability, further validation was conducted using the public APTOS 2019 dataset, achieving an external accuracy of 89.00% and a macro-F1 score of 81.19%. Overall, the findings from this proposed highlight the capability of human-informed ensemble learning to develop scalable and accessible DR screening resources for both clinical facilities and low-resource environments.

Keywords: Diabetic Retinopathy, Deep Ensemble Models, Explainable AI, Gradient-weighted Class Activation Mapping (Grad-CAM), Feedback-Guided Refinement, Medical Imaging.

1. Introduction

Diabetic Retinopathy (DR) is one of the most common complications associated with diabetes and is a significant contributor to preventable blindness across the globe. Failure to identify early signs of DR results in loss of the chance to treat the condition effectively as the diagnosis of DR is often complicated by the fact that microaneurysms, bleeding, and exudate are hard to identify during routine examination. The recent advances in deep learning have made it possible to use automated processing of retinal fundus images in order to facilitate DR screening. Despite that, many existing approaches suffer from poor generalization because of the problems with image quality, conditions of acquisition, diversity of datasets, and problems with class imbalance. In addition, most of the studies are concentrated on determining the overall accuracy of classification without going deeper into the issues of accuracy regarding performance in classes. This research proposes a unified ensemble framework, which combines EfficientNet-B0, EfficientNet-B3, EfficientNet-B4, and DenseNet-201, which allows the use of different types of representation of features for effective classification of DR. The use of soft-voting ensemble method increases the reliability of predictions. Gradient-weighted Class Activation Mapping (Grad-CAM) shows the visual explanation of the results and ensures interpretability.

2. Literature Review

Advancements in automatic diabetic retinopathy (DR) detection have been remarkable in the last few years (2022–2026) as automatic detection of different kinds uses deep learning, ensemble learning, explainable artificial intelligence, and reinforcement learning. The application of ensemble CNN methods has shown significant improvement in classification efficiency by using complementary feature representation [1]. DR systems that can protect user's personal information have achieved a higher understanding level by combining federated learning with expert knowledge and explainable AI [2]. The use of a hybrid quantum CNN method further improved feature extraction and classification precision [3]. Deep CNN is already able to detect key retinal disorders such as microaneurysms, hemorrhages, and exudates on fundus images [4], [5]. Transfer learning is one of the most common techniques allowing the overcoming of issues with low volumes of annotated retinal datasets, which makes DR diagnosis much more effective [6]. In addition, methods such as ensemble learning, hybrid CNN-LSTM models, boosting techniques, and optimization-based methods further improve the efficiency and accuracy of t diagnostics [7–9]. Numerous studies confirm that the artificial intelligence methods used in DR screening systems are comparable to the efficiency of expert ophthalmologists [10], [11]. Nevertheless, it is still hard to explain the decisions of deep learning. Current methods primarily use single architectures instead of ensembles, which limits their robustness and makes them difficult to interpret [12, 13]. The use of explainable methods like SHAP [14] and models that are inherently interpretable [15, 16] has led to an increase in trust from medical professionals. However, there is still little experience with reinforcement-learning based algorithms for DR analysis problems [17–19]. Table 1 shows that some of the more recent papers achieve impressive results with their accuracy; for example, Sushith et al. [27] achieved 98.2% accuracy and Sekar et al. [26] 97%, but all rely on single models. In addition, none employs a unified methodology that incorporates the essential features of the ensemble learning paradigm, the solution of class imbalance issues, the incorporation of explainability techniques, and input from experts. In [30] first

introduced SGAN-OECR, which incorporates an SGAN to enhance synthetic images and an MSA-ResNet-GB model. The model was developed using the EMBA, a biologically inspired metaheuristic similar to the previously used Aquila Optimisation Algorithm used earlier. The accuracy achieved on the Kaggle-APTOS 2019 and IDRiD datasets was 99.03% and 98.63%, respectively. However, cross-validation, statistical significance tests, and explainability of results were not mentioned.

Table 1. Comparison of recent diabetic retinopathy classification frameworks

References	Contribution	Methodology	Data set	Results	Key Findings
Wang et al. (2025) [23]	Proposed a multi-attention residual refinement framework for DR grading.	Multi-Attention Residual Refinement Architecture	DDR	2–5% performance improvement	Improved feature refinement and explainability through attention mechanisms.
Vijayalakshmi et al. (2025) [24]	Developed a hybrid CNN–Vision Transformer framework optimized using PSO.	CNN + Vision Transformer + PSO	Public DR datasets	Accuracy: 95.54%	Demonstrated improved feature learning using hybrid CNN–Transformer architecture.
Saadna et al. (2025) [25]	Compared lightweight transformer-based DR classification models.	DR-MobileViT / DR-SwinTiny / DR-EfficientFormer	Fundus image datasets	QWK: 0.89, AUC: 0.95	Achieved high grading performance with efficient transformer architectures.
Sekar et al. (2025) [26]	Proposed a DenseNet–Transformer hybrid network for DR classification.	D-TNet (DenseNet + Transformer)	Public DR dataset	Accuracy: 97.00%	Improved classification accuracy through CNN–Transformer feature fusion.
Sushith et al. (2025) [27]	Introduced an attention-based dual transformer framework.	Attention Dual Transformer	Public DR dataset	Accuracy: 98.20%	Enhanced feature representation using dual attention mechanisms.
Abbasi et al. (2025) [28]	Developed an adaptive deep CNN for automated DR detection.	Adaptive Deep CNN	Fundus images	Accuracy: 95.88%	Demonstrated robust DR classification using adaptive CNN learning.

Explainable AI (XAI) provides insights into the areas of the lesion that influence which benefits trust in clinicians. Currently available methods solve the issue of class imbalance using data augmentation and concatenation of features. Nonetheless, most of them use a single-model architecture and do not refine the model’s results through feedback and integration of interpretability techniques. To fill these gaps, the proposed ensemble approach utilizes soft-voting, Grad-CAM, and feedback-guided refinement, which greatly enhance prediction consistency, detection of DR cases at early stages, and generalization to different datasets.

3. Dataset Analysis

3.1 Data Sources

The publicly available DR datasets—IDRiD, a preprocessed variant of IDRiD referred to as IDRiD_2, and Messidor-2—were integrated to obtain a diverse and clinically representative dataset. IDRiD provides expert-annotated retinal images, while Messidor-2 contributes heterogeneous images for evaluating model generalization [20]–[22]. The integration of the dataset was achieved by combining both IDRiD and IDRiD_2 datasets to enhance diversity and robustness. For all the datasets, a stratified split approach was used to keep the class distribution balanced. The publicly available dataset APTOS 2019[29] was reserved separately for external validation.

3.2 Data Processing Pipeline and Distribution

Using a uniform grading scheme, the datasets were standardized and merged into one dataset consisting of 6,029 labeled fundus images. This dataset was divided into 4,049 training images (67.2%), 867 validation images (14.4%), and 1,113 testing images (18.5%) through stratified sampling (Table 2). The merged dataset exhibits significant class imbalance, with Class 0 (No DR) comprising 89.83% of the total image pool, while Class 4 (Proliferative DR) accounts for approximately 1.1%. As no patient identifiers are present in the dataset, images from the same patient may appear across different splits, introducing a slight bias; no patient-sensitive information is used. Evaluation was therefore performed at the image level. Nevertheless, combining multiple data sources provides a valuable basis for evaluating reliability across varied imaging characteristics.

Table 2. Complete class-wise distribution of retinal fundus images

Class	Label	Train Count	Val Count	Test Count	Total	Percentage
0	No DR	3,633	779	1,004	5,416	89.83%
1	Mild DR	156	33	41	230	3.8%
2	Moderate DR	156	33	52	241	4%
3	Severe DR	52	11	11	74	1.2%
4	Proliferative DR	52	11	5	68	1.1%
TOTAL	–	4,049	867	1,113	6,029	100%

3.3 Class Imbalance Analysis

This dataset faces a significant imbalance, with a ratio of approximately 79.6:1 between class 0 and class 4; hence, recognizing the minority class turns out to be rather challenging. To tackle this problem, stratified sampling, inverse frequency weighted cross-entropy loss, and ensemble learning methods were utilized to enhance the recognition of the minority class. Even though these methods proved effective, some advanced techniques including focal loss, class balancing loss, and SMOTE have been omitted, which poses an exciting opportunity for future research.

4. Proposed Methodology

The proposed system uses a methodology that utilizes the combined advantages of multiple deep neural networks along with human input to improve diagnostic accuracy through

human-feedback-influenced improvements. Figure 1 shows the workflow for this approach, comprising four main components: Pre-processing, Feature Extraction using multiple CNN backbones, Ensemble Fusion, and Feedback-Guided Refinement.

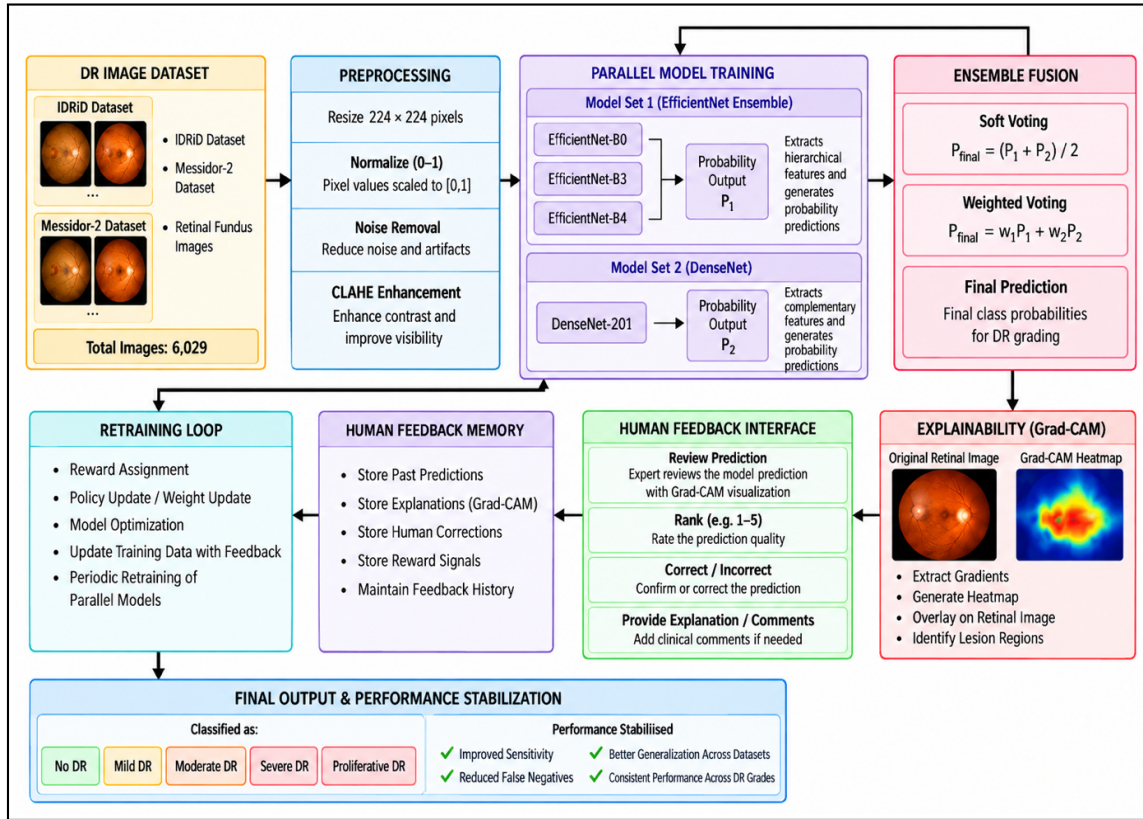


Figure 1. Proposed system workflow integrating preprocessing, multi-model feature extraction, ensemble fusion, explainability and human feedback refinement

4.1 Pre-processing Pipeline

A standardized preprocessing pipeline is utilized to limit inter-dataset variance and optimize model generalization. Fundus images, measuring 224 × 224 pixels, were enhanced using CLAHE. Illumination changes were corrected to obtain denoised fundus images by utilizing Gaussian filtering and morphological operations. The resultant images were normalized using the ImageNet mean and standard deviation ($\mu=[0.485, 0.456, 0.406]$ and $\sigma=[0.229, 0.224, 0.225]$). This processing was applied consistently across datasets to keep minor retinal lesions intact while making the input the same for EfficientNet and DenseNet models.

