Transfer Learning Based Diabetic Retinopathy Grading System

Samia Akhtar, Muhammad Sulman Bashir, Shabib Aftab
Department of Computer Science, Virtual University of Pakistan, Lahore, Pakistan
*Correspondence: shabib.aftab@vu.edu.pk

Citation | Akhtar. S, Bashir. M. S, Aftab. S, "Transfer Learning Based Diabetic Retinopathy Grading System", IJIST, Vol. 8 Issue. 9 pp 3782-3811, October 2026
Received | Oct 01, 2026 **Revised |** Oct 21, 2026 **Accepted |** Oct 25, 2026 **Published |** Oct 28, 2026.

Diabetes-related disease known as diabetic retinopathy damages the retinal blood vessels and leads to vision impairment. Analyzing diabetic retinopathy images manually is time-consuming and susceptible to mistakes. Traditional detection methods relying on manually engineered features often lack flexibility and accuracy. This work is dedicated to creating a system that is both automated and reliable for detecting diabetic retinopathy by overcoming the limitations of conventional techniques that depend on manually generated features. In this study, we present a unified ensemble-based deep learning framework for dual-task diabetic retinopathy grading addressing five-class severity classification and binary. Two publicly accessible datasets, IDRiD and APTOS, are fused. Preprocessing techniques including image resizing, Contrast Limited Adaptive Histogram Equalization (CLAHE) and Synthetic Minority Over-sampling Technique (SMOTE) are used to improve image quality and address class imbalance. Data augmentation methods including flipping, zooming and rotation are applied for further diversity. The classification model is constructed by the ensemble of two complementary transfer-learning architectures, ResNet50 and EfficientNetB2. After fine-tuning and hyper-parameter optimization, the models are trained and evaluated. The ensemble model yielded 96.83% test accuracy on multi-class while 99.04% on binary classification surpassing individual backbone models and previous state-of-the-art techniques. Its interpretability is further enhanced through Gradient-weighted Class Activation Mapping (Grad-CAM) visualizations which showed that the proposed framework focused on relevant retinal lesions.

Keywords: Diabetic Retinopathy, Transfer Learning, Data Fusion, Augmentation.

Introduction:

Diabetic retinopathy (DR) is a typical consequence of diabetes mellitus. It is a chronic eye condition [1]. Millions of individuals around the world are impacted by this disease [2]. It damages the retinal blood vessels. The retina is a photosensitive tissue in the eye that transfers visual data to the brain. In working adults, DR has emerged as the most common cause of vision loss and blindness due to the escalating global incidence of diabetes [3]. Notably, in countries with lower or middle incomes and limited access to healthcare facilities, the prevalence of DR is directly linked to the growing rate of diabetes [4]. The occurrence of diabetic retinopathy is primarily associated with prolonged exposure to high levels of blood sugar. Increased blood sugar gradually affects the blood vessels which provide the eye with nutrients and oxygen. This causes structural alterations and impaired functioning. Prolonged inflammation and vascular abnormalities can also lead to the formation of scar tissue, which may distort the structure of the retina and impair visual function. If DR is not treated, it may result in blindness or vision loss [5]. In the early stages of retinopathy, symptoms might not be apparent but can include blurred vision, floaters, difficulty seeing at night, and gradual vision loss.

It is crucial to detect DR early to preserve vision and prevent further damage to the retina. For people with diabetes, routine eye exams are advised to identify damage to the retina early on when options for treatment tend to be beneficial. People with diabetic retinopathy can develop several retinal abnormalities, including hemorrhages (blood leaking into the retina, forming spots or blotches), cotton wool spots (small patches caused by poor blood flow to nerve fibers), exudates (yellowish deposits from leaking fats and proteins), microaneurysms (tiny bulges in retinal blood vessels) and neovascularization (the growth of abnormal blood vessels) [6]. These distinctive features help doctors and healthcare workers to diagnose DR during the examination of fundus images [7].

Diabetic retinopathy progresses through two primary stages: Non-Proliferative Diabetic Retinopathy (NPDR) and Proliferative Diabetic Retinopathy (PDR). The NPDR stage marks the initial phase of the disease, distinguished by the presence of hemorrhages, lipid exudates and microaneurysms in the retina [8]. On the contrary, PDR is an advanced form of the condition that causes the development of abnormal retinal blood vessels and significant impairment in vision. To protect eyesight and stop the growth of diabetic retinopathy, early identification and treatment are essential [9]. Treatment options may include medication, surgery, lifestyle modifications, or laser therapy [10].

Deep Learning (DL), Transfer Learning (TL) and Artificial Intelligence (AI) are important fields in the detection and treatment of diabetic retinopathy. DL being a branch of AI makes use of artificial neural networks which take vast amounts of data and classify or predict things without the need for programming explicitly [11]. Transferring information from one activity or field to another similar activity is called transfer learning [12]. The primary method utilized for detecting diabetic retinopathy is through a retinal eye exam. This procedure entails an eye specialist examining the back of a patient's eye by peering through the pupil. During this examination, the specialist seeks out abnormalities such as micro aneurysms, hemorrhages etc. Furthermore, certain detection techniques utilize fundus photography to capture an image of the patient's retina, enabling the eye specialist to perform a similar assessment by analyzing the retinal image displayed on a computer screen. However, a shortage of ophthalmologists poses a significant challenge, potentially resulting in a substantial number of undiagnosed patients.

Traditionally, manual processing of images and exhausting procedures like manual feature extraction have been used for the diagnosis and treatment of diabetic retinopathy. These methods were difficult, susceptible to errors, and constrained by the practitioners' level of experience. However, DL and TL algorithms offer a more efficient and objective approach

to analyzing retinal images, enabling automated detection of distinctive features associated with diabetic retinopathy [13]. These advanced techniques allow large datasets containing fundus images to train deep neural networks, enabling them to assess the severity of DR and estimate the risk of disease progression [14]. These cutting-edge methods also enable the development of more accurate and dependable DR-detecting frameworks reducing the requirement for human involvement [15].

Recent studies have explored ensemble learning and transfer learning to improve the accuracy and reliability of DR grading. Yan et al. employed transfer learning with enhanced feature extraction and attention mechanisms for DR classification [16]. Darwish et al. proposed an ensemble-based DR classification approach that integrates an adaptive optimization algorithm for feature selection to identify discriminative retinal image features [17]. [18] an explainable deep learning framework for five-class DR grading that integrates Explainable Artificial Intelligence (XAI) techniques like Grad-CAM and SHAP. Similarly, many other studies have made use of deep and transfer learning to classify diabetic retinopathy. XAI techniques are also being employed to enhance model transparency by illustrating the key image features influencing classification decisions.

Research Objectives:

The primary objective of this study is to develop and evaluate a transfer learning-based ensemble framework for automated DR classification using retinal fundus images. The study addresses the following objectives:

Binary Classification: To distinguish between retinal images with diabetic retinopathy and those without DR.

Five-Class Classification: To classify retinal images into five DR severity categories: No DR, mild NPDR, moderate NPDR, severe NPDR and proliferative diabetic retinopathy (PDR).

Performance Evaluation: To assess the classification performance of the proposed framework using appropriate evaluation metrics for both binary and five-class classification.

Novelty Statement:

The novelty of this study lies in integrating transfer learning, cross-dataset fusion and ensemble learning within a unified framework for both binary and five-class DR classification. The proposed approach combines two fine-tuned pre-trained models and evaluates the ensemble against its individual components under consistent experimental conditions. Grad-CAM is further incorporated to examine the image regions contributing to the model's predictions. This study therefore focuses on the combined framework and its evaluation across both classification tasks rather than introducing the individual techniques as novel.

Research Contributions:

A transfer learning-based ensemble approach is developed in this study to categorize diabetic retinopathy into two classes (binary) and five classes (multi-class). We fine-tuned two pre-trained models and after training them, we combined their outputs to build the ensemble model for improved performance. The motivation behind our approach arises from the shortcomings of current approaches which frequently depend on handcrafted characteristics leading to reduced flexibility and classification accuracy. Furthermore, the generalizability of many deep learning models is limited because they are trained solely on a single dataset from one source. Some current approaches also struggle to address class imbalance effectively. Our methodology tackles these challenges by offering a more precise and automatic approach for grading the severity of DR that aims to be more adaptable and robust across varying data conditions. This work offers the following significant contributions:

A dual-task diabetic retinopathy grading framework is proposed that is capable of performing both fine-grained multi-class severity grading and binary diabetic retinopathy presence detection (DR vs. No DR) using a unified architecture.

This research uses a cross-dataset fusion strategy to merge two distinct DR Datasets. Achieving optimal accuracy on fused datasets becomes challenging as there are differences in data distribution among DR Datasets such as variations in imaging techniques, resolutions, and quality. However, we successfully merged two distinct datasets enhancing training data diversity and robustness. This approach overcomes the common limitation of deep learning models trained on a single source.

To enhance the images' overall contrast and highlight fine details of the retinal characteristics, we employed CLAHE (Contrast Limited Adaptive Histogram Equalization) and to rectify the class imbalance and ensure fair representation, we implemented the Synthetic Minority Over Sampling Technique (SMOTE). It was used to establish a well-balanced dataset ensuring a more objective model evaluation.

We integrated two pre-trained models and constructed an ensemble model leveraging their complementary strengths to achieve higher accuracy in diabetic retinopathy classification tasks.

A comparative ensemble evaluation demonstrating that averaging complementary transfer-learning models yields consistent performance gains over individual networks under identical preprocessing and training conditions.

The study incorporates explainable AI (XAI) technique known as Grad-CAM (Gradient-weighted Class Activation Mapping) to enhance model interpretability by visually highlighting retinal regions that are clinically and pathologically relevant to the model's predictions.

Research Organization:

This study is structured into the following sections: In "Related Work" section, a summary of previous studies is presented, along with the methods and various classifiers that were employed. "Methodology" section provides details on the proposed pipeline of the research. In addition to comparing this methodology with earlier studies, the outcomes of the proposed approach are discussed in "Results and Discussion" section utilizing a range of performance evaluation measures. The "Conclusion" section summarizes the key findings of the study.

Related Work:

Many studies have focused on testing a diverse set of methodologies, algorithms, and methods aimed at achieving the goals of detection of DR severity and its early identification. Their dedication has led to significant advancements and great enhancements in the detection and understanding of this disease.

Researchers of [19] have proposed a multi-classification framework for classifying DR through the concepts of transfer learning. The dataset used in their research is APTOS 2019 Blindness Detection. For pre-processing, researchers resized the images to 512 x 512 pixels and then normalized the images by dividing each pixel by 255. Over-sampling was done to take care of class imbalance issues. Techniques like flipping, scaling etc. were applied for data augmentation. Six pre-trained convolutional neural networks (CNN) were utilized in the research. Out of all these models, an ensemble of the top 3 best models was built using the simple averaging technique. The ensemble model showed the best results achieving 82.4% QWK.

[20] have presented an ensemble learning-based framework for severity classification of diabetic retinopathy. They have worked with ensemble-based method using decision tree and gray-level intensity algorithm. The dataset used in their study is APTOS 2019. Pre-processing techniques followed by feature extraction and selection techniques were employed prior to classification. Images were resized first and then augmented to deal with class imbalance problems. Images were also sharpened and contrast was enhanced. The ensemble

algorithm used for classification is XGBoost. With accuracy hitting 94.20%, the model also attained 93.51% F-measure and 92.69% recall.

In [21], the researchers developed a TL-based binary classification framework for the classification of DR. Deep learning is combined with PySpark, a Big Data framework, for dataset processing. A retinopathy dataset is used in the research, which is the Indian Diabetic Retinopathy image Dataset. Firstly, this dataset was cleaned as a pre-processing step by removing dark images. Then it is cropped to remove the unwanted spaces and resized. An LR classifier was used for classification, while DL Pipelines were applied on Apache Spark. The data was divided in an 80:20 ratio. Results indicate that, among the three TL models used, InceptionV3 showed the best performance, achieving 95% accuracy, an AUC of 94.98%, and an F1-score of 95%.