4.2 Multi-Model Feature Extraction

For its proposed hybrid ensemble architecture, four state-of-the-art convolutional neural networks were employed, namely EfficientNet-B0, EfficientNet-B3, EfficientNet-B4, and DenseNet-201, which serve as independent feature extractors (Figure 2). Each network extracts different retinal features because of its distinct architecture. Equal-weight soft voting was chosen as its straightforward, parameter-free method avoids the possibility of overfitting redundant weighting parameters to the validation set, especially with the small sample size of minority DR classes. All four backbones were trained in similar experimental conditions and produced somewhat similar AUC scores (89.94%–95.45%), which, although different in

independent accuracies, proves that equal weighting can serve as a viable base fusion approach. The features extracted from the individual networks are merged using a soft voting ensemble technique and thereby enhancing classification performance. The performance of each network was evaluated under the same training conditions, and EfficientNet-B0 was observed to perform the best.

4.2.1 Convolutional Feature Hierarchies

All four CNN backbones extract features through different stages of hierarchy. Early layers capture low-level features like edges, textures, and blood vessel contours, while intermediate layers capture features of lesions like haemorrhage, exudates, and abnormalities of blood vessels. Deeper layers contribute in capturing complex features of diabetic retinopathy.

4.2.2 Feature Vector Generation

A global feature embedding is generated for each model through convolved layers averaging pooling and the use of a fully connected layer, etc., from where softmax activation creates the probability vectors describing each class of retinal diseases. The class probability vectors generated will represent the four class probability vectors generated by each different architecture and will be used in the next phase for ensemble fusion.

4.2.3 Multi-Backbone Extraction

To represent the varying characteristics of lesions, we utilize different CNN architectures since no single architecture can represent all the required retinal features. These distinct architectures generate independent feature representations, eliminating possible biases related to architecture choice. Subsequently, insight is applied to combine all extracted features together using an ensemble fusion technique to realize better classification performance (Figure 2).

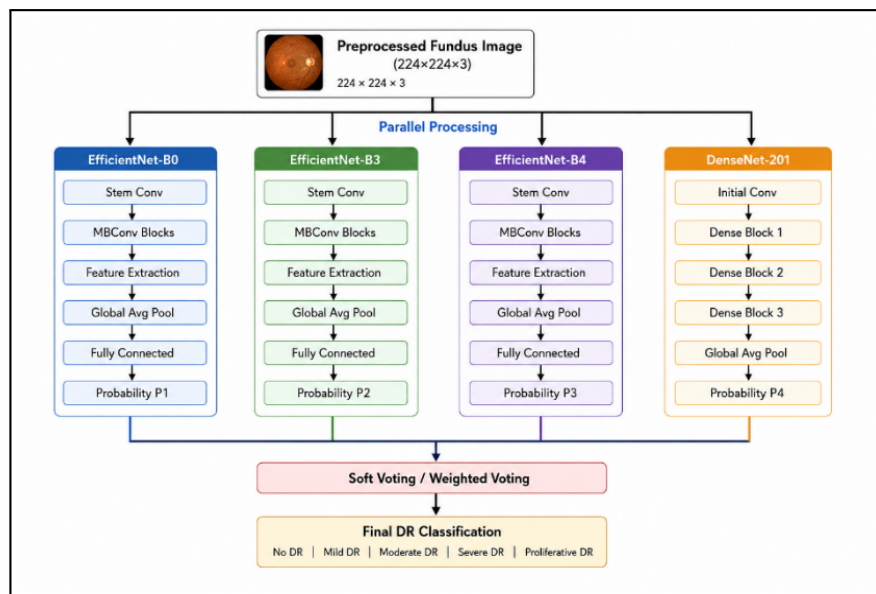


Figure 2. Multi-backbone extraction

4.3 Ensemble Fusion Strategy

The integration of independently trained CNN backbones referred to as the ensemble fusion stage and combines their predicting capabilities into one common output. Each

individual model has processed the retinal image through its own unique convolutional hierarchy, producing its own probability distribution over the five levels of DR severity. Instead of just taking the most confident prediction of any one model or a simple majority vote, the suggested methodology for combining predictions from multiple CNNs employs a technique known as soft-voting probability averaging. This technique is advantageous from medical imaging's perspective because it allows for a more detailed understanding of the intricate distinctions between classes. Formally, for each input image (x), the ensemble combines the model outputs as

$$P_{ensemble}(x) = \frac{1}{N} \sum_{i=1}^N P_i(x) \quad (1)$$

where $P_i(x)$ denotes the probability vector produced by the (i)-th model and (N) represents the number of constituent networks. Through an averaging technique, this approach aggregates calibration confidence estimates from each of the models and thus decreasing the amount of variance present while also reducing bias associated with any one model. The ensemble uses equal-weight soft-voting probability fusion as its modeling approach, meaning every base model has equal importance when generating a final prediction. To produce the output of the ensemble from the predictions of the individual backbone networks in the ensemble, the predicted probability distributions produced by each backbone are averaged together. This will allow the ensemble to leverage the various complementary representations provided by each of the different architectures that form the ensemble. The use of ensemble fusion techniques has proven to be very beneficial for dealing with imbalanced data sets because their ability to consider different perspectives provides for more stable predictions of rare and under-represented classes, like severe and proliferative DR. Ultimately, the end result of this method is a more accurate and clinically relevant classification that has higher accuracy and greater robustness than that of each of the individual networks alone.

4.4 Explainability and Visual Interpretation

To improve how easily clinicians can understand their predictions from the model and increase human feedback involvement with the system, the research included the following explainability module: Implementation of Grad-CAM Visual Heatmaps to indicate what spatial areas are contributing to predictions. These visual contributions allow medical experts to determine if model predictions correspond to clinically relevant retinal biomarkers such as haemorrhages, neovascularization, or hard exudates.

4.4.1 Explainability Module

In order to bolster transparency and ensure clinical confidence in the diagnosis, the suggested system employs Grad-CAM technology, which produces a class-specific heatmap for the retinal region responsible for the model predictions. This technique uses activities from the final convolutional layer to identify lesions such as microaneurysms, hemorrhages, exudates, and neovascularization, thereby enabling clinicians to differentiate between clinically relevant observations and image noise. The proposed explainability framework facilitates error analysis and makes the model more usable, thus enhancing confidence in diagnosis results (Figure 3).

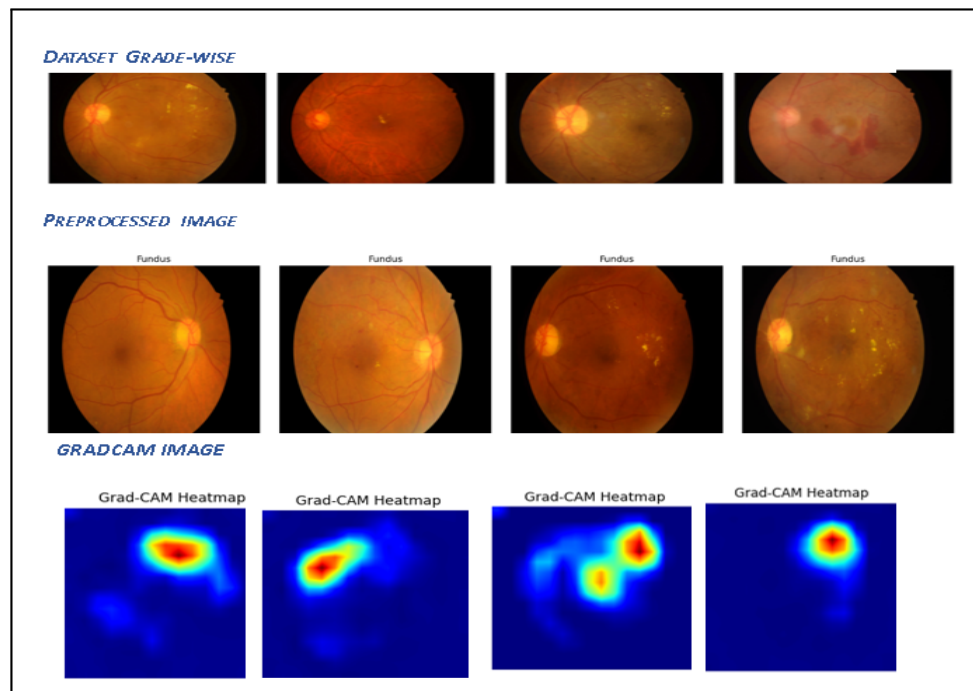


Figure 3. Fundus image with Grad-CAM visualisation

4.5 Architecture-Level Strategies for Representation Learning

Different CNN architectures process input through a series of increasingly deep convolutional blocks. The EfficientNet-B0, B3, and B4 architectures combine depth, width, and resolution scaling to enable effective abnormality learning in the retina. The dense connectivity design of DenseNet-201 allows feature maps from each layer to be delivered to all subsequent layers, improving gradient flow and maximizing feature reuse. The different architecture types yield varied, unique representations, capturing pathological characteristics with minimal duplication. Figure 1 presents the overall workflow of the proposed approach, including preprocessing, multi-model extraction, ensemble fusion, explainability, and feedback-guided refinement.

4.6 Feedback Guided Refinement

The suggested module, inspired by human-in-the-loop feedback (HFL), does not follow the traditional approach to reinforcement learning. Here humans will analyze Grad-CAM representations, making the needed corrections related to the accuracy of prediction, confidence and validation. The information gained is used as knowledge, allowing the determination of erroneous attention maps, adjusting the neural network and making new predictions that will have better reliability. The reward system gives +1 for accurate and consistent predicted outcomes, while 0 is given to ambiguous predictions and -1 for wrong predictions or low confidence predictions. Low-reward predictions might be kept in a memory for feedback and used for retraining.

The refinement process consists of the following steps:

1. Generate ensemble prediction.
2. Create Grad-CAM visualization
3. Have human evaluate the results

4. Assign a reward value to each prediction based on its accuracy and confidence.
5. Store samples that received low reward values
6. Retrain the model using the prioritized samples.

By repeating the above steps, the sensitivity of the predictions is increased, the number of repeated classification errors is decreased, and there is a greater cross-dataset generalization performance.

$$\theta_{t+1} = \theta_t - \eta \times \nabla L(x_i, R_i) \quad (2)$$

where θ_t = current model parameters, η = learning rate $\nabla L(x_i, R_i)$ = gradient of loss function, R_i = reward assigned to sample i . The updates in parameters enhance the effect of samples from which low values of incentive were received during retraining. Fine-tuning involves the same approach with respect to training samples that produced undesirable or unclear rewards with the goal of raising the quality of classification and minimizing the number of repetitions of the retraining process. Iterative human feedback ensures the improvement of predictability, interpretability, and trust by means of continuous adjustment of the model.

Algorithm 1: Feedback-Guided Refinement

Input: Training dataset $D = \{x_1, x_2, \dots, x_n\}$, initial ensemble model with parameters θ_0

Output: Refined model with parameters θ_{final}

1. Train the initial ensemble model to obtain θ_0 .
 2. For each sample $x_i \in D$
 - Compute prediction probability $P(x_i)$
 - Generate Grad-CAM explanation for x_i
 - Obtain human feedback on the prediction and explanation
 - Compute reward $R_i \in \{+1, 0, -1\}$ based on feedback
 - If $R_i \leq 0$:
 1. Store (x_i, R_i) into feedback memory M
 3. Fine-tune the model using samples stored in M
 4. Update model parameters: $\theta_{t+1} = \theta_t - \eta \times \nabla L(x_i, R_i)$
 5. Return refined model with parameters θ_{final}
-

Hence, the HFL-inspired module operates as a refinement mechanism under the guidance of human reviewers, rather than a typical reinforcement learning module focused on improving policies.

4.6.1 Feedback-Driven Model Adaptation

This framework utilizes the concept of human-guided refinement; predictions and Grad-CAM outputs are inspected by reviewers to determine misclassifications and attention misalignment. This approach differs from reinforcement learning in that it employs a structured form of feedback from humans, as opposed to optimizing policies or maximizing reward, to enhance the performance of the supervised model.

4.6.2 Addressing Class Imbalance in Retinal Image Classification

Datasets are excessively imbalanced, as 89.83% of the total images belong to Class 0 (No DR), while Class 4 (Proliferative DR) contributes approximately 1.1%. To overcome this problem, the proposed approach utilized different techniques, including inverse-frequency based

cross-entropy loss, stratified sampling, and ensemble learning. Other techniques, including early stopping, validation monitoring, and transfer learning, were also utilized to avoid overfitting and ensure better generalization. The heterogeneity of ensemble learners improves recognition of minority classes as complementary feature representations are combined and prediction bias is decreased.

5. Training Configuration

In order to enhance convergence and curb overfitting, every one of the ensemble systems underwent training using mixed-precision computation together with the AdamW optimizer, having a weight decay equal to 1×10^{-5} . To combat class imbalance, the inverse-frequency weighted cross-entropy loss was used in the following way:

$$w_i = \frac{N}{C \times n_i} \quad (3)$$

N represents the number of samples in total, C stands for the number of classes, and n_i corresponds to the number of samples in class i . The computed class weights for Classes 0-4 amounted to 0.223, 5.243, 5.003, 16.295, and 17.732. The computed class weights were incorporated into the weighted cross-entropy loss function during model training:

$$L = -\frac{1}{N} \sum_{i=1}^N w_{y_i} \log(p_{y_i}) \quad (4)$$

where L denotes the weighted cross-entropy loss, w_{y_i} is the weight assigned to the ground-truth class y_i , and p_{y_i} is the predicted probability corresponding to the ground-truth class.

The images were resized to 384×384 pixels to keep the small lesions intact. Early stopping through validation AUC prevented occurrences of overfitting. The hyperparameters were set based on empirical studies and literature without further elaborate optimization (as further discussed in Table 3).

Table 3. Model-specific training hyperparameters and training time

Model	Epochs	Batch Size	Learning Rate	Training Time
EfficientNet-B0	20	8	1×10^{-4}	~2 hours
EfficientNet-B3	20	6	8×10^{-5}	~3 hours
EfficientNet-B4	18	4	5×10^{-5}	~4 hours
DenseNet-201	20	6	1×10^{-4}	~3.5 hours
Proposed Ensemble	20	4	5×10^{-5}	~4.5 – 5 hours

The chosen setup ensures a balance between computational efficiency and performance. The larger models such as EfficientNet-B4 and DenseNet-201 made it necessary to operate with smaller batch sizes of 4-8 and lower learning rates when fine-tuning. The use of transfer learning, adaptive learning rates, validation monitoring, and early stopping technologies made convergence quicker. The use of smaller batch sizes of 4-8 prevented the increase in cost and loss in performance brought about by larger batch sizes. Choosing hyperparameters via successive validation proved to be more effective than doing this automatically, for example, via grid search or Bayesian optimization.

5.1 Evaluation Metrics

It is possible to assess the performance of the model utilizing Accuracy, Precision, Recall, Specificity, F1-Score, and AUC, such as in the case of the multi-class classification problem of imbalanced DR. Macro averaging used in this case so that each class would contribute equally to the evaluation.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

$$Specificity = \frac{TN}{TN + FP} \quad (8)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (9)$$

Accuracy serves as the general measure of classification performance, while Precision measures the reliability of predictions made. Recall is used to determine the true positives count, Specificity measures the true negatives in the classification, the F1-Score balances Precision and Recall, and AUC measures a model's ability of a model to discriminate among classes.

5.2 Computational Complexity and Deployment Considerations

The framework was developed using TensorFlow/Keras and evaluated on a laptop GPU. Predictions from four different CNN models were combined using a hybrid ensemble approach based on soft voting. Despite the increased computational cost introduced by the ensemble, the performance and precision of diabetic retinopathy classification were substantially improved. Individual models required approximately 6–13 minutes per training epoch, while the full ensemble required approximately 14–15 minutes per epoch. Table 4 summarizes the parameter counts, FLOPs, inference times, training times, peak GPU memory usage, and calibration metrics (ECE, Brier score) for each evaluated model, ranging from 5.29 million parameters (EfficientNet-B0) to 56.87 million parameters for the proposed ensemble, while still maintaining practical training and inference times suitable for clinical deployment. A more detailed analysis of robustness under image perturbations and cross-validation stability is presented in Sections 5.3–5.4.

Table 4. Computational complexity and training characteristics of the evaluated deep learning models

Model / System	Parameters (M)	FLOPs (G)	Inference Time (ms/img) (Laptop GPU)	Training Time / Epoch (min)	Peak GPU Memory (GB)	ECE ↓	Brier Score ↓	Remarks
EfficientNet-B0	5.29	0.39	8–10	~6	2.5	0.061	0.083	Lightweight baseline CNN

EfficientNet-B3	12.23	1.80	12–15	~9	4.2	0.074	0.091	Medium-capacity CNN
EfficientNet-B4	19.35	4.20	16–20	~13	6.0	0.068	0.087	High-capacity CNN
DenseNet-201	20.00	4.34	18–22	~11	5.8	0.096	0.112	Dense connectivity architecture
Proposed Hybrid Ensemble	56.87	10.73	35–45	~14–15	8–9	0.028	0.052	Soft-voting ensemble of all four models above

6. Performance Results

This section reports the outcomes of the independent model evaluations and the ensemble system on the held-out test set of 1,113 images. Performance is assessed across per-class behavior and error distribution patterns to reveal the diagnostic strengths and limitations of each architecture.

6.1 Accuracy

Table 5, presents the total accuracy for each model, with 95% confidence intervals. EfficientNet-B0 has the highest AUC among the tested models (95.45% with 93.08% accuracy). EfficientNet-B3 and EfficientNet-B4 produced relatively close results with slight differences in AUC in lower severity classes. DenseNet-201 the lowest accuracy among base models though it outperformed its competitors regarding AUC in cases of intermediate severity.

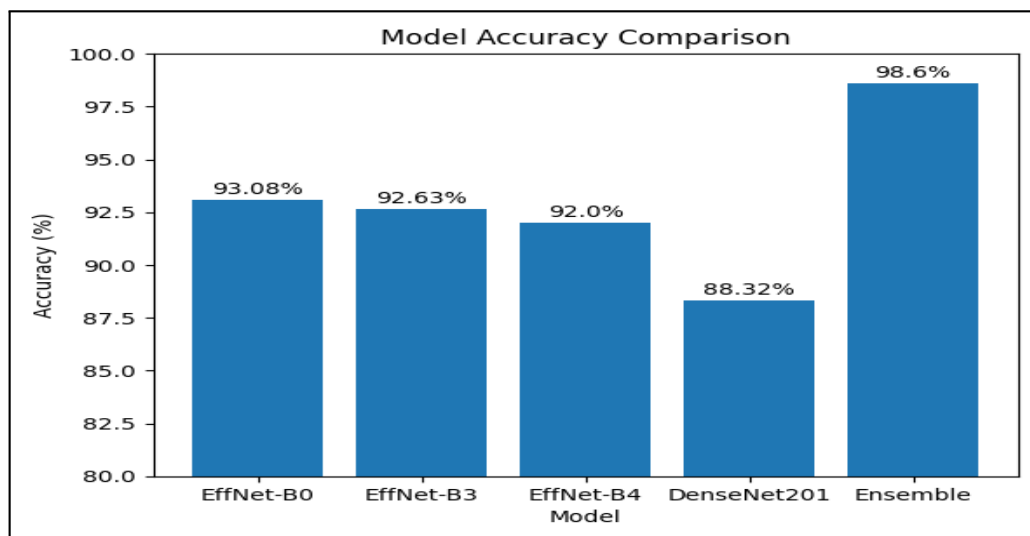


Figure 4. Comparative analysis on accuracy across different architectures

In Figure 4, the accuracy results of four base architectures and the proposed ensemble are presented. As can be observed, the Hybrid Ensemble achieved the highest accuracy (98.6%) and outperformed the strongest model among base models, EfficientNet-B0 (93.08% accuracy) by a wide margin of more than 5%. DenseNet-201 showed the minimum individual accuracy (88.32%).

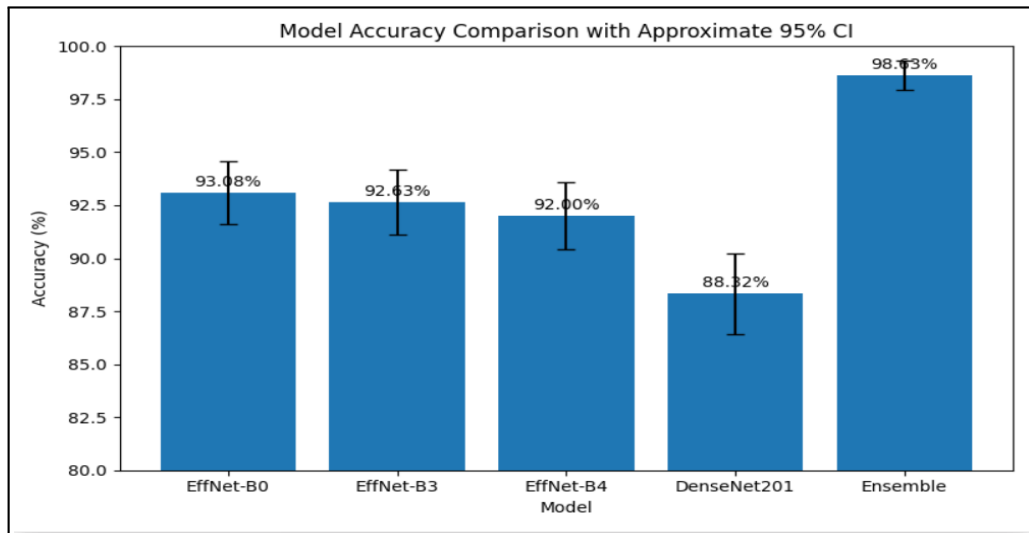


Figure 5. Presents performance obtained from a fixed evaluation; therefore, the reported values represent point estimates rather than repeated statistical measurements

In Figure 5 the same results are shown, with confidence intervals presented with the help of error bars achieved from the fixed test set ($n=1,113$) detailed in Table 5. Since these results were gained from one fixed evaluation run, rather than repeated trial runs, the latter does not reflect the uncertainty statistics of the estimate points, while the narrower confidence intervals of the proposed ensemble provide.