The study in [22] proposed a deep learning-based binary classification system for DR classification. Convolutional neural networks (CNNs) were employed to categorize the fundus images into two groups: those with no DR and those with DR. APTOS 2019 is the dataset used in this research. Images underwent a pre-processing phase where they were cropped and resized. Then CLAHE was utilized to improve image contrast. Researchers have developed a CNN model for classification containing convolutional, pooling, and fully-connected layers. Their deep learning model achieved a testing accuracy of 91%. The model has 86% sensitivity and 96% specificity.

The work in [23] develops a binary classification system employing ResNet-50 to grade diabetic retinopathy. Authors have made use of transfer learning and modified the ResNet50 model in a way to boost the prediction accuracy of DR. A Kaggle dataset with 35,126 retinal images was employed for classification. Images were cropped and the unwanted border was removed. Then Histogram Equalization was applied. ResNet-50 was implemented to perform the classification task. An adaptive learning rate was set for the model and regularization was employed. A comparison was made between ResNet-50 and other pre-trained models like Xception, AlexNet, and VGGNet-16, and the results showed the performance of the revised ResNet-50 stood better than the others with an accuracy of 74.32%.

In [24], a binary classification approach leveraging Convolutional Neural Networks was introduced for DR detection. They have implemented VGG (visual geometry group) architecture. It has 4 blocks where a convolutional layer is present in each block, followed by max pooling layers. In the layers, ReLU was the chosen activation function. For binary classification, the sigmoid activation function was utilized in the output layer. Adam served as the optimizer, while binary cross entropy was employed as the loss function. The dataset was extracted from diabetic-retinopathy-classified through Kaggle. Images were augmented to add diversity and improve the process of generalization. 95.5% accuracy was achieved by the model introduced in this study.

To classify DR into five categories, the authors of [25] leveraged 3 hybrid models. (Model f, Model a and Model c). Five CNN models— EfficientNetB4, InceptionResNetV2, NASNetLarge, EfficientNetB5, and Xception—were utilized to construct these hybrid models. To train the base models, two loss functions were used, and their outputs were later fed into the hybrid models. The study employed three datasets for experimentation: DeepDR, EyePACS and APTOS. Model training was preceded and accompanied by pre-processing, with image enhancement being the primary focus. Hybrid-c model exhibited the highest efficiency among the hybrid models assessed in this research, reaching an accuracy of 86.34%.

A novel binary classification framework is proposed in [26], where researchers incorporated Convolutional Neural Networks (CNNs) for the detection of diabetic retinopathy. Two different datasets were incorporated in this study: Messidor, which consists of 1,200 images, and DR1, which includes 1,014 images. Three distinctive methods have been applied by the authors to carry out classification. The models are pre-trained on the ImageNet

dataset. Images were cropped first and later they were resized. Then classification was carried out by 3 different experiments i.e. 1) Full fine-tuning, 2) Layer-wise fine-tuning and 3) Pre-trained convolutional neural networks. The highest accuracy was achieved using Layer-wise fine-tuning. On the Messidor dataset, the model achieved 92.01% accuracy, whereas on the DR1 dataset, it reached 94.52%.

Authors in [27] addressed the growing public health challenge of type 2 diabetes mellitus by developing interpretable machine learning models for early risk prediction. The study focused on an underrepresented population using data from the Dena Cohort in Iran. The dataset included 3,203 adults aged between 35 and 70 years. Data preprocessing involved outlier removal, median and mode imputation, and feature selection using LightGBM. Class imbalance was handled using the SMOTE. XGBoost and CatBoost models were trained with extensive hyperparameter tuning and evaluated through 10-fold cross-validation. The XGBoost model achieved the best performance, reporting 96.07% accuracy and an AUC of 99.29%, while SHAP analysis revealed key clinical and lifestyle predictors and improved model interpretability.

In [28], authors proposed an automated deep learning framework that combines customized ResNet-50 and EfficientNetB0 architectures for diabetic retinopathy detection. The model addressed class imbalance by applying data augmentation along with SMOTE on the training set. Feature extraction was performed using fine-tuned ResNet-50 and EfficientNetB0 networks with LeakyReLU activations and global average pooling. These fused representations were passed to four individually tuned classifiers for both binary and multi-class classification tasks. The experiments were conducted on APTOS 2019 dataset. The proposed approach achieved 98.5% accuracy for binary classification and 92.73% accuracy for multi-class classification.

Authors in [29] investigated automated diabetic retinopathy detection by evaluating five severity classes using deep learning models. The study initially analyzed 19 state-of-the-art architectures on the APTOS 2019 Blindness Detection Challenge dataset. However, most existing models achieved validation accuracies below 80%, indicating limited generalization on the selected data. To overcome this limitation, the authors proposed DR-RetinaNet, a customized framework built upon the DenseNet201 backbone. The model incorporates selective freezing of early convolutional layers and batch normalization layers, while deeper layers are fine-tuned to learn DR-specific features more effectively. As a result, DR-RetinaNet achieved 99% training accuracy and 94% validation accuracy, outperforming competing methods.

All of the studies reviewed have used DL approaches for detecting and classifying diabetic retinopathy utilizing numerous pre-processing techniques and classification methodologies on different DR Datasets. These approaches involved transfer learning, hybrid models and convolutional neural networks yet they continued to face challenges related to class imbalance, generalization and accuracy. In comparison, we introduce a solution that enhances both efficiency and robustness. Many previous techniques focused solely on either multi-class or binary classification but our approach successfully handles both tasks while ensuring greater flexibility and versatility. Rather than using a single dataset or various datasets individually as done in most studies, we fused two diabetic retinopathy datasets to make the model more resilient to diverse data conditions. To improve visual clarity and boost model adaptability, our framework incorporates advanced preprocessing methods like CLAHE. Our approach integrates SMOTE to adjust class distribution and minimize bias in the model's predictions. Our methodology incorporates transfer learning. A major highlight of this study is the adoption of an ensemble approach, where two transfer learning models are combined to improve the efficiency and accuracy of DR detection.

Methodology:

An ensemble classification framework leveraging transfer learning is proposed in this research for diabetic retinopathy detection across five and two classes. For the purpose of classifying this disease, two popular datasets are merged together and an ensemble of two transfer learning models is built. TL models are extensively applied and acknowledged in visual data analysis and healthcare diagnostics [30][31]. We explored both binary and multi-class classification, treating them as distinct tasks within the same framework. Using the same pre-trained and ensemble model, we trained and evaluated each task independently which allowed us to harness the strength of the models for both classification needs. This approach enabled us to achieve high accuracy in both tasks while maintaining efficiency within a unified model. Figure 1 shows the proposed ensemble classification framework. It presents each step of the methodology in an orderly manner. An extensive description of the dataset used in this work is given below along with the corresponding procedures that were carried out for the classification of DR.

Dataset Description:

Two distinct datasets are utilized in this research for classification of DR. First one is IDRiD (Indian Diabetic Retinopathy Image Dataset) and the second one is APTOS (Asia Pacific Tele-Ophthalmology Society) 2019 Blindness Detection. A detailed description of each dataset is given below.

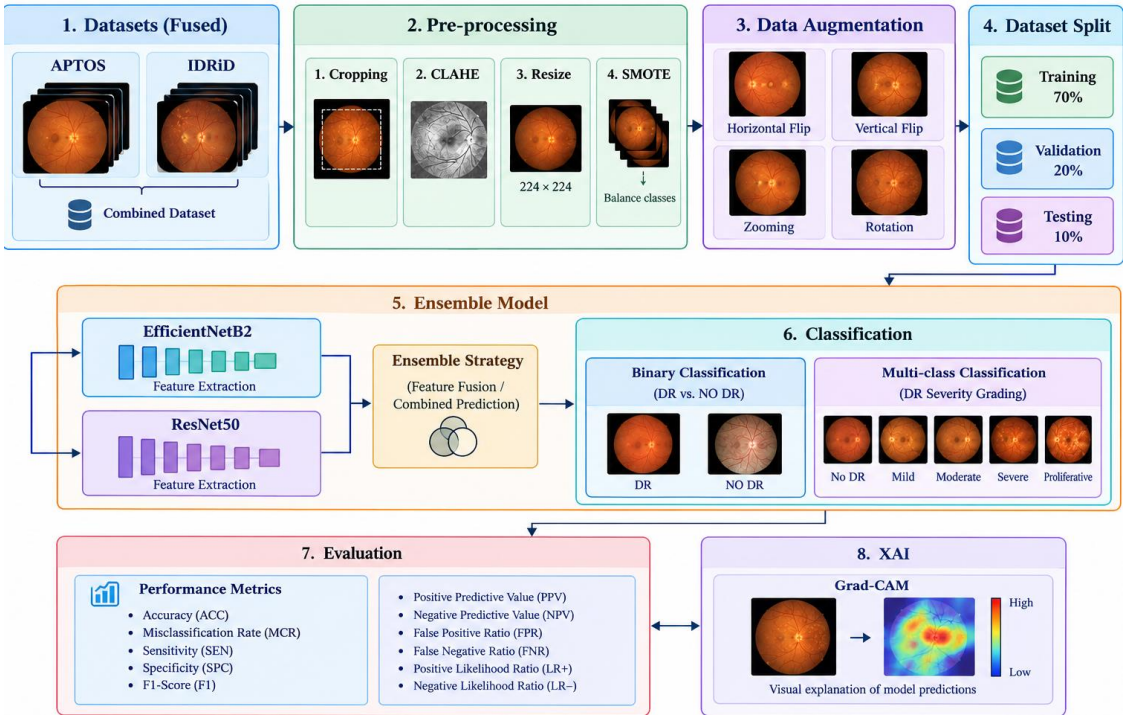


Figure 1. Proposed ensemble framework.

Indian Diabetic Retinopathy Image Dataset marks an important milestone as an early collection focusing on diabetic retinopathy cases in Indian individuals [32]. It is distinguished by its inclusion of both normal retinal features and diabetic retinopathy lesions, making it unique among datasets. Each image in this dataset is carefully labeled to indicate DR severity levels and diabetic macular edema. Given its capability to detect diabetic retinopathy at an early stage, IDRiD serves as an excellent resource for the development and evaluation of image processing techniques. Organized into three core sections, the dataset covers Segmentation, Localization and Disease Grading. The central aim of our research is disease grading, which involves determining the diabetic retinopathy severity. A collection of 516 retinal images are provided in the dataset. It contains 413 training images with 103 testing images, each with

corresponding labels. Just like other diabetic retinopathy datasets, IDRiD classifies images into five grades: Grade 0 (No DR), while Grades 1 to 4 represent mild, moderate, severe, and proliferative stages respectively. Examples from all the grades of this dataset are depicted in Figure 2.

The APTOS 2019 Blindness Detection dataset can be accessed on Kaggle [33]. Captured at an Indian Eye Hospital named "Aravind", the image count of the dataset is 3662. The images were captured in a variety of environments. In remote regions where access to clinical evaluation is limited, the hospital's contributions in providing diagnosis and treatment are particularly valuable. On a scale ranging from 0 to 4, hospital clinicians have classified the degree of DR in every fundus image. Each grade represents a different level of disease severity. In this grading system, 0 signifies No DR, 1 indicates Mild, 2 represents Moderate, 3 stands for Severe, and 4 denotes Proliferative DR. Figure 2 showcases samples from each severity grade, providing visual insight into the different stages of DR.

The image distribution across each class is unbalanced in both the IDRiD and APTOS datasets. Both datasets are merged in this research to increase the overall size of the dataset. After combining, the total number of fundus images increased to 4178. All five grades of the fused dataset were used for the multi-classification task. The dataset was split into two classes for binary classification labeled as grade 0 and grade 1. Here Grade 0 represents all those normal eye images with no signs of diabetic retinopathy while grades 1 to 4 were merged into a single category, Grade 1, representing the presence of diabetic retinopathy to support the 2-stage classification task. A summary of image distribution across grades and the overall image count for IDRiD, APTOS, and the merged dataset is presented in Table 1. The image distribution in the combined dataset across various grades is displayed in Table 2, detailing counts for both multi-class and binary classification tasks.