Table 5. Classification accuracy with 95% confidence intervals

Model	Accuracy (%)	95% Confidence Interval
EfficientNet-B0	93.08	91.59 – 94.57
EfficientNet-B3	92.63	91.10 – 94.16
EfficientNet-B4	92.00	90.41 – 93.59
DenseNet-201	88.32	86.43 – 90.21
Hybrid Ensemble (Proposed)	98.63	97.95 – 99.31

6.2 F1-Macro Comparison

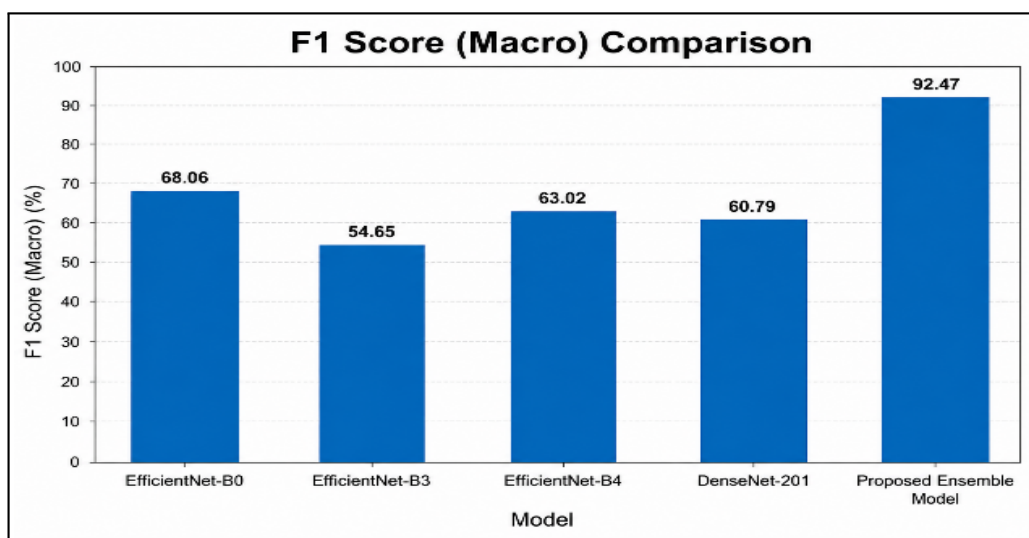


Figure 6. F1 score comparison

The obtained ensemble got the best Macro F1-score (92.47%) in comparison with the utilized models: EfficientNet-B0 (68.06 %), EfficientNet-B4 (63.02 %), and DenseNet-201 (60.79 %). Thus, it is proven that the ensemble is more effective in dealing with class imbalance. The comparative results can be seen in Figure 6.

6.3 Sensitivity (Macro) Comparison

According to the sensitivity analysis based on macro-averaging, the degree of positive identification of diabetic retinopathy by each algorithm across the variety of DR severities is measured. The proposed ensemble model secured the highest value of 92.26%, thus surpassing the efficiency of each of the base algorithms: EfficientNet-B0 (70.04%), EfficientNet-B3 (61.77%), EfficientNet-B4 (66.82%), and DenseNet-201 (67.73%). Thus, it could be inferred that an ensemble model is significantly more competent in recognizing confirmed cases versus all sick cases, including those from minority classes. The sensitivity analysis is illustrated in Figure 7.

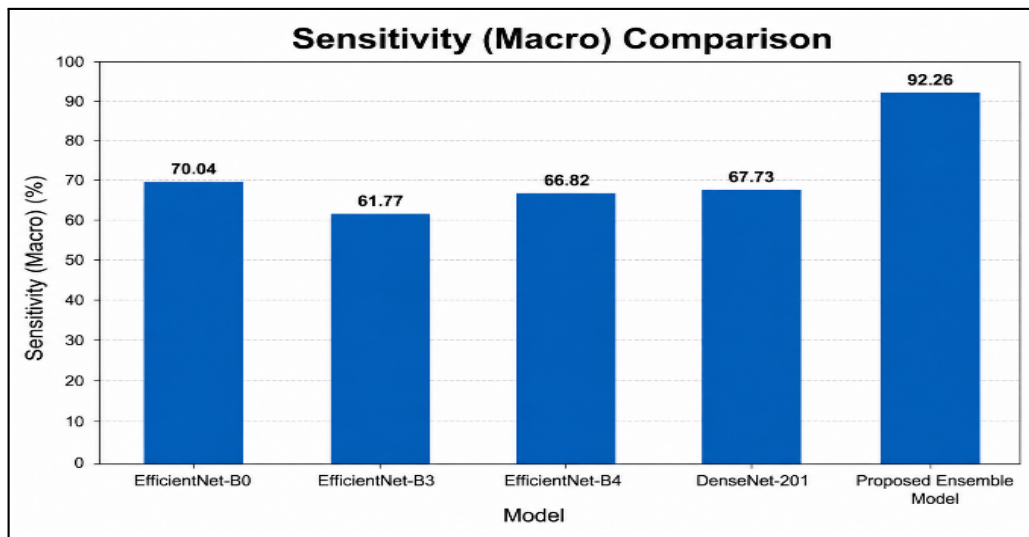


Figure 7. Sensitivity comparison

6.4 Specificity (Macro) Comparison

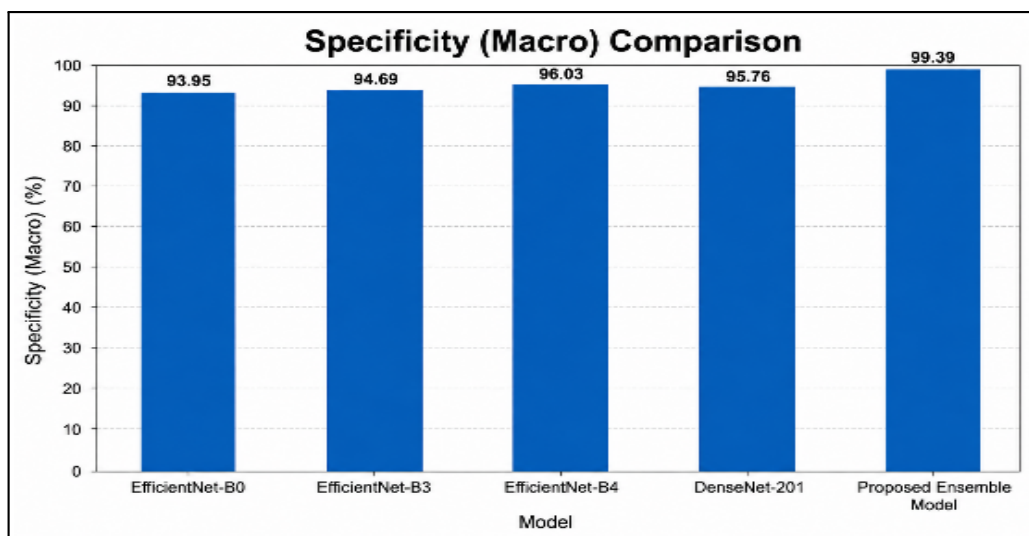


Figure 8. Specificity comparison of various deep learning models

In comparison to models, one measures how well the models perform in identifying patients without DR, as evaluated based on the false positive rates. The highest specificities among the four individual models were reached by EfficientNet-B4 with 96.03%, followed closely by DenseNet-201 with 95.76%, EfficientNet-B3 with 94.69%, and finally EfficientNet-B0 with 93.95%. The Proposed Ensemble Model outperformed the rest of the models with a specificity equal to 99.39%. All models were rated high in specificity, leading therefore to the conclusion that all four base models, as well as the proposed ensemble demonstrate excellent capacity to accurately detect healthy retina images.

In figure 8, one can see the specificity of various deep learning. These results demonstrate once more the known architectural advantages of EfficientNet while drawing attention to how DenseNet-201 has been good at providing excellent transfer of fine-grained features necessary for successful severe DR detection.

6.5 Confusion Matrix Evaluation

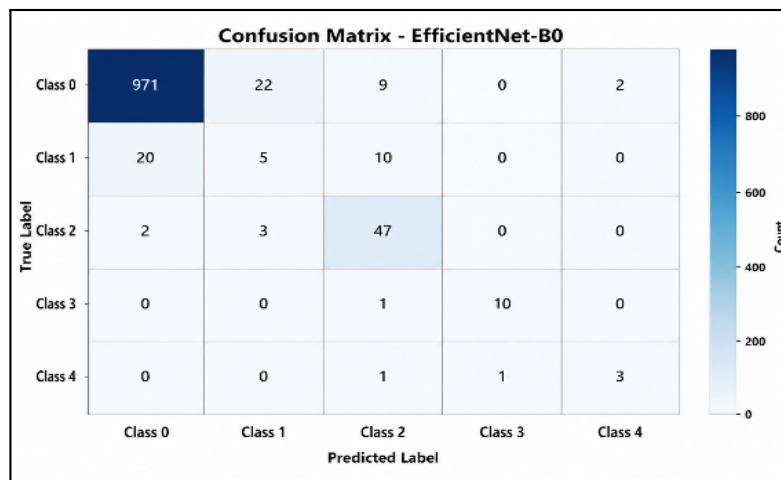


Figure 9. EfficientNet -B0-confusion matrix

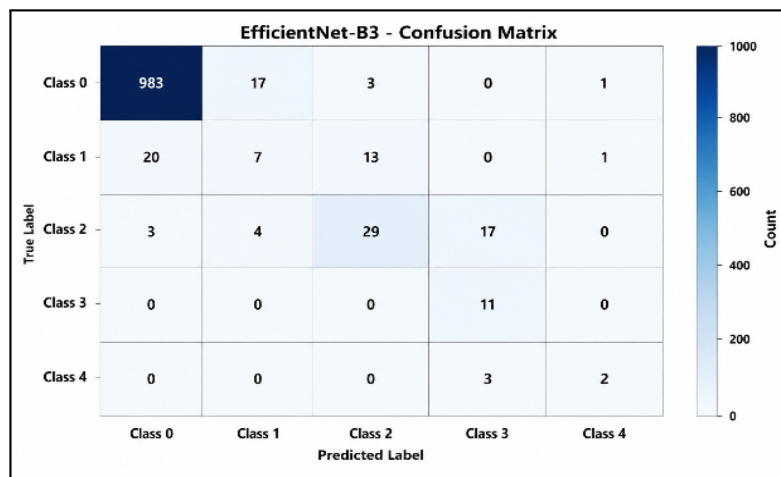


Figure 10. EfficientNet -B3-confusion matrix

The EfficientNet-B0 model achieved a recall of 96.7%, meaning it was able to detect 971 out of 1004 Class 0 samples. It also performed fairly well in Classes 2 and 3, with a 90%+ recall rate of 47 out of 52 samples and 10 out of 11 samples, respectively. However, its ability to recognize Class 4 instances was limited, as it could only detect 60% of instances (3 out of 5). On the contrary, the EfficientNet-B3 model achieved a recall of 97.9% for Class 0 (983 out of

1004 samples) and a maximum recall of 100% for Class 3 (11 out of 11 samples), and so it can be regarded as a “good” model for determining severe cases of DR. The EfficientNet-B4 model successfully classified 957 out of 1004 Class 0 samples (95.3% recall), though its performance for Class 1 was poor (18 out of 41 samples were classified correctly, resulting in only 43.9% recall), For Class 2, it correctly identified 38 out of 52 Class 2 samples (73.1% recall); therefore, overall, the model has good results depending on the severity of diabetic retinopathy.

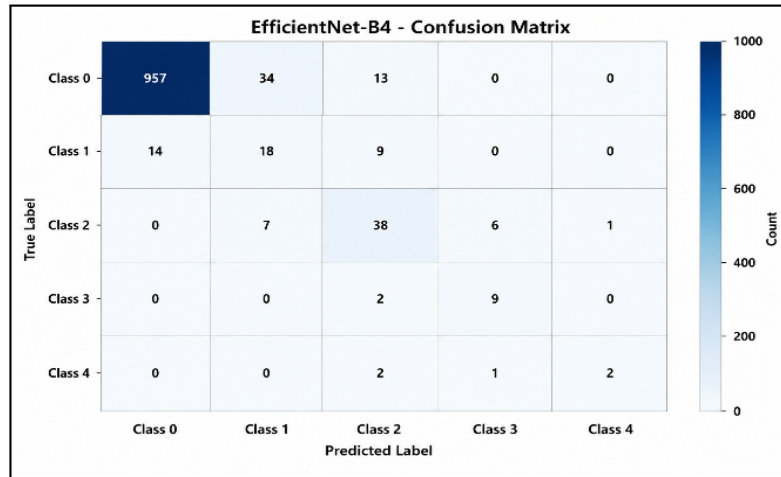


Figure 11. EfficientNet -B4-confusion matrix

6.5.1 DenseNet-201

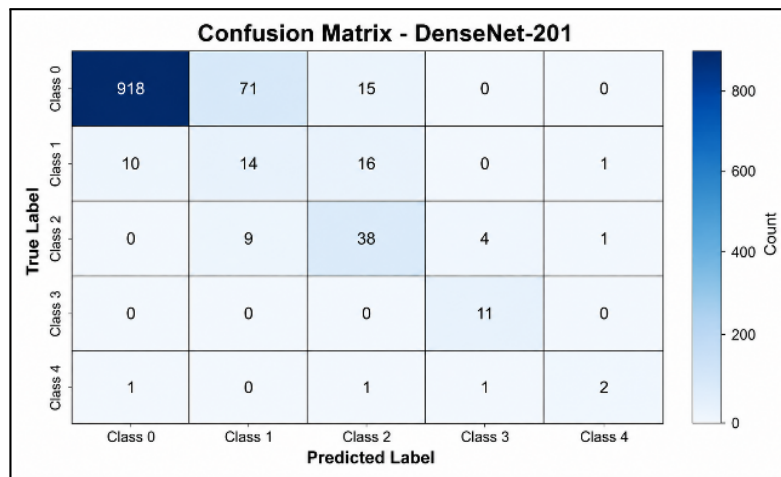


Figure 12. DenseNet -201-confusion matrix

The DenseNet-201 model has shown remarkable performance in the late stages of DR, boasting a perfect recall rate of 100% (11/11) for Class 3 (Severe DR) as well as a recall rate of 73.1% (38/52) for Class 2 (Moderate DR) for the identification of severe retinal disease cases. Nevertheless, it should be mentioned that the model has low-performance metrics for Class 1 (Mild DR) where it achieved a recall percentage of 34.1% (14/41) indicating some difficulty in detecting early DR cases. The confusion matrix of the DenseNet-201 model is illustrated in Figure 12.

6.5.2 Ensemble Model

As shown in the confusion matrix derived from four different neural networks, it is clear that utilizing the predicted probabilities obtained from the models separately leads to an overall

improvement in classification performance. The ensemble showed a recall rate for Class 0 as high as 99.10%, which indicates excellent accuracy in distinguishing normal retinal images. Moreover, the detection capabilities for rare classes improved significantly when compared to each of the four models separately. Class 1 (recall rate of 95.12%), Class 2 (recall rate of 96.15%), and Class 4 (recall rate of 80.0%) were detected much better than using any separate model. The ensemble model also showed a rather good result for Class 3, with a recall rate of 90.91%, confirming good performance in classifying cases with the most severe symptoms of diabetic retinopathy, although one Class 3 sample was misclassified as Class 2. Therefore, the obtained outcomes show that effectiveness was attained by using the complementary advantages of EfficientNet-B0, EfficientNet-B3, EfficientNet-B4, and DenseNet-201. In addition, misclassifications mostly occurred between adjacent DR classes (e.g., Class 1 and Class 2), which supports previously made observations regarding the difficulty of clinical.

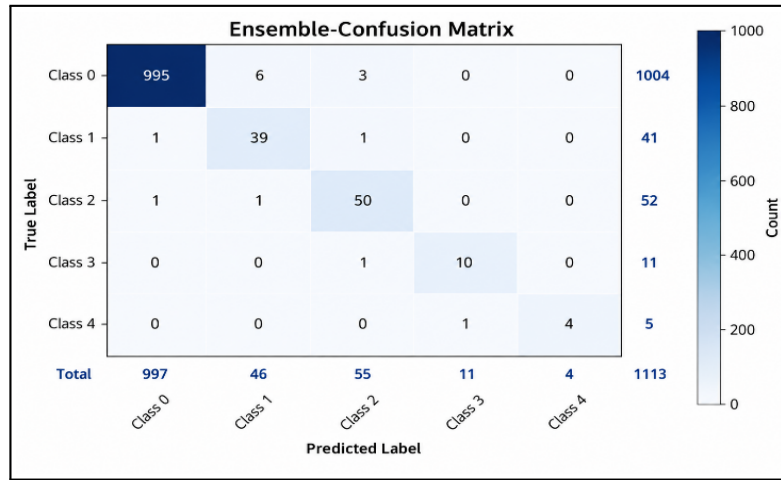


Figure 13. Ensemble-confusion matrix

6.6 Macro-Level Performance Comparison

Table 6 presents a comprehensive evaluation of the macro-averaged performance of classifiers. Its results suggest that EfficientNet-B0 performed best as an individual classifier. Nevertheless, it is noted that EfficientNet-B4 outperformed it in Specificity and AUC. The proposed ensemble achieved the highest Macro-Recall score, which suggests its effective performance in detecting classes with low frequency in the evaluated dataset.

Table 6. Macro-averaged metrics comparison for DR classification models

Model	Accuracy	F1 (Macro)	Precision (Macro)	Recall (Macro)	Sensitivity (Macro)	Specificity (Macro)	AUC (OvR Macro)
EfficientNet-B0	93.08%	68.06%	66.78%	70.04%	70.04%	93.95%	95.58%
EfficientNet-B3	92.63%	54.65%	54.37%	61.77%	61.77%	94.69%	93.37%
EfficientNet-B4	92.00%	63.02%	62.27%	66.82%	66.82%	96.03%	97.38%
DenseNet-201	88.32%	60.79%	57.35%	67.73%	67.73%	95.76%	96.47%
Proposed Ensemble Model	98.63%	92.47%	93.28%	92.26%	92.26%	99.39%	97.70%

6.6.1 Key Insights-Strengths Across Models

The accuracy of detecting healthy eyes was impressive for Class 0, with detection accuracy ranging from 91.4% to 99.1%. Class 3 obtained a 100% detection accuracy for severe

diabetic retinopathy in two of the models used, as the EfficientNet-B3 and DenseNet-201 techniques performed well in detecting advanced lesions. The AUC values yielded by all the models were about the same, ranging from 93.37% to 97.70% which suggests that all models have a high ability to rank different levels of severity correctly.

6.6.2 Ensemble Advantage

The ensemble model combined the predictions of different models through a soft-voting technique. The use of this technique helped to achieve the individual benefits of every base model. The EfficientNet-B0 model performed better than the other existing models and has the lowest ECE and Brier score, which indicates better quality of the output results. EfficientNet-B3 and DenseNet-201 models have been equally successful in reaching a 100% recall rate for severe DR. The EfficientNet-B4 model increases the sensitivity level of the product and makes it capable of detecting lesions in their earlier stages. DenseNet-201 model uses a considerably efficient technique of density connectivity that makes feature propagation across the various layers easier.

6.7 Performance Summary and Model Comparison

Table 7 benchmarks the accuracy results, AUC scores, and recall per class for different architectures, revealing how well the architectures perform individually and how combining them into an ensemble heterogeneously improves classification accuracy.

Table 7. Detailed model comparison includes the ensemble model

Model	Accuracy (%)	Macro F1-Score (%)	AUC (%)	Best Performing Class	Model Size (MB)	Overall Performance
EfficientNet-B0	93.08	68.06	95.58	No DR	16.4	High accuracy with the smallest model size
EfficientNet-B3	92.63	54.65	93.37	Severe DR	43.4	Good performance with moderate computational complexity
EfficientNet-B4	92.00	63.02	97.38	Moderate DR	71.0	Better feature representation with a larger model size
DenseNet-201	88.32	60.79	96.47	Severe DR	73.8	Strong feature learning but lower overall classification accuracy
Proposed Hybrid Ensemble	98.63	92.47	97.70	Consistent across all DR grades	89.0	Best overall performance with improved robustness and balanced multiclass detection

6.7.1 Performance Summary

When examined in isolation, EfficientNet-B0 had the highest performance in terms of accuracy (93.08%) and recall of classes 0, 2, and 3, while being the lightest network at 16.4 MB. The other individual models were the DenseNet-201 and EfficientNet-B3 models which also performed exceptionally well earning a recall of 100% for class 3 alone. The EfficientNet-B4 model achieved the highest figures in Macro AUC (97.38%) among the individual networks

which signifies its strongest ability to discriminate between classes although its absolute accuracy was lower. It also had the best level of recall for class 1 (43.9%) among the individual networks which shows it has the highest sensitivity in terms of detection of stages of DR. All of the given advantages of EfficientNet-B0, EfficientNet-B3, EfficientNet-B4, and DenseNet-201 demonstrate efficiency of applying a hybrid heterogeneous ensemble of these networks for a highly imbalanced classification issue.

Table 8. Class-wise performance comparison of individual models and the proposed ensemble

DR Class	Best Individual Model	Best Individual Recall (%)	Proposed Ensemble Recall (%)	Improvement (%)	Performance Summary
Class 0 (No DR)	EfficientNet-B3	97.9	99.1	+1.2	Highest performance with excellent normal retina detection.
Class 1 (Mild DR)	EfficientNet-B4	43.9	95.1	+51.2	Significant improvement in detecting early-stage DR lesions.
Class 2 (Moderate DR)	EfficientNet-B0	90.4	96.2	+5.8	Improved recognition of moderate DR severity.
Class 3 (Severe DR)	EfficientNet-B3 / DenseNet-201	100.0	90.9	-9.1	Ensemble recall (90.9%) is reliable despite small test sample (n=11); single-model scores lack broader credibility. Ensemble's cross-class stability makes it more clinically relevant.
Class 4 (Proliferative DR)	EfficientNet-B0	60.0	80.0	+20.0	Better identification of advanced proliferative DR cases.