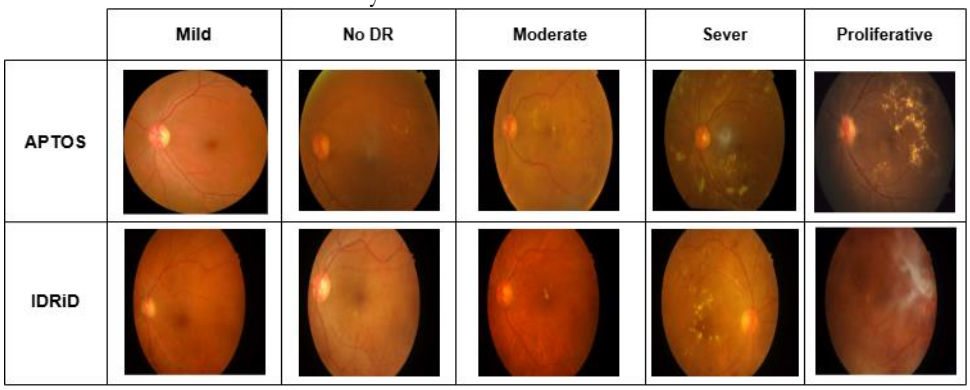


Figure 2. Example fundus images corresponding to different grades in the APTOS and IDRiD datasets [33][32].

Table 1. Grade-wise image count for each dataset

Datasets	No DR	Mild	Moderate	Severe	Proliferative DR	Total
IDRiD	168	25	168	93	62	516
APTOS	1805	370	999	193	295	3662
Merged Dataset	1973	395	1167	286	357	4178

Table 2. Distribution of images by grades in the merged dataset for multi-class and binary classification task.

Grade	Multi-class Classification	Binary Classification
0	1973	1973
1	395	2205 (images of grades 1 to 4 combined)
2	1167	
3	286	
4	357	

Data Preprocessing:

To ensure that deep learning projects are successful, data preprocessing is essential for providing high-quality data that enables accurate and reliable model predictions [34]. Publicly available datasets for diabetic retinopathy detection often contain various imperfections such as noise, inconsistencies in resolution, and uneven contrast. Data preprocessing helps in addressing these challenges and preparing the data for effective model training. Techniques such as denoising filters, contrast enhancement, and standardization of image resolutions are commonly applied.

Preprocessing for this research follows a four-step approach: 1) Cropping, 2) CLAHE, 3) Resizing and 4) SMOTE. All pre-processing steps are the same for both multi-class and binary classification except for the application of SMOTE. It was not used for binary classification. We found that there wasn't a significant difference in sample sizes between the classes when separating images with and without DR. Initially, the dataset was thoroughly analyzed considering the fact that images present within the dataset are from diverse clinics and hospitals and captured using various cameras and different angles. Since two different datasets are fused together in this research, the images from both datasets are different from each other in terms of background, contrast and resolution. A few images have noise and extraneous black regions, while some exhibit dullness and poor contrast which conceals the visibility and fine details of retinal features like tiny blood vessels and micro aneurysms.

To address these issues, the first step is cropping. Some of the dataset's images have an unwanted black background. Giving a model such a dataset with a wide range of backgrounds could cause problems with performance and generalization. This will reduce our model's accuracy. The images undergo cropping from the point where the eye is visible to remove unwanted black areas and achieve the required dimensions. This step helped get rid of the extra unwanted regions in the images, bringing them all to have the same background.

In the second step, we applied Contrast Limited Adaptive Histogram Equalization. This approach improved the contrast, enhanced the visibility of details in images and minimized the amplification of noise. This step solved the problem of dullness and low contrast that was analyzed earlier in our merged dataset.

Then for the third step, these enhanced images undergo resizing to ensure that every image has the same dimensions of 150 x 150 x 3. The fused dataset has images with different resolutions and dimensions due to differences in the original sources. Standardizing the image resolution contributed to a consistent data format which allowed the models to generalize well across diverse image sources. The first 3 key steps in the preprocessing sequence are illustrated in Figure 3.



Figure 3. A depiction of the initial three phases of pre-processing.

The final step of pre-processing i.e. step 4 is the use of Synthetic Minority Over-sampling Technique to balance the dataset. It was only used for multi-class classification as image distribution across each grade is unbalanced. SMOTE is an over-sampling technique that creates artificial instances for the underrepresented classes in order to even out the distribution of classes. It generates new, artificial samples of the minority classes based on the existing instances. The fused dataset's unbalanced image distribution between classes is shown in Table 1. As a result, bias can emerge in our models when they consistently favor the class with the highest sample count due to its dominance in the dataset. Mitigating the imbalance among classes is crucial for achieving optimal model performance.

The goal is to amplify the sample count in underrepresented classes without modifying the majority class. Since Grade 0 has the highest image count in the fused dataset, it will be kept unchanged. We only augmented the sample size of the remaining grades, bringing them closer to the image count in Grade 0. Table 3 presents the revised image distribution across grades following SMOTE implementation. To address class imbalance in our dataset, SMOTE was applied directly to the pixel-level representation of the retinal images. Each image was flattened into a one-dimensional pixel vector and provided to SMOTE for synthetic sample generation. Following this process, the generated samples were reshaped to the required image dimensions for subsequent model training. The SMOTE procedure was configured with k-neighbors = 5, while the remaining parameters were maintained at their default settings.

Table 3. New Image count per grade after the application of SMOTE.

Grade	Initial Image Count	Post-SMOTE Image Count
No DR	1973	1973
Mild	395	1900
Moderate	1167	1800
Severe	286	1750
Proliferative DR	357	1750
Total	**4178**	**9173**

Data Augmentation:

Once preprocessing was completed, the images underwent augmentation using various methods. This augmentation aims to expand the dataset size, enabling the model to detect a greater range of characteristics and patterns in the images. By doing so, the models can achieve improved generalization capabilities.

Data augmentation is a well-established strategy for expanding datasets by modifying existing samples in multiple ways [35]. These alterations can involve rotating, resizing, flipping, translating, shearing, zooming, and other adjustments. By generating new samples from existing ones through augmentation, the model becomes more robust to handle unseen data during testing. Augmentation has many types and forms but for our proposed methodology, we have worked with geometric augmentation techniques. Geometric augmentation techniques encompass a set of image transformations that modify the positioning and structure of image elements such as scale, rotation etc.

For this research, we utilized vertical and horizontal flipping, along with zooming and rotation. The application of these techniques to the dataset is done randomly. The augmentation process is designed to increase the dataset size by generating twice as many images. After using SMOTE, the new total of images for the multi-class increased to 9173. Consequently, after applying augmentation, it became 18,346. As for binary classification, the application of augmentation resulted in a dataset size of 8,356 images. A representation of the augmentation process applied in this study is illustrated in Figure 4. Table 4 presents the used augmentation techniques.

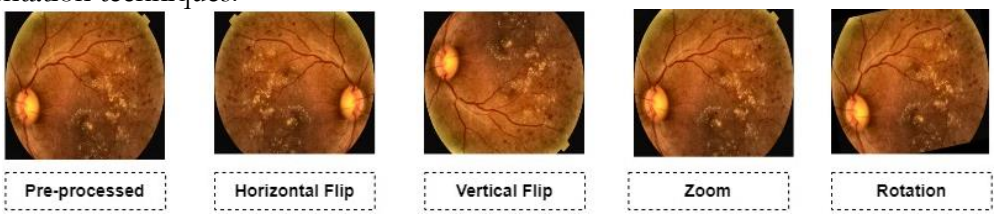


Figure 4. Illustration of the geometric techniques utilized in the augmentation process.

Table 4. Utilized Augmentation Approaches

Techniques	Values
Flip	Vertical
Flip	Horizontal

Rotation Range	10
Zoom Range	0.2

Dataset Split:

After preparing the dataset, it was partitioned into testing, validation, and training before the model training procedure. We separated our dataset into three sections: 70% for the training, 20% for the validation, and 10% for the testing. Once the models were trained using the training set, we used the testing set to assess their performance. Throughout the training phase, the models' performance was evaluated using the validation set.

Transfer Learning Models:

Convolutional Neural Networks are extensively trained using the ImageNet dataset as it has a varied and broad range of images [36]. ImageNet is a huge collection of millions of labeled images, covering a wide range of categories. Several well-known models, such as AlexNet, DenseNet, Xception, and EfficientNet, are trained on ImageNet. Excelling in the ImageNet challenge, these models have proven their superior capability in precise object recognition.

EfficientNetB2 is a highly optimized convolutional neural network that builds upon the principles of compound scaling. This architecture balances depth, width, and resolution to improve both accuracy and efficiency. EfficientNetB2 utilizes layer-wise separable filters which split the convolution operation into two steps: depth-wise and point-wise convolutions. This approach preserves excellent performance while significantly decreasing the overall parameter count. The model's architecture includes various stages with increasing depth, each contributing to learning more complex features, which makes it particularly effective for image classification tasks. Additionally, its architecture enables efficient processing, making it ideal for use in environments with limited resources.

ResNet50, on the other hand, is renowned for its innovative use of residual connections that help minimize the disappearing gradient issue that deeper networks frequently face. The architecture contains fifty layers comprising convolutional layers with ReLU as an activation function, batch normalization and many other layers. ResNet50's skip connections are its primary characteristic; they enable gradients to pass across the network without degrading, making it easier to train extremely deep models. This design enables ResNet50 to effectively capture intricate features within the images, leading to improved classification accuracy.

For this study, we have used 2 transfer learning-based models that are described above. Before training, these models were adjusted by making specific modifications to the initial and final layers. Adjustments were made to the input layer to accommodate images of size $150 \times 150 \times 3$ ensuring that the models could effectively process our dataset. Then after the base transfer learning layer, we incorporated multiple layers in between. We added a Global Average Pooling (GAP) layer to minimize the complexity of the feature maps without losing any significant data. This step helps minimize overfitting by averaging the spatial dimensions, leading to a more generalized model. Dropout layers were also included to avoid overfitting. These layers helped the model acquire more resilient features by arbitrarily adjusting a portion of the units used for input to zero. After that, we added a flattened layer to prepare the data to be used in fully connected (FC) layers by transforming the two-dimensional feature maps into a single-dimensional vector. The FC layers then learned high-level features by connecting the outputs of the previous layer to produce more complex representations. Non-linearity was introduced in the hidden layers using the ReLU activation function, enabling better extraction of fine image details. Next, an output layer was added. For multi-class classification, we applied the softmax activation function. On the contrary, we used the sigmoid activation function for binary classification. The refined architectures of these models, after fine-tuning, are shown in Figure 5.

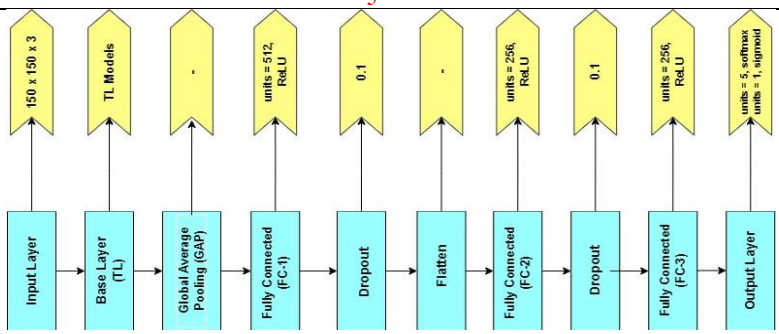


Figure. 5. Architectural breakdown of the fine-tuned models.

Ensemble Model:

Ensemble learning integrates several models to enhance the predictive performance beyond that of a single model. Rather than depending on a single model, ensemble methods merge predictions from different models, helping to minimize errors and boost overall performance. This approach provides several benefits over individual models such as reducing overfitting, increasing stability, and enhancing predictive capabilities. In the context of transfer learning, ensemble techniques merge outputs from various pre-trained models to strengthen classification accuracy [37]. Ensemble learning can be implemented using various techniques with the most commonly used approaches including averaging, bagging, stacking and boosting.