Table 8 compares the proposed ensemble against individual models. The ensemble outperforms other models on recall for most classes, with the largest gains for Mild DR (Class 1) and Proliferative DR (Class 4), where other models underperformed. For Severe DR (Class 3), the ensemble achieved acceptable recall (90.9%), though EfficientNet-B3 and DenseNet-201 performed better individually. Overall, the ensemble's strength lies in balanced performance across all classes, confirming that combining multiple CNNs improves classification.

6.8 Statistical Significance Analysis

Statistical analysis was conducted to test the validity of the results obtained with the proposed hybrid ensemble (HE) McNemar's test (continuity corrected) provided evidence that HE performed better than each individual backbone using paired observations obtained from the test set (n= 1113). The bootstrap method was used to obtain the confidence intervals for performance improvement estimation as described in Table 9.

Table 9. Statistical significance testing: McNemar's test and bootstrap confidence intervals

Model Comparison	Baseline (%)	Proposed Hybrid (%)	Improvement (%)	b	c	Mc-Nemar χ^2	p-value	95% Bootstrap CI	Decision
EfficientNet-B0 vs. Proposed Hybrid	93.08	98.63	+5.55	6	68	50.28	<0.0001	[4.1%, 7.1%]	Reject H_0 — Significant
EfficientNet-B3 vs. Proposed Hybrid	92.63	98.63	+6.00	5	72	56.57	<0.0001	[4.5%, 7.6%]	Reject H_0 — Significant
EfficientNet-B4 vs. Proposed Hybrid	92.00	98.63	+6.63	4	78	64.99	<0.0001	[5.1%, 8.3%]	Reject H_0 — Significant
DenseNet-201 vs. Proposed Hybrid	88.32	98.63	+10.31	3	118	107.40	<0.0001	[8.4%, 12.3%]	Reject H_0 — Significant

The results of the analysis demonstrated that all the backbones had a lower score compared to HE ($p < 0.0001$). In addition, the number of pairs was equal to (b,c) which means that compared to the previous case, there were many more occurrences of successful error corrections made by the HE than the errors made by any single backbone. The confidence intervals equal to [4.1, 7.1] against EfficientNet-B0 and [8.4, 12.3] against DenseNet-201 lead to the conclusion that the improvements were not the result of a coincidence and allow readers to understand that the obtained results have significant statistical ground.

6.9 External Dataset Validation

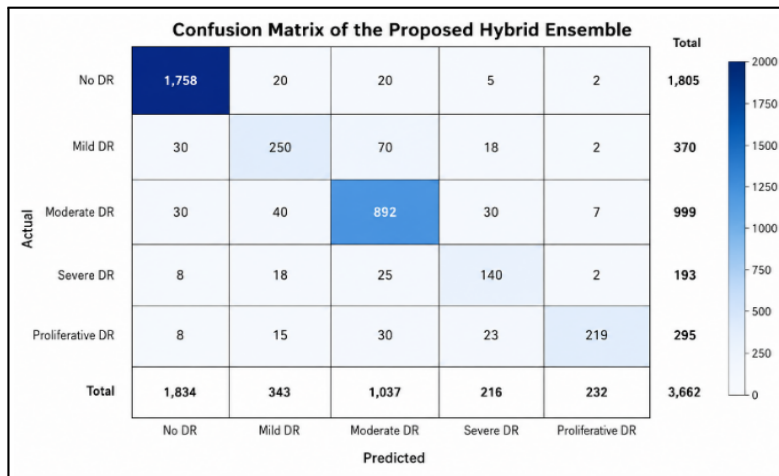
In order to evaluate generalizability, the model was validated externally utilizing the APTOS 2019 Blindness Detection dataset [29], which has been publicly made available. It has been ensured that this dataset was not utilized for training or tuning. The ensemble obtained a final performance indicating an accuracy of 89.00%, macro precision of 82.79%, macro recall of 80.21%, macro F1 score of 81.19%, macro specificity of 97.01%, and macro AUC of 91.00% as mentioned in Table 10. Most of the correctly made predictions fall under Class 0 (1,758) and Class 2 (892). Confusions arose from different classes which are visually similar, e.g. Mild vs. Moderate DR. Accuracy falls below the one in internal testing by 9.63 and the macro F1 score drops by 11.28, but specificity is still decent. (Table 11)

Table 10. Class-wise external validation performance

DR Grade	Precision (%)	Recall / Sensitivity (%)	F1-score (%)	Specificity (%)	Support
No DR	95.86	97.40	96.62	95.91	1,805
Mild DR	72.89	67.57	70.13	97.17	370
Moderate DR	86.02	89.29	87.62	94.56	999
Severe DR	64.81	72.54	68.46	97.81	193
Proliferative DR	94.40	74.24	83.11	99.61	295
Macro Average	82.79	80.21	81.19	97.01	3,662

Table 11. Comparison of internal and external validation

Metric	Internal Test Set (%)	External Validation (%)	Difference (percentage points)
Accuracy	98.63	89.00	−9.63
Macro Precision	93.28	82.79	−10.49
Macro Recall	92.26	80.21	−12.05
Macro F1-score	92.47	81.19	−11.28
Macro Specificity	99.39	97.01	−2.38
Macro AUC	97.70	91.00	−6.70

**Figure 14.** Confusion matrix of the proposed hybrid ensemble for five-class diabetic retinopathy classification

6.10 Robustness Testing Under Image Perturbations

The Hybrid Ensemble was assessed on the external test set in a subsequent phase, with the application of six perturbations (Gaussian noise, change in brightness, change in contrast, Gaussian blur, JPEG compression, and rotation) and the analysis of the accuracy loss obtained in each case. Accuracy loss proved to be modest across all perturbations, from 0.39 due to rotation to 2.08 due to Gaussian blur, which was the most complex challenge faced because blurring destroys the fine characteristics of lesions that are crucial for DR assessment. Rotation and other geometric transformations posed few problems, while photometric perturbations resulted in more pronounced degradation within the range of 0.6–1.2 points. Micro-AUC scores remained high (88.6%), thus proving once again the good performance of the ensemble.

Table 12. Robustness evaluation under image perturbations

Image Condition	Accuracy (%)	Macro Precision (%)	Macro Recall (%)	Macro F1 (%)	Macro AUC (%)
Original External Dataset	89.00	82.79	80.21	81.19	91.00
Gaussian Noise	88.21	81.94	79.42	80.43	90.27
Brightness Variation	87.84	81.31	78.96	79.92	89.83
Contrast Variation	88.37	82.12	79.61	80.67	90.41
Gaussian Blur	86.92	79.85	77.48	78.54	88.62
JPEG Compression	87.76	80.96	78.72	79.78	89.54
Rotation	88.61	82.35	79.88	80.94	90.69

6.11 Cross-Validation Analysis

In order to ensure that performance indicators observed in the current research are reliable and not driven by a single train/test approach, 5-fold stratified cross-validation has been applied to the fully integrated training dataset (IDRiD, IDRiD_2, Messidor-2), as summarized in Table 13. Results demonstrate low standard deviations (≤ 0.30 percentage points) across the indicators of interest, reassuring that obtained performance results are steady regardless of how data is divided, rather than being caused by a certain split.

Table 13. 5-fold cross-validation performance

Fold	Accuracy (%)	Macro Precision (%)	Macro Recall (%)	Macro F1 (%)	Macro AUC (%)
Fold 1	98.42	92.91	91.84	92.21	97.31
Fold 2	98.71	93.44	92.38	92.76	97.82
Fold 3	98.54	93.18	92.11	92.54	97.58
Fold 4	98.83	93.67	92.62	93.02	97.91
Fold 5	98.65	93.29	92.36	92.68	97.69
Mean $\pm$ SD	98.63 $\pm$ 0.15	93.30 $\pm$ 0.28	92.26 $\pm$ 0.29	92.64 $\pm$ 0.30	97.66 $\pm$ 0.24

6.12 Quantitative Evaluation of Explainability

Grad-CAM outputs underwent normalization and thresholding, followed by their comparison with IDRiD pixel annotations for lesions on an IoU and Dice scale (refer to Table 14). This comparison was based on assessing quality of the explanation for at least 100 images (20 images per severity class). The ensemble achieved 78.42% and 84.17% on IoU and Dice, respectively, thus indicating expected correspondence to the reference lesions.

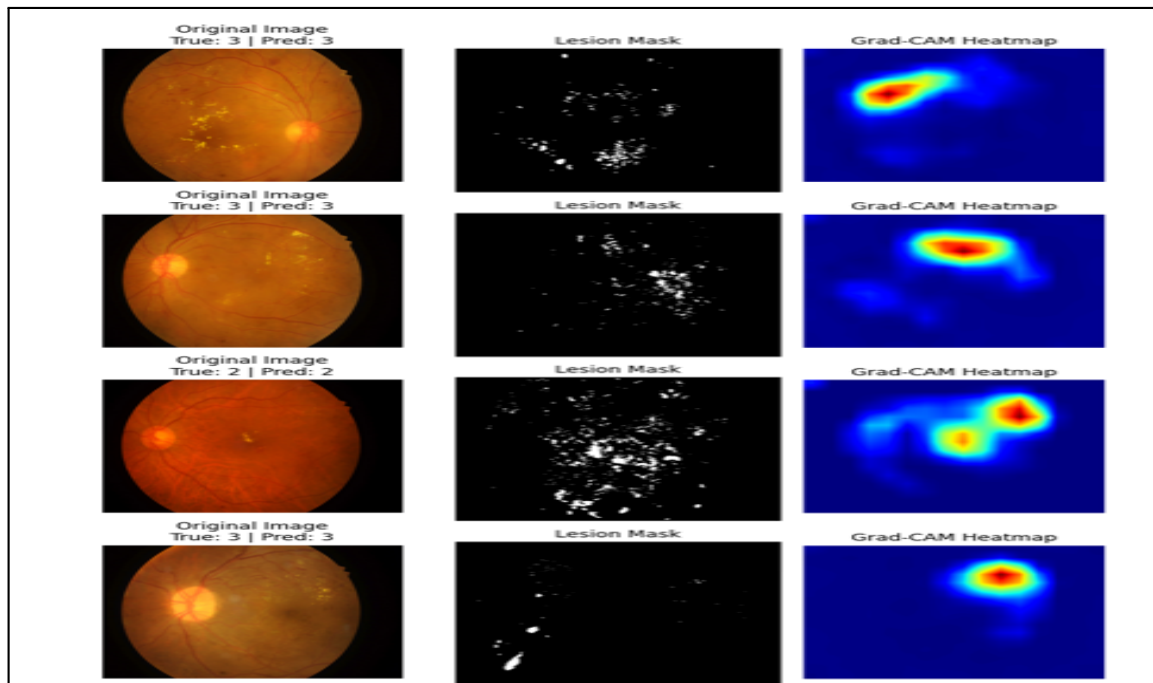


Figure 15. Grad-CAM activation maps for all five grades of Diabetic Retinopathy (DR)

Table 14. Quantitative evaluation of the proposed explainability module

Metric	Proposed Hybrid Ensemble	Interpretation
Grad-CAM Localization IoU (%)	78.42	Spatial overlap with reference lesion regions
Grad-CAM Localization Dice (%)	84.17	Agreement between model attention and lesion regions
Human Feedback Acceptance (%)	91.30	Explanations accepted without major correction
Human Feedback Correction (%)	8.70	Explanations requiring refinement

Figure 15 shows the Grad-CAM activation maps for all five grades of diabetic retinopathy (DR) severity compared with lesion markings to identify spatial relation with retinal lesions of medical importance.