The ensemble technique used in this study is Ensemble Averaging, where we combine predictions from multiple models to enhance overall accuracy. We created an ensemble model by combining the predictions of EfficientNetB2 and ResNet50. In this approach, the predicted probabilities for each class from both models were averaged. The final classification decision for each sample was based on the class with the highest mean probability. Each model assigned a probability score to a given class. The final prediction was determined by averaging these probabilities across all models. We built a strong ensemble model by choosing the class with the highest average probability. This approach combines the advantages of both architectures, enhancing classification performance in our study.

Performance Evaluation:

Performance evaluation metrics were used to assess the classification performance of the proposed models on the training and test datasets [38]. These metrics include misclassification rate (MCR), accuracy (ACC), specificity (SPC), sensitivity (SEN), F1-score (F1), positive predictive value (PPV), negative predictive value (NPV), false-positive rate (FPR), false-negative rate (FNR) and positive and negative likelihood ratios (LR+ and LR−).

$$Misclassification\ Rate\ (MCR) = \frac{FP+FN}{TP+TN+FP+FN} \quad (1)$$

$$Accuracy\ (ACC) = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$Specificity\ (SPC) = \frac{TN}{TN+FP} \quad (3)$$

$$Sensitivity\ (SEN) = \frac{TP}{TP+FN} \quad (4)$$

$$F1-Score\ (F1) = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

$$Positive\ Prediction\ Value\ (PPV) = \frac{TP}{TP+FP} \quad (6)$$

$$Negative\ Prediction\ Value\ (NPV) = \frac{TN}{TN+FN} \quad (7)$$

$$False\ Positive\ Ratio\ (FPR) = 1 - \text{Specificity} \quad (8)$$

$$False\ Negative\ Ratio\ (FNR) = 1 - Sensitivity \quad (9)$$

$$Positive\ Likelihood\ Ratio(LR+) = \frac{\text{Sensitivity}}{(1 - Specificity)} \quad (10)$$

$$Negative\ Likelihood\ Ratio\ (LR-) = \frac{(1 - \text{Sensitivity})}{Specificity} \quad (11)$$

In binary classification, the evaluation is based on four outcomes: true positive (TP), true negative (TN), false positive (FP) and false negative (FN). TP represents a positive sample correctly classified as positive whereas TN represents a negative sample correctly classified as negative. FP refers to a negative sample incorrectly classified as positive while FN represents a positive sample incorrectly classified as negative. Based on these outcomes, the performance metrics are calculated using Equations (1)–(11).

For five-class DR classification, the one-versus-rest (OvR) strategy was used to calculate class-wise metrics. Under this strategy, each DR severity class was treated as the positive class while the remaining four classes were grouped as the negative class. TP, TN, FP and FN were determined separately for each class using this approach. The relevant class-wise metrics were then macro-averaged across the five classes to obtain the corresponding overall results. Accuracy was calculated from the overall classification outcomes while the remaining metrics were calculated according to their respective definitions and the adopted averaging strategy.

In Equations (1)–(11), TP, TN, FP and FN denote true positives, true negatives, false positives and false negatives respectively. Precision and recall in Equation (5) correspond to PPV and SEN respectively. MCR represents the proportion of incorrectly classified samples whereas ACC represents the proportion of correctly classified samples. SPC measures the ability to correctly identify negative samples and SEN measures the ability to correctly identify positive samples. PPV indicates the proportion of predicted positive samples that are truly positive while NPV indicates the proportion of predicted negative samples that are truly negative. FPR and FNR represent the proportions of negative and positive samples incorrectly classified respectively. LR+ measures how much a positive prediction changes the odds in favor of the positive class whereas LR− measures how much a negative prediction changes those odds.

Experimental Setup:

Python programming language was used for both model training and testing within Kaggle Notebooks. These notebooks offer RAM of 29GB, 15GB of GPU and 73GB of storage space. The models utilized in this study are EfficientNetB2 and ResNet50. Various training parameters were experimented with for model training including optimizers, batch size and more to improve the training process and analyze which parameters best suit the process.

We experimented with learning rates of 0.0001, 0.001, and 0.01, determining that a rate of 0.001 struck the optimal balance between fast convergence and training stability. Higher learning rates resulted in instability, while lower ones significantly slowed the model's learning process. For batch size, we tested values of 16, 32, and 64, and ultimately chose 32, as it offered the best compromise between memory efficiency and model generalization. Smaller batch sizes increased training time, while larger ones risked overfitting. Initially, we trained the model for up to 100 epochs but decided on 50, as the model showed convergence by this point. When evaluating optimizers, we tested Stochastic Gradient Descent (SGD) and RMSProp, eventually selecting RMSProp because its adaptive learning rate provided more consistent

performance across varying learning conditions, resulting in faster and more reliable convergence. To avoid overfitting, this study implements regularization techniques: ReduceLROnPlateau and EarlyStopping methods. These techniques are aimed at stopping training when the model achieves its optimal performance so that overfitting can be avoided. Training is stopped by EarlyStopping if validation accuracy either stagnates or shows no improvement for 10 consecutive epochs. If the loss does not change for two consecutive epochs, the learning rate is adjusted. The ReduceLROnPlateau technique decreases it by a factor of 0.2 to improve training stability. Table 5 presents the training parameters used in this research.

Table 5. Training hyper-parameters used for both types of classification.

Hyper Parameters	
No. of Epochs	50
Loss Function	Categorical cross-entropy, Binary cross-entropy
Learning Rate	0.001
Input Shape	150 x 150 x 3
Optimizer	RMSProp
EarlyStopping	patience = 10
Batch Size	32
ReduceLROnPlateau	factor = 0.2, patience = 2

Binary and Five-Class Classification Setup:

The proposed framework was designed to support both binary and five-class DR classification using a consistent methodological pipeline. The same model architectures, layer configurations and training parameters, including batch size and number of epochs, were maintained for both tasks. Separate training runs were performed for the two classification tasks. The only task-specific changes were the output layer and loss function: two output units with binary cross-entropy were used for binary classification whereas five output units with categorical cross-entropy were used for five-class classification. SMOTE was not applied to binary classification because the No DR class was classified against the four DR severity classes combined. All other methodological stages including preprocessing, model training, ensemble prediction and performance evaluation followed the same framework.

Results and Discussion:

In automated diabetic retinopathy screening systems, reliable detection of disease presence and accurate differentiation of severity levels are both critical for clinical decision-making. The experimental results discussed in this section provide insights into class-wise error patterns, discriminative capability and the benefits of ensemble learning for both multi-class and binary classification tasks.

Confusion Matrix and Class-Wise Error Analysis:

When evaluated for multi-class diabetic retinopathy grading, the individual and ensemble models demonstrated strong overall performance; however, a closer inspection of the confusion matrices reveals meaningful class-wise error patterns that highlight the inherent challenges of fine-grained DR severity classification. The confusion matrices of training and testing set of each model for severity grading are given in Figure 6.

For EfficientNetB2, the training phase yielded a high accuracy of 99.40%, indicating effective feature learning. The corresponding training confusion matrix in figure 6 shows near-perfect diagonal dominance across all five severity classes with minimal misclassifications. The model yielded 96.18% accuracy on testing set. The confusion matrix for testing set shows that the No DR (Class 0) category is the most accurately classified class, exhibiting the highest number of correct predictions with negligible confusion. In contrast, the Moderate DR (Class 2) category demonstrates the highest level of misclassification. Errors for this class are

predominantly distributed toward adjacent classes, namely Mild DR (Class 1) and Severe DR (Class 3). The class-wise performance of this model on testing set is summarized in Table 6.

Table 6. Class-wise performance of EfficientNetB2 on testing set (multi-class).

Class	EfficientNetB2			
	Precision	**Recall**	**F1-score**	**Support**
0	0.9750	0.9774	0.9762	399
1	0.9840	0.9609	0.9723	384
2	0.9396	0.9201	0.9297	338
3	0.9491	0.9779	0.9633	362
4	0.9579	0.9688	0.9633	352
Macro Avg	0.9611	0.9610	0.9610	1835
Weighted Avg	0.9620	0.9619	0.9618	1835

Similarly, ResNet50 achieved 98.28% training accuracy but showed a larger performance drop during testing, with accuracy decreasing to 92.47%. A similar trend is observed but with increased confusion across intermediate classes. While No DR (Class 0) again remains the best-classified category during testing, ResNet50 exhibits a noticeably higher misclassification rate for Moderate DR (Class 2) and Severe DR (Class 3) compared to EfficientNetB2. The confusion matrix indicates that images from these classes are frequently misidentified as neighboring severity levels, suggesting that ResNet50 is more sensitive to subtle inter-class variations and less robust to changes in lesion appearance and imaging conditions. Table 7 presents the class-wise performance of this model on testing set.

Table 7. Class-wise performance of ResNet50 on testing set (multi-class).

Class	ResNet50			
	Precision	**Recall**	**F1-score**	**Support**
0	0.9539	0.9850	0.9692	399
1	0.9560	0.9063	0.9305	384
2	0.9057	0.8521	0.8780	338
3	0.9339	0.9365	0.9352	362
4	0.8704	0.9347	0.9014	352
Macro Avg	**0.9240**	**0.9229**	**0.9228**	1835
Weighted Avg	**0.9255**	**0.9248**	**0.9246**	1835

The ensemble model, constructed by integrating EfficientNetB2 and ResNet50, mitigated many of these limitations. While EfficientNetB2 excels at capturing fine-grained texture details, ResNet50 contributes robust hierarchical feature representations. The ensemble leverages complementary strengths from both architectures, resulting in improved generalization and reduced class confusion. This is evident in the ensemble's testing confusion matrix, which shows fewer misclassifications across all classes and a stronger diagonal concentration compared to the individual models. Consequently, the ensemble achieved the highest test performance, with 96.83% accuracy, 96.76% sensitivity, and 99.21% specificity. Table 8 shows the class-wise performance of ensemble model on testing set.

Table 8. Class-wise performance of Ensemble Model on testing set (multi-class).

Class	Ensemble			
	Precision	**Recall**	**F1-score**	**Support**
0	0.9753	0.9900	0.9826	399
1	0.9919	0.9583	0.9748	384
2	0.9489	0.9349	0.9419	338
3	0.9649	0.9862	0.9754	362
4	0.9579	0.9688	0.9633	352
Macro Avg	**0.9678**	**0.9676**	**0.9676**	1835

Weighted Avg	**0.9685**	**0.9684**	**0.9683**	1835

In binary diabetic retinopathy (DR vs. No DR) classification, all models achieved high accuracy but showed different error behaviors. Figure 7 represents the confusion matrix of each model for binary classification. EfficientNetB2 reached 99.40% training accuracy and 98.56% test accuracy. It produced very few false positives, but some false negatives remained mainly in images with subtle DR signs or low image contrast. ResNet50 achieved 99.50% training accuracy and 97.49% test accuracy. It showed more misclassifications in both classes. The ensemble model achieved the best performance with 99.48% train and 99.04% test accuracy. It misclassified only four images per class in the test set. This shows its improved robustness and balanced error behavior. These binary misclassification patterns reflect the challenges of multi-class DR grading, where neighboring severity levels share similar visual features and real-world image variability further complicates classification. The class-wise performance on testing set for all the models is provided in Table 9.

Table 9. Class-wise performance of all the models on testing set (Binary classification).

Model	Class	Precision	Recall	F1-score
EfficientNetB2	Class 0	0.9902	0.9807	0.9854
	Class 1	0.9812	0.9905	0.9858
	Macro Avg	0.9857	0.9856	0.9856
	Weighted Avg	0.9857	0.9856	0.9856
ResNet50	Class 0	0.9781	0.9710	0.9745
	Class 1	0.9718	0.9787	0.9752
	Macro Avg	0.9749	0.9748	0.9749
	Weighted Avg	0.9749	0.9749	0.9749
Ensemble	Class 0	0.9903	0.9903	0.9903
	Class 1	0.9905	0.9905	0.9905
	Macro Avg	0.9904	0.9904	0.9904
	Weighted Avg	0.9904	0.9904	0.9904

Training and Validation Performance: The training and validation accuracy and loss curves provide important insights into the convergence behavior, generalization capability and stability of the evaluated models for both multi-class and binary diabetic retinopathy classification tasks. The curves for all models are presented in Figure 8 in graphical form for multi-class classification. The graphs demonstrate that the models were effectively trained which is visible through their minimum validation loss value. In multi-class classification, EfficientNetB2 scored 96.41% validation accuracy and 0.1454 validation loss. ResNet50 recorded 89.67% validation accuracy and 0.4238 validation loss. The ensemble model attained 96.12% validation accuracy while maintaining 0.1257 validation loss. Figure 9 shows similar graphs for binary classification. Here, EfficientNetB2 attained 98.14% validation accuracy and 0.1563 validation loss. ResNet50 achieved 97.23% validation accuracy, recording 0.3592 validation loss. The ensemble model delivered 99.45% validation accuracy while maintaining 0.1145 validation loss. As shown in Table 10, the ensemble model achieves the lowest validation loss and highest validation accuracy for both multi-class and binary classification, indicating improved generalization compared to individual models. It showed the most stable behavior in the multi-class scenario with smoother convergence patterns and lower validation loss compared to individual models.

Table 10. Validation Accuracy and Validation Loss for Multi-Class and Binary Classification

Model	Multi-class		Binary	
	Validation Accuracy (%)	Validation Loss	Validation Accuracy (%)	Validation Loss
EfficientNetB2	96.41	0.1454	98.14	0.1563

ResNet50	89.67	0.4238	97.23	0.3592
Ensemble Model	96.12	0.1257	99.45	0.1145

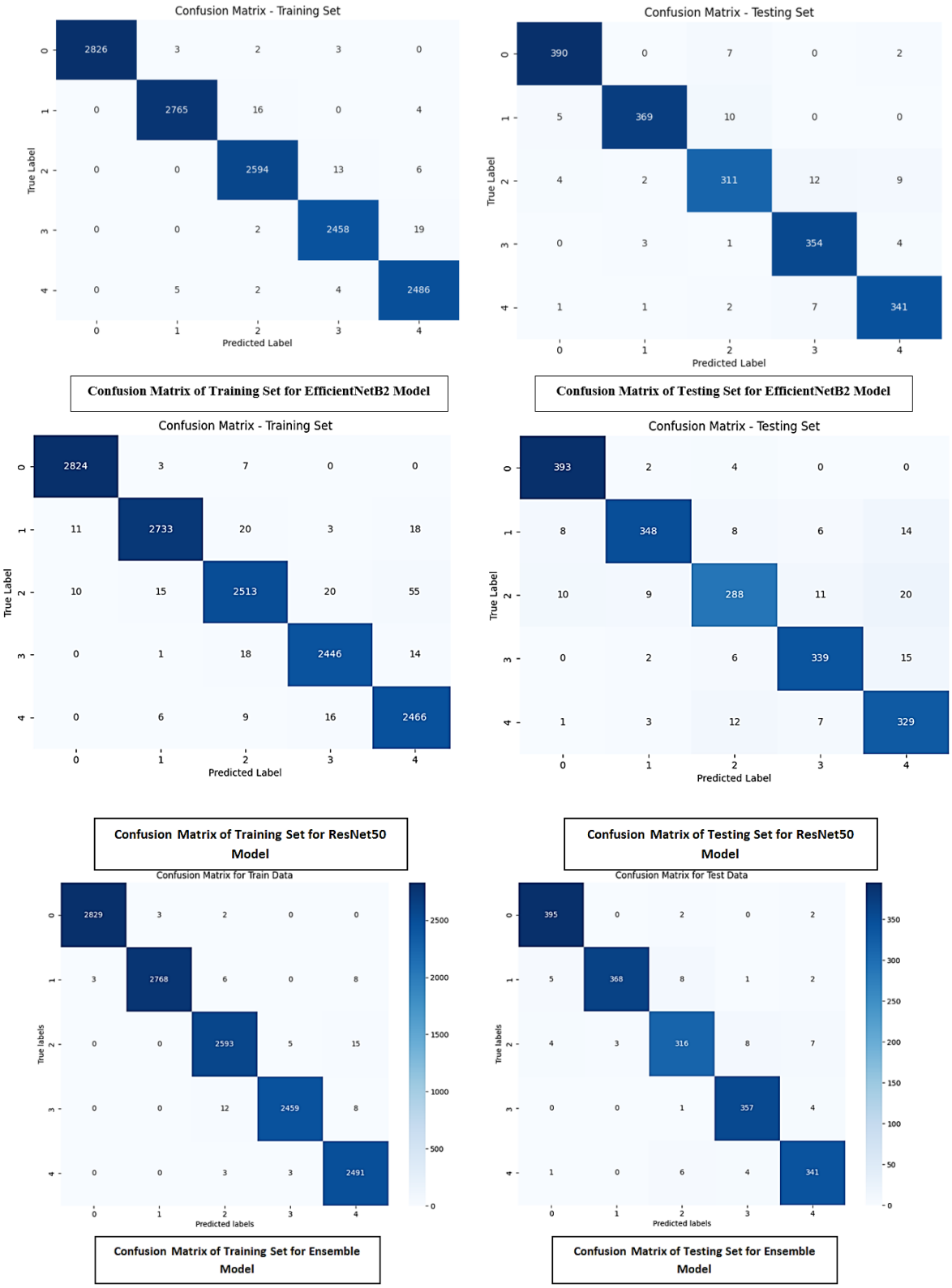
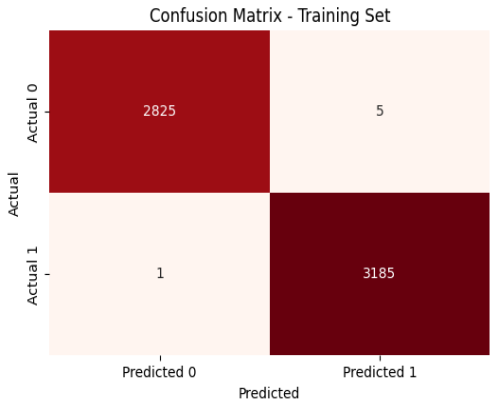
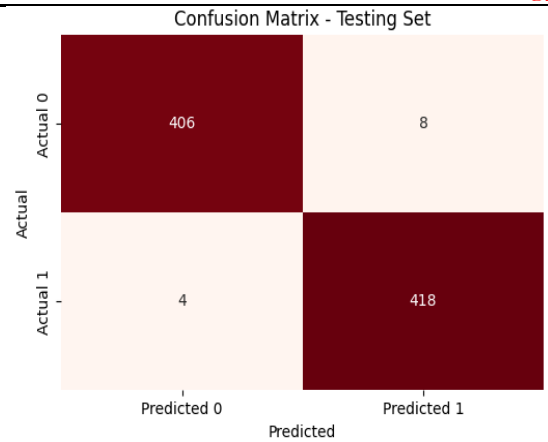


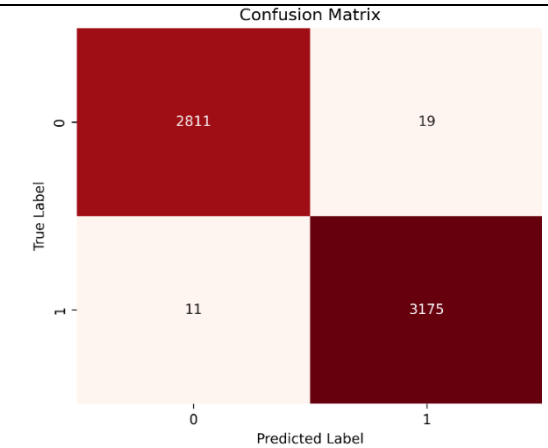
Figure 6. Multi-class Classification: Confusion Matrix for Individual and Ensemble Model



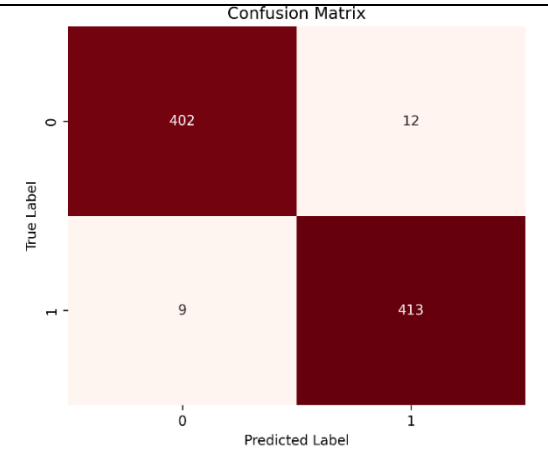
Training Data Confusion Matrix for EfficientNetB2 Model



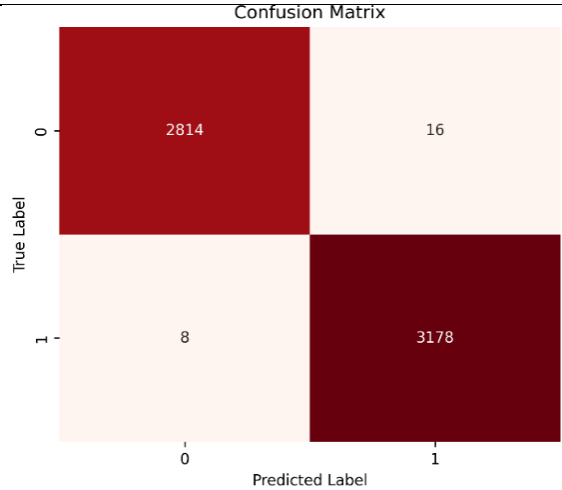
Testing Data Confusion Matrix for EfficientNetB2 Model



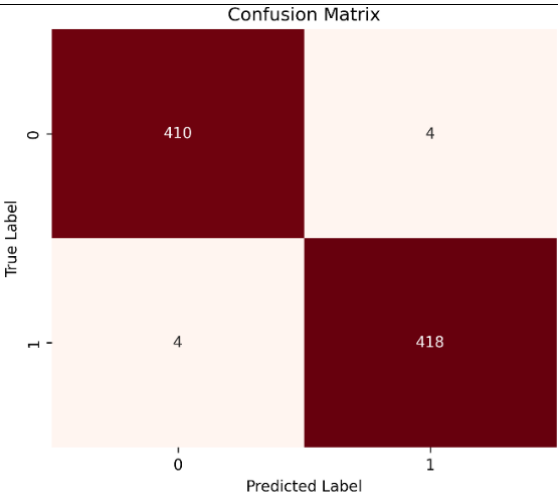
Training Data Confusion Matrix for ResNet50 Model



Testing Data Confusion Matrix for ResNet50 Model

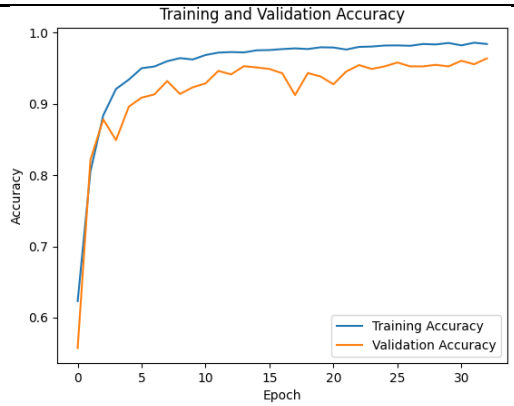


Training Data Confusion Matrix for Ensemble Model

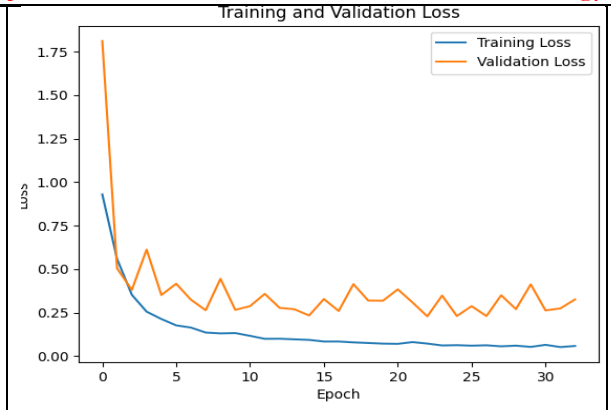


Testing Data Confusion Matrix for Ensemble Model

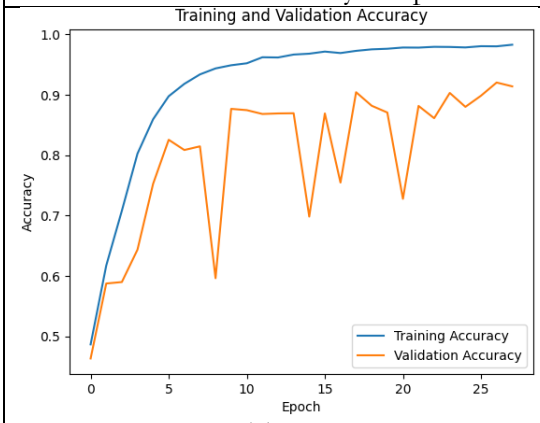
Figure 7. Binary Classification: Confusion Matrix for Individual and Ensemble Models