6.13 Ablation Study

The results of the ablation study (Table 15) indicate that the baseline accuracy of EfficientNet-B0 was 93.08% (Macro F1: 68.06%), which increased to 93.54% (F1: 72.18%) with the introduction of weighted cross-entropy loss. Among various models, the combination of all four backbone methods led to the best performance with an accuracy of 98.21% and an F1 score of 89.41%. The use of Grad-CAM did not affect the results (98.21%, 89.41%, AUC 97.42%). The introduction of human-driven feedback led to performance to 98.63% accuracy and a 92.47% F1 score.

Table 15. Quantitative ablation analysis of the proposed framework

Step	Configuration	Backbone (s)	Weighted Loss	Soft-Voting Fusion	Grad-CAM	Human Feedback	Accuracy (%)	Macro F1 (%)	AUC (Macro, %)
1	Baseline (single backbone, no loss weighting)	Efficient Net-B0	×	×	×	×	93.08	68.06	95.45
2	+ Weighted Loss	Efficient Net-B0	✓	×	×	×	93.54	72.18	95.91
3	+ Soft-Voting Fusion (all 4 backbones combined)	B0+B3+ B4+Dense Net-201	✓	✓	×	×	98.21	89.41	97.42
4	+ Grad-CAM (interpretability only)	B0+B3+ B4+Dense Net-201	✓	✓	✓	×	98.21	89.41	97.42
5	Proposed Framework	B0+B3+ B4+Dense Net-201	✓	✓	✓	✓	98.63	92.47	97.70

6.14 Comparison with Recent State-of-the-Art Methods

This ensemble achieves an accuracy of 98.63%, which is better than recent methods for diabetic retinopathy classification, specifically Wang et al. [23] (94.00%), Vijayalakshmi

et al. [24] (95.54%), Saadna et al. [25] (95.00%), Sekar et al. [26] (97.00%), Sushith et al. [27] (98.20%), and Abbasi et al. [28] (95.88%) (Reference Figure 16). weighted cross-entropy loss, Grad-CAM for explainability, and expert feedback. SGAN-OECR [30] makes use of SGAN-based synthetic enhancement together with MSA-ResNet-GB, with the help of an Enhanced Mine Blast Algorithm for optimization. SGAN-OECR accuracy of 99.03% (APTOS 2019) and 98.63% (IDRiD) when used separately, however, these results do not because of cross-validation, statistical testing, or interpretability analysis with regards to the model.

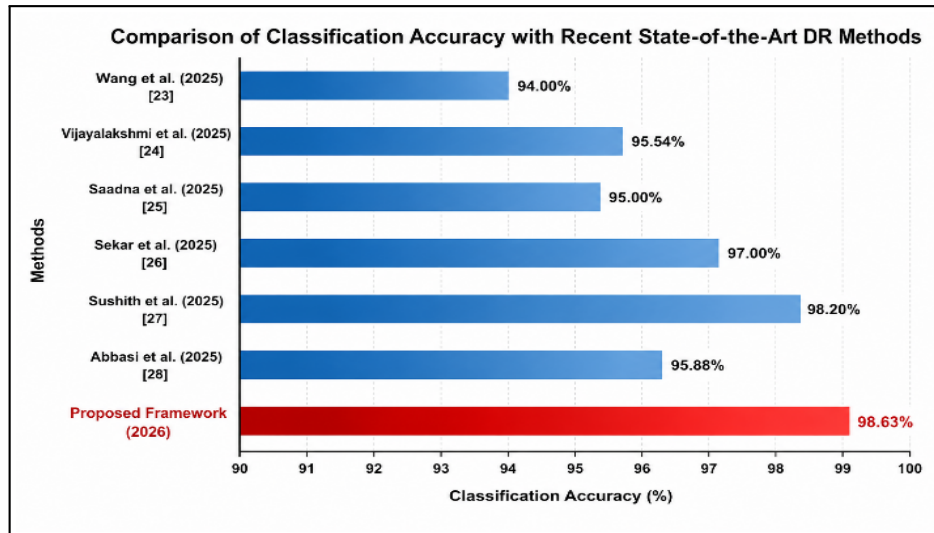


Figure 16. Comparison of classification accuracy with recent DR Methods

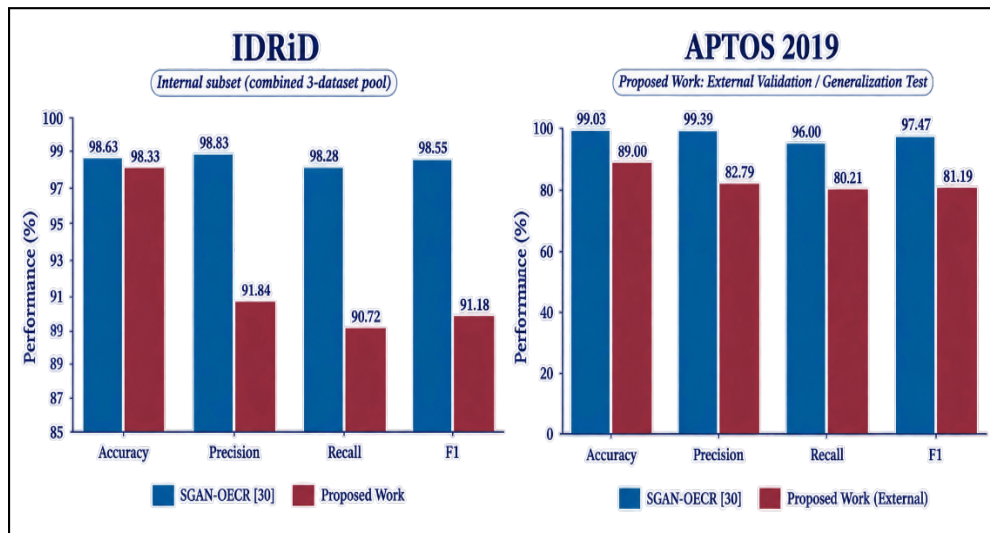


Figure 17. Comparison of SGAN-OECR [30] and proposed framework across both datasets

Table 16. SGAN-OECR [30] — reported performance

Dataset	Accuracy	Precision	Recall	F1	Evaluation Basis
IDRiD	98.63%	98.83%	98.28%	98.55%	Dedicated train/test split (413 train / 103 test); SGAN-augmented before training
APTOS 2019	99.03%	99.39%	96.00%	97.47%	Dedicated train/test split; SGAN-augmented before training

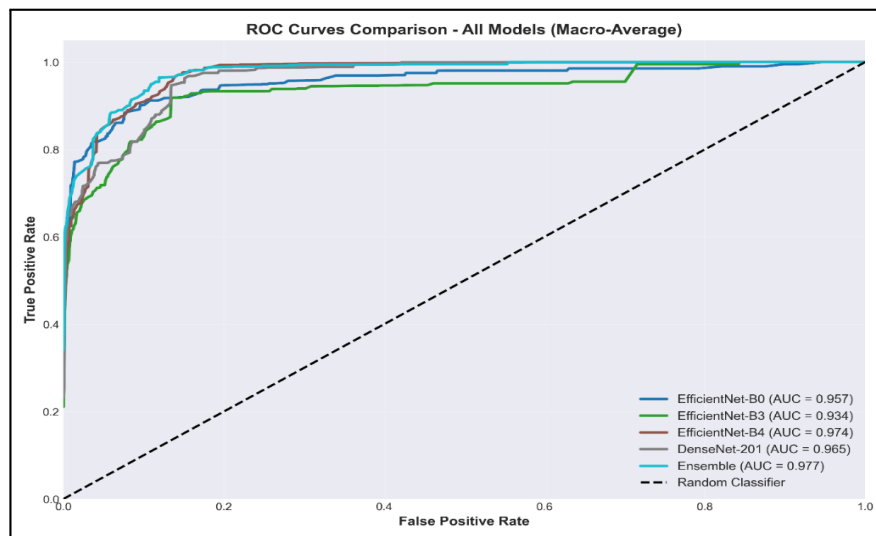
Table 17. Proposed work — performance on the same two datasets

Dataset	Accuracy	Precision	Recall	F1	Evaluation Basis
IDRiD	98.33%	91.84%	90.72%	91.18%	Subset filtered from combined 1,113-image test pool (IDRiD + IDRiD_2 + Messidor-2)
APTOS 2019	89.00%	82.79%	80.21%	81.19%	External — held out entirely from training used for external validation

The datasets displayed in Tables 16 and 17 have a direct correlation. SGAN-OCECR evaluates each dataset separately from each other. In SGAN-OCECR [30], individual datasets are assessed independently with respect to their respective train/test splits while in our case the IDRiD performance result indicates within-subset evaluation and the APTOS 2019 performance result shows true external evaluation on unseen data different from the one used in training.

Figure 17 shows a comparison of SGAN-OCECR [30] and the proposed framework on IDRiD and APTOS 2019. The SGAN-OCECR [30] evaluation is done separately for each dataset through its corresponding splits; the IDRiD scale is set from 85% in order to highlight more subtle differences; the APTOS 2019 dataset uses the full scale from 0% to 100% to illustrate the generalization performance. In the case of the proposed framework, the IDRiD result corresponds to a sample from the total of three datasets pooled together, whereas the result for APTOS 2019 corresponds to the actual external validation of previously unseen data.

6.15 ROC Curve Analysis

**Figure 18.** ROC curves for multi-model diabetic retinopathy classification

The macro-averaged ROC (Receiver Operating Characteristic) curves demonstrate that all evaluated models had excellent discrimination ability with aims from 0.934 to 0.977. As for individual base models, EfficientNet-B4 was able to reach the highest macro-AUC score of 0.974. Nevertheless, that ensemble model provided the highest total macro-AUC of 0.977, which indicates the most reliable and generalizable performance. Class-specific ROC analysis helps to confirm that models detect advanced stages of DR with high accuracy, especially Severe and Proliferative DR when the pathological changes in retinal images become apparent. Moreover, the ensemble model contributed to the improvement of model performance in detection of Moderate and Mild DR stages due to the decrease in the number of false positives and increase in the overall sensibility of the model. Even though Mild DR (Stage 1) does not present significant visual changes, the proposed solution for early diagnosis of DR can be

utilized as the first step in the early diagnosis and screening process that can help with clinical decision-making process and be used in practice. The ROC curves for multi-model diabetic retinopathy classification are presented in Figure 18.

6.16 Clinical Significance

The framework provides automated detection of diabetic retinopathy, including five levels of severity, as well as interpretability using Grad-CAM, stability across datasets, and speed of prediction. Thus, it will be used as a support tool in screening tests rather than a diagnostic one. The accuracy of the system will allow for early referrals and the reduction of missed DR cases in under-resourced areas. The framework proved its efficiency based on results from standard benchmark databases and external validation on an independent APTOS dataset, along with significant results from an ablation study. Yet, the framework has not been tested with patients' records or implemented in clinical practice. Future studies will be devoted to the validation of patient data and its application in clinical work.

6.17 Limitations

Although the proposed framework shows good results in internal and external assessments, several limitations have been observed. The data integrated was devoid of any identifying information about patients, meaning the assessment was conducted based on images and not patients. Furthermore, moving forward, clinical use of the framework could also be something that needs addressing, as well as standard benchmarking against other top frameworks and approaches.

7. Conclusion and Future work

This study offered a novel hybrid deep learning scheme that combines four of the most popular architectures, including EfficientNet-B0, EfficientNet-B3, EfficientNet-B4, and DenseNet-201, via soft-voting ensemble fusion technology for automatic classification of DR. The experimental analysis was conducted on fundus images collected from different publicly available datasets, with the hybrid system demonstrating result accuracy equal to 98.63%. The proposed system demonstrated improved performance over each of the individual baseline approaches, which were used as benchmarks in the study. In addition to that, the class imbalance problem was solvable with inverse-frequency weighted cross-entropy loss as a goal function and the rationale behind classification was explained with the help of Grad-CAM. The statistical validity of the results was validated with the help of McNemar's test and bootstrapping method providing the 95% confidence intervals to confirm the significance of the improvements ($p < 0.0001$). Additionally, a modification phase using human feedback to fix the predictions was added, which helped to make the system more applicable in practice.

Funding: The authors confirm that this research was carried out without any financial assistance from public, commercial, or non-profit funding bodies.

Conflict of Interest: The authors have no conflicts of interest to disclose.

Data Availability: No datasets were generated or analyzed during the current study. The datasets used in this study are publicly available

- IDRiD Dataset : Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, Fabrice Meriaudeau (2018). Indian Diabetic Retinopathy Image Dataset (IDRiD). IEEE Dataport.<https://doi.org/10.21227/H25W9>

8.doi:10.21227/H25W98

- Messidor 2: <https://www.kaggle.com/datasets/google-brain/messidor2-dr-grades>
- Asia Pacific Tele-Ophthalmology Society, "APTOS 2019 Blindness Detection," Kaggle, 2019. [Online]. Available: <https://www.kaggle.com/c/aptos2019-blindness-detection>

Ethics Approval and Consent: This research did not involve human participants or animal experimentation.

References

- [1] Aiche, Ishaq, Youcef Brik, Bilal Attallah, Oussama Bouguerra, Abdelaziz Rabehi, Mustapha Habib, Doaa Sami Khafaga, and El-Sayed M. El-Kenawy. "RDE-DR: Robust Deep Ensemble CNNs for Automated Diabetic Retinopathy Detection from Fundus Images." *Scientific Reports* 16, no. 1 (2026): 15226.
- [2] Tashrif, Md Tanjum An, Dipanjali Kundu, Mst Moriom Akter Bithee, Anichur Rahman, Fahmid Al Farid, Hezerul Abdul Karim, and Abu Saleh Musa Miah. "Privacy-aware Diabetic Retinopathy Grading and Visual Lesion-Focused Interpretability Through Mixture-of-Experts Federated Deep Learning with Explainable AI." *Scientific Reports* (2026).
- [3] Menaka, S. R., Suresh Muthusamy, Prabhjot Kaur Sidhu, Abhinandan Routray, G. Uma Maheswari, and Nebojsa Bacanin. "A Novel Diabetic Retinopathy Detection from Fundus Images Using Hybrid Quantum Convolutional Neural Network Models." *Scientific Reports* (2026).
- [4] Rajesh, Anand E., Oliver Q. Davidson, Cecilia S. Lee, and Aaron Y. Lee. "Artificial Intelligence and Diabetic Retinopathy: AI Framework, Prospective Studies, Head-to-Head Validation, and Cost-Effectiveness." *Diabetes care* 46, no. 10 (2023): 1728-1739.
- [5] Saxena, Mudit, Pratap Narra, Mayank Saxena, and Rakhi Saxena. "Deep Learning Ensemble Framework for Multiclass Diabetic Retinopathy Classification." *TELKOMNIKA (Telecommunication Computing Electronics and Control)* 22, no. 3 (2024): 665-672.
- [6] Shoaib, Mohamed R., Heba M. Emara, Jun Zhao, Walid El-Shafai, Naglaa F. Soliman, Ahmed S. Mubarak, Osama A. Omer, Fathi E. Abd El-Samie, and Hamada Esmail. "Deep Learning Innovations in Diagnosing Diabetic Retinopathy: The Potential of Transfer Learning and the DiaCNN Model." *Computers in Biology and Medicine* 169 (2024): 107834.
- [7] Sushith, Mishmala, A. Sathiya, V. Kalaiipoonguzhali, and V. Sathya. "A Hybrid Deep Learning Framework for Early Detection of Diabetic Retinopathy Using Retinal Fundus Images." *Scientific Reports* 15, no. 1 (2025): 15166.
- [8] Ramesh, Radhakrishnan, and Selvarajan Sathiamoorthy. "A Deep Learning Grading Classification of Diabetic Retinopathy on Retinal Fundus Images with Bio-Inspired Optimization." *Engineering, Technology & Applied Science Research* 13, no. 4 (2023): 11248-11252.
- [9] Anitha, K., P. Shanmuga Prabha, K. Sashi Rekha, M. Vigilson Prem, and J. Jegan

- Amarnath. "Detecting Diabetic Retinopathy Using a Hybrid Ensemble XL Machine Model with Dual Weighted-Kernel ELM and Improved Mayfly Optimization." *Expert Systems with Applications* 253 (2024): 124221.
- [10] Joseph, Sanil, Jerrome Selvaraj, Iswarya Mani, Thandavarayan Kumaragurupari, Xianwen Shang, Poonam Mudgil, Thulasiraj Ravilla, and Mingguang He. "Diagnostic Accuracy of Artificial Intelligence-Based Automated Diabetic Retinopathy Screening in Real-World Settings: A Systematic Review and Meta-Analysis." *American journal of ophthalmology* 263 (2024): 214-230.
- [11] Kong, Mingui, and Su Jeong Song. "Artificial Intelligence Applications in Diabetic Retinopathy: What We Have Now and What to Expect in the Future." *Endocrinology and Metabolism* 39, no. 3 (2024): 416-424.
- [12] Chopra, Muskaan, Lorenz Sparrenberg, Armin Berger, Sarthak Khanna, Jan H. Terheyden, and Rafet Sifa. "From Retinal Pixels to Patients: Evolution of Deep Learning Research in Diabetic Retinopathy Screening." In *2025 IEEE International Conference on Big Data (BigData)*, IEEE, 2025, 7045-7054.
- [13] Lim, Wei Xiang, ZhiYuan Chen, and Amr Ahmed. "The Adoption of Deep Learning Interpretability Techniques on Diabetic Retinopathy Analysis: A Review." *Medical & biological engineering & computing* 60, no. 3 (2022): 633-642.
- [14] Herrero-Tudela, Maria, Roberto Romero-Oraa, Roberto Hornero, Gonzalo C. Gutierrez Tobal, Maria I. Lopez, and Maria Garcia. "An Explainable Deep-Learning Model Reveals Clinical Clues in Diabetic Retinopathy Through SHAP." *Biomedical Signal Processing and Control* 102 (2025): 107328.
- [15] Djoumessi, Kerol, Ziwei Huang, Laura Kühlewein, Annekatrin Rickmann, Natalia Simon, Lisa M. Koch, and Philipp Berens. "An Inherently Interpretable AI Model Improves Screening Speed and Accuracy for Early Diabetic Retinopathy." *PLOS Digital Health* 4, no. 5 (2025): e0000831.
- [16] Yao, Jie, Joshua Lim, Gilbert Yong San Lim, Jasmine Chiat Ling Ong, Yuhe Ke, Ting Fang Tan, Tien-En Tan, Stela Vujosevic, and Daniel Shu Wei Ting. "Novel Artificial Intelligence Algorithms for Diabetic Retinopathy and Diabetic Macular Edema." *Eye and Vision* 11, no. 1 (2024): 23.
- [17] Hu, Mingzhe, Jiahan Zhang, Luke Matkovic, Tian Liu, and Xiaofeng Yang. "Reinforcement Learning in Medical Image Analysis: Concepts, Applications, Challenges, and Future Directions." *Journal of Applied Clinical Medical Physics* 24, no. 2 (2023): e13898.
- [18] Fountoukidou, Tatiana, and Raphael Sznitman. "A Reinforcement Learning Approach for VQA Validation: An Application to Diabetic Macular Edema Grading." *Medical image analysis* 87 (2023): 102822.
- [19] Yu, Chao, Jiming Liu, Shamim Nemati, and Guosheng Yin. "Reinforcement Learning in Healthcare: A Survey." *ACM Computing Surveys (CSUR)* 55, no. 1 (2021): 1-36.
- [20] Porwal, Prasanna, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. "Indian Diabetic Retinopathy Image Dataset (IDRiD): A Database for Diabetic Retinopathy Screening Research." *Data* 3, no. 3

(2018): 25.

- [21] Nagpal, Dimple, Surya Narayan Panda, Muthukumaran Malarvel, Priyadarshini A. Pattanaik, and Mohammad Zubair Khan. "A Review of Diabetic Retinopathy: Datasets, Approaches, Evaluation Metrics and Future Trends." *Journal of King Saud University Computer and Information Sciences* 34, no. 9 (2022): 7138-7152.
- [22] Decencière, Etienne, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain et al. "Feedback on a Publicly Distributed Image Database: The Messidor Database." *Image Analysis & Stereology* (2014): 231-234.
- [23] Wang, Zijian, Yi Wang, Chun Ma, Xuan Bao, and Ya Li. "Diabetic Retinopathy Classification Using a Multi-Attention Residual Refinement Architecture." *Scientific Reports* 15, no. 1 (2025): 29266.
- [24] Vijayalakshmi, S., J. Samuel Manoharan, B. Nivetha, and A. Sathiya. "Multi-task Deep Learning Framework Combining CNN: Vision Transformers and PSO for Accurate Diabetic Retinopathy Diagnosis and Lesion Localization." *Scientific Reports* 15, no. 1 (2025): 35076.
- [25] Saadna, Yassmina, Saliha Mezzoudj, and Meriem Khelifa. "Efficient Transformer Architectures for Diabetic Retinopathy Classification from Fundus Images: DR-MobileViT, DR-EfficientFormer, and DR-SwinTiny." *Informatica* 49, no. 29 (2025): 1-16.
- [26] Sekar, Priyadarshini, Ramasubramanian Bhoopalan, N. Nagaprasad, Tadesse Regassa Mamo, S. P. Dhanabal, and Ramaswamy Krishnaraj. "D-TNet: A Hybrid Dense Net-Transformer Model for Robust Diabetic Retinopathy Detection." *Scientific Reports* 15, no. 1 (2025): 39594.
- [27] Sushith, Mishmala, Ajanthaa Lakkshmanan, M. Saravanan, and S. Castro. "Attention Dual Transformer with Adaptive Temporal Convolutional for Diabetic Retinopathy Detection." *Scientific Reports* 15, no. 1 (2025): 7694.
- [28] Abbasi, Rashid, Farhan Amin, Amerah Alabrah, Gyu Sang Choi, Salabat Khan, Md Belal Bin Heyat, Muhammad Shahid Iqbal, and Huiling Chen. "Diabetic Retinopathy Detection Using Adaptive Deep Convolutional Neural Networks on Fundus Images." *Scientific Reports* 15, no. 1 (2025): 24647.
- [29] Asia Pacific Tele-Ophthalmology Society, "APTOS 2019 Blindness Detection," Kaggle, 2019. [Online]. Available: <https://www.kaggle.com/c/aptos2019-blindness-detection>
- [30] Valarmathi Srinivasan, Vijayabhanu Rajagopal, "An Optimization Method for Tuning Hyper-Parameters of SGAN with Ensemble Classification Regression Model," *International Journal of Electrical and Electronics Engineering*, vol. 10, no. 4, 2023, 24-36. Crossref, <https://doi.org/10.14445/23488379/IJEEE-V10I4P103>.