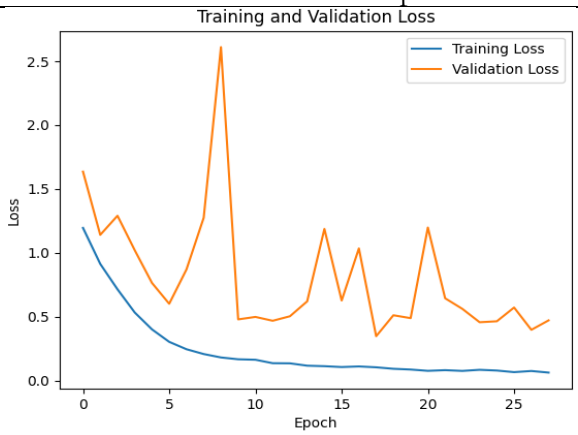
EfficientNetB2 model: Training vs. Validation Accuracy Graph.



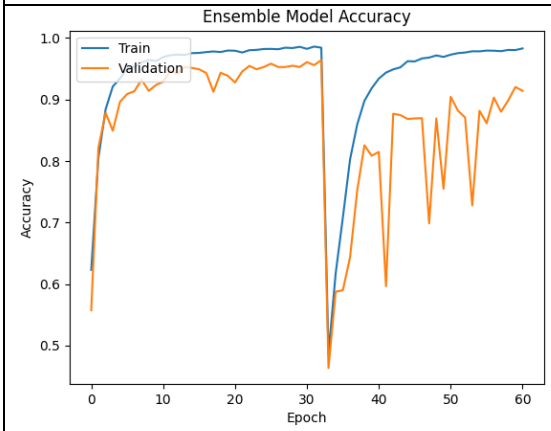
EfficientNetB2 model: Training vs. Validation Loss Graph



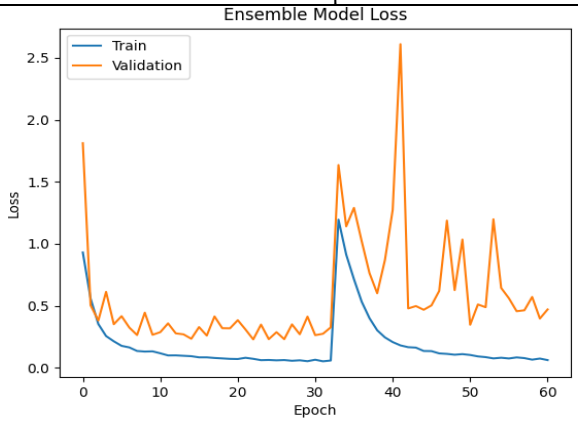
ResNet50 model: Training vs. Validation Accuracy Graph



ResNet50 model: Training vs. Validation Loss Graph

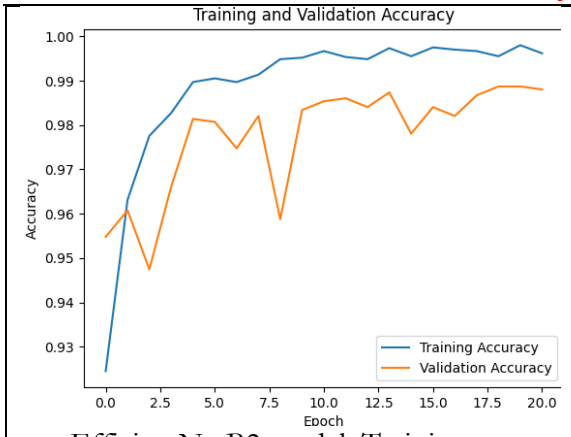


Ensemble model: Training vs. Validation Accuracy Graph

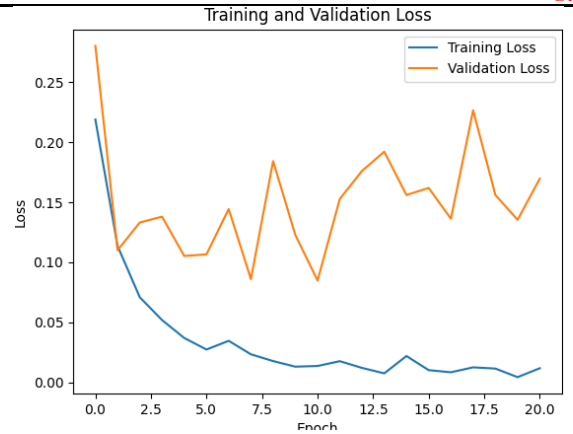


Ensemble model: Training vs. Validation Loss Graph

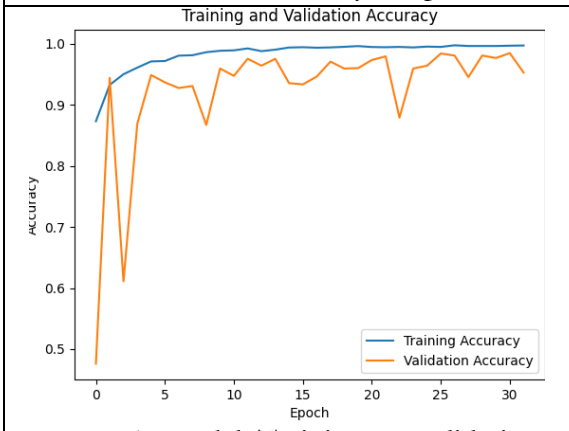
Figure 8. Training/validation accuracy and loss graphs for TL models and ensemble model (multi-class classification).



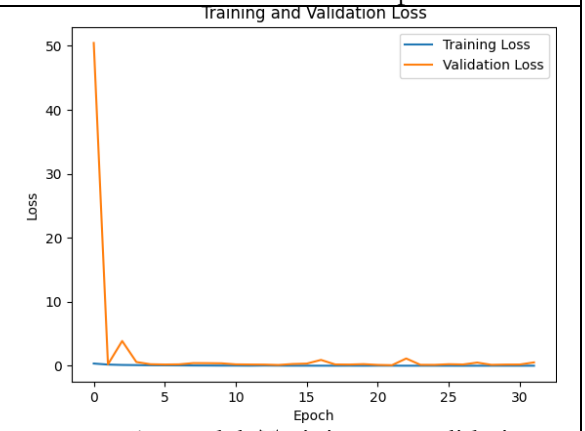
EfficientNetB2 model: Training vs. Validation Accuracy Graph



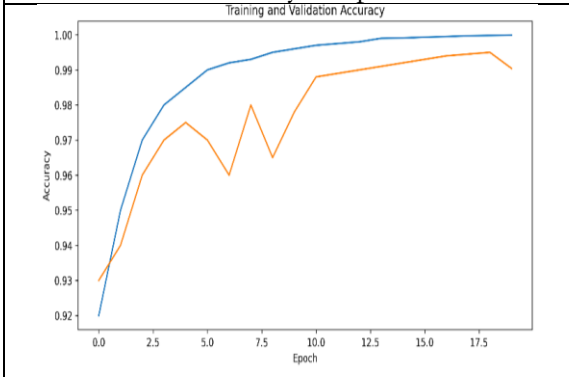
EfficientNetB2 model: Training vs. Validation Loss Graph



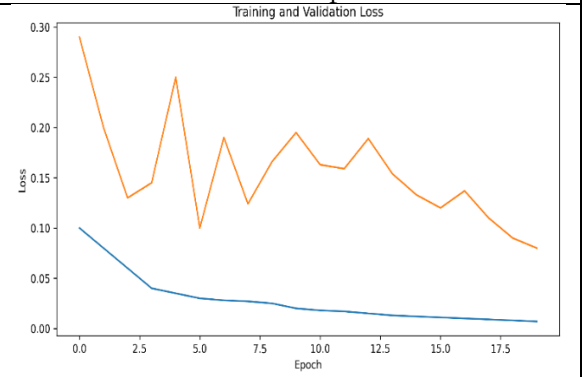
ResNet50 model: Training vs. Validation Accuracy Graph



ResNet50 model: Training vs. Validation Loss Graph



Ensemble model: Training vs. Validation Accuracy Graph



Ensemble model: Training vs. Validation Loss Graph

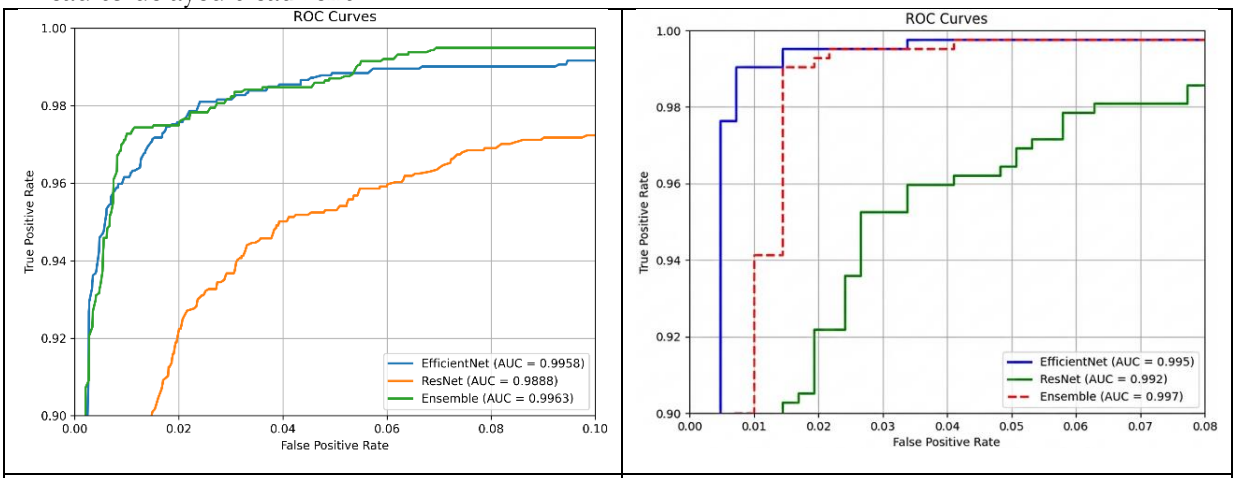
Figure 9. Training/validation accuracy and loss graphs for TL models and ensemble model (Binary-classification).

ROC Curve and AUC Analysis:

Figure 10 illustrates the macro-averaged ROC curves for multi-class diabetic retinopathy severity grading. Macro-averaging was adopted to ensure equal contribution from all severity classes, thereby mitigating bias toward majority classes. EfficientNetB2 achieved a high AUC value of 0.9958 indicating strong separability across DR severity levels. ResNet50 demonstrated comparatively lower performance with an AUC of 0.9888 reflecting its increased confusion among intermediate stages such as Moderate and Severe DR. The ensemble model outperformed both individual architectures, achieving the highest AUC of

0.9963. The consistently high true positive rates at low false positive rates further emphasize the robustness of the ensemble model for fine-grained severity grading.

The ROC curves for binary classification (DR vs. No DR) are also presented in Figure 10. All models demonstrated excellent discriminative performance, reflecting the comparatively simpler nature of binary DR detection. Ensemble model achieved the highest AUC of 0.997, followed by EfficientNetB2 with an AUC of 0.995 and ResNet50 with an AUC of 0.992. The steep ascent of the ROC curves near the y-axis indicates high sensitivity at low false positive rates which is critical for clinical screening scenarios where missed DR cases can lead to delayed treatment.



(a). Macro-averaged ROC curves for multi-class diabetic retinopathy severity grading, comparing EfficientNetB2, ResNet50, and the proposed ensemble model

(b). ROC curves for binary diabetic retinopathy classification (DR vs. No DR) showing the performance of EfficientNetB2, ResNet50, and the ensemble model

Figure 10. ROC curve graphs for all models (multi-class + Binary).

Quantitative Performance Comparison:

We calculated metrics for both training and testing. It can be seen that the Ensemble model outperformed in both multi-class and binary classification. Among individual models, EfficientNetB2 showed an amazing performance in classifying diabetic retinopathy into five and two classes. The role of ResNet50 was also crucial in DR classification and detection. Table 11 provides an analysis of the multi-classification results in detail. The binary classification results are shown in Table 12.

Table 11. Comparison of the models using various performance metrics (multi-classification).

Multi-class Classification												
TL Models		**ACC**	**MCR**	**SPC**	**SEN**	**QWK**	**PPV**	**NPV**	**FPR**	**FNR**	**LR+**	**LR-**
EfficientNetB2	**Training**	0.9940	0.0059	0.9985	0.9939	0.9961	0.9938	0.9985	0.0014	0.0060	∞	0.0060
	Testing	0.9618	0.0381	0.9905	0.9610	0.9736	0.9611	0.9905	0.0094	0.0389	122.64	0.0393
ResNet50	**Training**	0.9828	0.0171	0.9957	0.9827	0.9874	0.9825	0.9957	0.0042	0.0172	296.62	0.0173
	Testing	0.9247	0.0752	0.9812	0.9228	0.9409	0.9239	0.9813	0.0187	0.0771	56.981	0.0786
Ensemble Model	**Training**	0.9948	0.0051	0.9987	0.9948	0.9964	0.9947	0.9987	0.0012	0.0051	1807.42	0.0052
	Testing	0.9683	0.0316	0.9921	0.9676	0.9762	0.9677	0.9921	0.0078	0.0323	179.10	0.0326

Multi-class Classification:

Table 12. Comparison of the models using various performance metrics (binary classification).

Binary Classification												
TL Models		**ACC**	**MCR**	**SPC**	**SEN**	**QWK**	**PPV**	**NPV**	**FPR**	**FNR**	**LR+**	**LR-**
EfficientNetB2	**Training**	0.9990	0.0010	0.9984	0.9996	0.9980	0.9982	0.9997	0.0016	0.0004	637.77	0.0004
	Testing	0.9856	0.0144	0.9812	0.9902	0.9713	0.9807	0.9905	0.0188	0.0098	52.730	0.0099
ResNet50	**Training**	0.9950	0.0050	0.9941	0.9961	0.9900	0.9933	0.9965	0.0059	0.0039	167.45	0.0039
	Testing	0.9749	0.0251	0.9718	0.9781	0.9498	0.9710	0.9787	0.0282	0.0219	34.641	0.0225
Ensemble Model	**Training**	0.9960	0.0040	0.9950	0.9972	0.9920	0.9943	0.9975	0.0050	0.0028	199.05	0.0028
	Testing	0.9904	0.0096	0.9905	0.9903	0.9809	0.9903	0.9905	0.0095	0.0097	104.48	0.0098

Comparison with Existing Studies:

The best results of this study which were obtained using an ensemble model, are compared with earlier research regarding test accuracy performance. Numerous studies have focused on grading and detecting diabetic retinopathy. For multi-class classification, authors of [39] have fused two different models together CNN512 and YOLOv3 and used this for the classification of DR. In [16][25] have worked with hybrid architectures after combining deep learning architectures. Researchers of [12][40][34] have proposed Convolutional Neural Networks (CNN) architectures of their own. In [41][38] authors have utilized transfer learning models to achieve an optimal accuracy. Different datasets like APTOS, Messidor, EyePACS, and DeepDR are used by all the researchers along with various and distinctive pre-processing techniques, data augmentation and dataset split ratios. Similarly for binary classification, researchers in [22][24][42] developed their own deep learning models (CNNs) from scratch while researchers in [21][43][44] used pre-trained models like ResNet and InceptionV3. Among all these studies, the highest accuracy was observed on both types of classification using our proposed ensemble model. A comprehensive comparison of our ensemble model with previous research in multi-class classification is shown in Table 13. Table 14 presents this comparison in terms of binary classification.

Table 13. Analysis of test accuracy and multi-class classification performance: Ensemble model vs. previous studies

Study	Technique	Accuracy
[12]	CNN Model	92.65%
[38]	DenseNet-169	90%
[39]	Model Fusion	89%
[25]	Hybrid-model-c	86.34%
[40]	CNN Model	88-89%
[41]	ResNet-50	93.6%
[34]	CNN Model	94.44%
[16]	Hybrid Architectures	90.60%.
This Study	Ensemble Model	96.83%

Table 14. Analysis of test accuracy and Binary classification performance: Ensemble model vs. previous studies

Study	Technique	Accuracy
[22]	CNN Model	91%
[21]	InceptionV3	95%
[24]	CNN Model	95.5%
[42]	CNN Model	96%
[43]	ResNet34	96.35%
[44]	ResNet50	96.59%
[45]	EDLDR	96.98%
This Study	Ensemble Model	99.04%

Ensemble Model Explainability (Grad-CAM Analysis):

Explainable Artificial Intelligence (XAI) techniques aim to make the decision-making process of deep learning models transparent and interpretable which is critical in medical imaging where clinical trust is essential [46]. By providing insight into which regions of an image influence a model's predictions, XAI allows clinicians to verify that automated decisions are based on meaningful pathological features rather than irrelevant artifacts [47]. Gradient-weighted Class Activation Mapping is a widely used XAI method that generates visual heatmaps to highlight the areas of an image most influential for a model's prediction [48]. Grad-CAM leverages the gradients of a target class flowing into the final convolutional layer

to produce a coarse localization map of important regions, providing an intuitive understanding of model focus [49]. In this study, Grad-CAM was applied to our Ensemble model across all diabetic retinopathy stages. Figure 11 presents representative Grad-CAM visualizations illustrating the decision-making behavior of the proposed model across correctly classified samples from all five classes. For each class, three views are shown: the original retinal fundus image, the corresponding Grad-CAM heatmap, and the superimposed visualization highlighting the discriminative regions influencing the model's prediction.

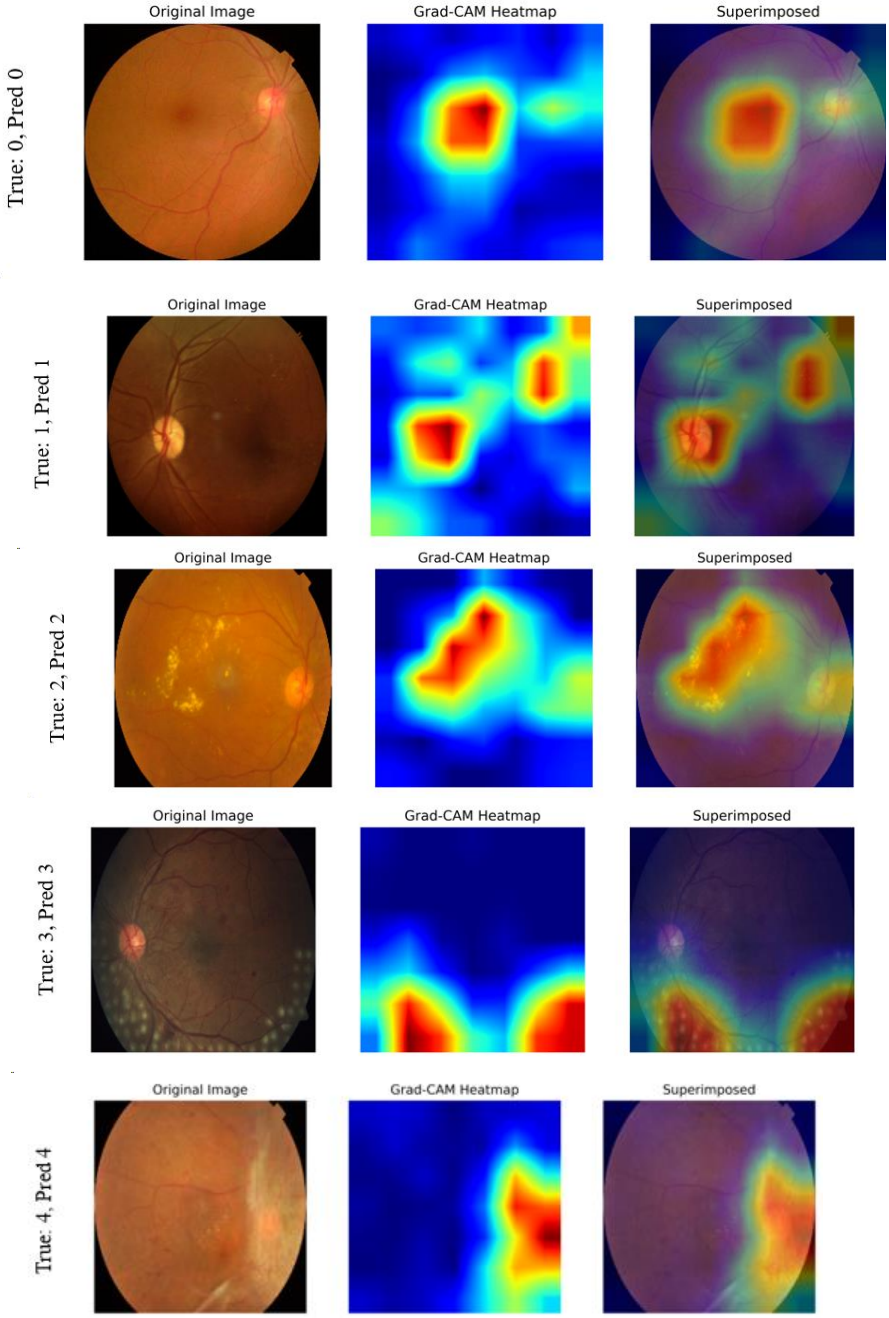


Figure 11. Grad-CAM–based visual explanations for correctly classified retinal fundus images across all five classes. For each class, the original image (left), the Grad-CAM heatmap (middle), and the superimposed visualization (right) are shown. The samples are ordered as: Class 0 (True 0, Predicted 0), Class 1 (True 1, Predicted 1), Class 2 (True 2, Predicted 2), Class 3 (True 3, Predicted 3), and Class 4 (True 4, Predicted 4). Warmer colors indicate regions that contribute more strongly to the model's prediction).

For Class 0 (True 0, Predicted 0), the activation map mainly focuses on the central retinal area. This indicates that the model correctly attends to normal retinal structures and does not rely on irrelevant peripheral regions, which is consistent with the absence of visible pathological signs.

In Class 1 (True 1, Predicted 1), the Grad-CAM highlights the optic disc and nearby blood vessels. This suggests that the model uses subtle structural and intensity changes in these regions to identify early-stage disease features.

For Class 2 (True 2, Predicted 2), the activation regions extend toward the lower and peripheral areas of the retina, where moderate lesions and texture variations are present. This shows that the model effectively captures disease-related features that are more widely distributed across the retina.

In the case of Class 3 (True 3, Predicted 3), strong and clearly defined activations appear over areas with dense lesion patterns, especially in the inferior retinal regions. The higher intensity of the heatmap reflects the model's focus on more prominent pathological features associated with severe disease.

Finally, for Class 4 (True 4, Predicted 4), the Grad-CAM visualization shows concentrated attention along the peripheral retinal regions. These areas correspond to extensive structural damage and severe pathological changes indicating that the model successfully identifies features linked to advanced disease stages.

Overall, these visualizations demonstrate that the model bases its predictions on meaningful retinal regions rather than background information supporting both the interpretability and reliability of the classification results.

Visual inspection indicates that the highlighted regions correspond to areas exhibiting retinal abnormalities in the examined images. However, these observations are qualitative as expert clinical validation and quantitative localization analysis were not performed.

Discussion:

The results of our study demonstrate that the proposed ensemble framework, combining EfficientNetB2 and ResNet50, achieves high accuracy and robust performance in multi-class and binary diabetic retinopathy grading. The ensemble's strong performance is particularly notable in differentiating between higher severity levels where pathological features are either absent or prominent leading to clearer visual patterns that are easier to classify. However, intermediate severity levels particularly Moderate and Severe DR remain challenging due to the transitional nature of the disease. Lesion characteristics such as microaneurysms, small hemorrhages, and exudates begin to overlap between classes which led to increased confusion. This trend is consistent with clinical practice, where differentiating adjacent DR stages is subject to inter-observer variability even among experienced ophthalmologists. Moreover, variations in imaging conditions including illumination, field of view, camera resolution, and contrast can obscure subtle pathological features, disproportionately affecting intermediate DR classes.

To improve data diversity and enhance model generalization, we fused two independent datasets, IDRiD and APTOS. Both these datasets are collected from different clinics and patient populations. IDRiD does not provide patient identifiers and APTOS contains only one image per patient, making patient-level splitting infeasible. While source-level splitting could further reduce potential bias from imaging devices or acquisition batches, the fusion of multiple sources was intended to better reflect real-world variability in fundus imaging. Our random image-level splitting preserves dataset independence and ensures meaningful evaluation of model performance.

The ensemble approach mitigated most of the challenges by combining complementary feature representations from EfficientNetB2 and ResNet50. The model demonstrated reduced misclassification between adjacent classes compared to individual

networks, highlighting its robustness for fine-grained DR grading. Residual misclassifications primarily occur between neighboring severity levels and reflect inherent ambiguities in retinal imaging rather than systematic model failure. This underscores the importance of ensemble-based architectures for addressing variability in retinal images and improving reliability in large-scale screening programs.

Clinically, the high AUC values observed in both multi-class and binary classification tasks confirm that the ensemble model provides stable and threshold-independent predictions across varying decision criteria. Such performance is crucial in practical DR screening, where operating points may vary depending on screening priorities, available resources and referral protocols. Overall, our results demonstrate that combining complementary deep feature representations can substantially reduce class confusion, improve threshold-independent performance, and offer a reliable solution for real-world DR screening programs.

Practical Implications and Recommendations:

The proposed ensemble framework has potential applications in computer-aided diabetic retinopathy screening and severity assessment using retinal fundus images. Its ability to perform both binary classification and five-class severity grading offers flexibility for different screening and assessment requirements. Binary classification may support preliminary identification of individuals with DR, while five-class grading provides more detailed information about disease severity. Furthermore, Grad-CAM visualizations offer visual insight into the image regions influencing the model's predictions. These capabilities may support the development of AI-assisted screening systems to facilitate retinal image assessment and assist eye-care professionals in the decision-making process.

Limitations and Future Work:

Despite the strong performance of the proposed framework, certain limitations exist. Datasets for diabetic retinopathy are subject to inherent biases arising from variations in patient populations, imaging devices, acquisition protocols and class distributions [50]. Such factors can influence model behavior and may affect generalization in real-world screening scenarios [51]. In this study, data augmentation techniques, including flipping, rotation, and zooming, were applied to the training data to reduce bias caused by limited samples and to encourage more robust feature learning across different severity grades. Despite these efforts, the confusion matrix analysis indicates that advanced disease stages still exhibit relatively higher misclassification rates compared to early or non-referable cases. This reflects the intrinsic complexity of severe and proliferative DR patterns, as well as the residual imbalance and visual overlap among higher-grade categories. These observations highlight that, while augmentation helps mitigate bias and improves robustness, it does not fully eliminate challenges associated with minority and advanced classes. Future work will focus on addressing these limitations through more targeted strategies for handling advanced disease stages with the aim of further improving fairness and clinical reliability. We also plan to perform independent evaluations on IDRiD and APTOS. Statistical testing and cross-dataset validation will be incorporated to further validate and strengthen the generalization capability of the proposed framework. Also, we will explore class-imbalance handling in the learned feature embedding space for better alignment with emerging medical AI preprocessing guidelines. Feature based implementation of SMOTE will also be explored.

Conclusion:

This research implements a transfer learning-based ensemble strategy to identify and classify diabetic retinopathy into two and five classes. Diabetic retinopathy damages the retinal blood vessels and leads to vision impairment if not treated on time. The key contribution of this research lies in data fusion where two DR datasets named IDRiD and APTOS 2019 were merged. This created a larger and more diverse dataset to enhance model training. Ensemble Learning was used to combine the predictions of individual models to develop a more robust

and accurate system. The retinal fundus images from this combined dataset underwent extensive pre-processing, including cropping to remove irrelevant regions and enhancing image quality using CLAHE. To address class imbalance, we applied the SMOTE technique to achieve an equitable distribution of images across all classes. This study fine-tuned two pre-trained models, ResNet50 and EfficientNetB2. Then the predictions of both models were averaged to construct the final ensemble model. This ensemble approach achieved a test accuracy of 96.83% on multi-classification and a test accuracy of 99.04% on binary classification using the fused dataset. Grad-CAM was used to highlight the areas of an image most influential for a model's prediction. In comparison with previous studies, this research's ensemble model showed superior performance highlighting the advantage of combining predictions from multiple models.

Authors Contributions:

Samia Akhtar and Shabib Aftab have collected data from different resources and contributed to writing—original draft preparation. Samia Akhtar. performed formal analysis and Simulation, Shabib Aftab and Muhammad Sulman Bashir; writing—review and editing, Samia Akhtar, Muhammad Sulman Bashir and Shabib Aftab; performed supervision, Shabib Aftab.; drafted pictures and tables, Samia Akhtar.; performed revisions and improve the quality of the draft, Muhammad Sulman Bashir. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest:

The authors have no conflict of interest.

References:

[1] P. M. Ebin and P. Ranjana, "A Variant Binary Classification Model for No-DR Mild-DR Detection using CLAHE Images with Transfer Learning," *Proc. Int. Conf. Comput. Commun. Secur. Intell. Syst. IC3SIS 2022*, 2022, doi: 10.1109/IC3SIS54991.2022.9885665.

[2] A. Bora *et al.*, "Predicting the risk of developing diabetic retinopathy using deep learning," *Lancet Digit. Heal.*, vol. 3, no. 1, pp. e10–e19, Jan. 2021, doi: 10.1016/S2589-7500(20)30250-8.

[3] Z. W. & H. Z. Feng Li, Yuguang Wang, Tianyi Xu, Lin Dong, Lei Yan, Minshan Jiang, Xuedian Zhang, Hong Jiang, "Deep learning-based automated detection for diabetic retinopathy and diabetic macular oedema in retinal fundus photographs," *Eye*, vol. 36, pp. 1433–1441, 2022, [Online]. Available: https://www.nature.com/articles/s41433-021-01552-8

[4] W. L. Alyoubi, W. M. Shalash, and M. F. Abulkhair, "Diabetic retinopathy detection through deep learning techniques: A review," *Informatics Med. Unlocked*, vol. 20, p. 100377, Jan. 2020, doi: 10.1016/J.IMU.2020.100377.

[5] S. Akhtar *et al.*, "A deep learning based model for diabetic retinopathy grading," *Sci. Reports 2025 151*, vol. 15, no. 1, pp. 3763-, Jan. 2025, doi: 10.1038/s41598-025-87171-9.

[6] Samia Akhtar,Shabib Aftab., "A Framework for Diabetic Retinopathy Detection using Transfer Learning and Data Fusion," *Res. gate*, vol. 16, no. 6, pp. 61–73, doi: 10.5815/ijitcs.2024.06.05.

[7] M. K. Jabbar, J. Yan, H. Xu, Z. U. Rehman, and A. Jabbar, "Transfer Learning-Based Model for Diabetic Retinopathy Diagnosis Using Retinal Images," *Brain Sci.*, vol. 12, no. 5, p. 535, May 2022, doi: 10.3390/BRAINSCI12050535.

[8] S. Ghosh and A. Chatterjee, "Transfer-Ensemble Learning based Deep Convolutional Neural Networks for Diabetic Retinopathy Classification," *2023 3rd Int. Conf. Adv. Electron. Commun. Eng. AECE 2023*, pp. 489–493, Aug. 2023, doi: 10.1109/AECE59614.2023.10428233.

[9] E. AbdelMaksoud, S. Barakat, and M. Elmogy, "A computer-aided diagnosis system for detecting various diabetic retinopathy grades based on a hybrid deep learning technique," *Med. Biol. Eng. Comput. 2022 607*, vol. 60, no. 7, pp. 2015–2038, May 2022, doi:

10.1007/S11517-022-02564-6.

[10] S. Akhtar, S. Aftab, M. Ahmad, and A. Akhtar, "Diabetic Retinopathy Severity Grading Using Transfer Learning Techniques," *Int. J. Eng. Manuf.*, vol. 14, no. 6, pp. 41–53, Dec. 2024, doi: 10.5815/IJEM.2024.06.04.

[11] Samia Akhtar,Shabib Aftab., "Diabetic Retinopathy Severity Classification Using Data Fusion and Ensemble Transfer Learning," *J. Softw. Eng. Appl.*, vol. 18, no. 1, pp. 1–23, doi: 10.4236/jsea.2025.181001.

[12] B. Ebrahimi *et al.*, "Optimizing the OCTA layer fusion option for deep learning classification of diabetic retinopathy," *Biomed. Opt. Express*, vol. 14, no. 9, p. 4713, Sep. 2023, doi: 10.1364/BOE.495999.

[13] G. Çınarer, K. Kılıç, and T. Parlar, "A Deep Transfer Learning Framework For The Staging of Diabetic Retinopathy," *J. Sci. Reports-A*, no. 051, pp. 106–119, Dec. 2022, Accessed: Oct. 21, 2026. [Online]. Available: https://dergipark.org.tr/en/pub/jsr-a/article/1166858

[14] S. Akhtar, S. Aftab, S. Kousar, A. Rehman, M. Ahmad, and A. Q. Saeed, "A Severity Grading Framework for Diabetic Retinopathy Detection using Transfer Learning," *2024 Int. Conf. Decis. Aid Sci. Appl. DASA 2024*, 2024, doi: 10.1109/DASA63652.2024.10836441.

[15] B. Menaouer, Z. Dermane, N. El Houda Kebir, and N. Matta, "Diabetic Retinopathy Classification Using Hybrid Deep Learning Approach," *SN Comput. Sci. 2022 35*, vol. 3, no. 5, pp. 357-, Jul. 2022, doi: 10.1007/S42979-022-01240-8.

[16] S. Yan, M. Feng, and Y. Cai, "An integrated deep learning model with enhanced EfficientNetB0 and MobileNetV1 for diabetic retinopathy grading," *Biomed. Signal Process. Control*, vol. 113, p. 108915, Mar. 2026, doi: 10.1016/J.BSPC.2025.108915.

[17] S. M. Darwish, K. G. Milad, and R. E. E. D. Ibrahim, "An enhanced diabetic retinopathy detection approach using optimized deep learning technique," *Sci. Reports 2026 161*, vol. 16, no. 1, pp. 9825-, Mar. 2026, doi: 10.1038/s41598-026-41998-y.

[18] Deepak Chandra Uprety, Jyotirmay Patel, Manju Ramrao Bhosle, Navpreet Kaur, Suresh Kumar, and Jamal Akhtar Khan, "An Explainable AI-Based Predictive Modeling Framework for Early Detection of Diabetic Retinopathy," *Int. J. Comput. Inf. Syst. Ind. Manag. Appl.*, vol. 18, no. 12s, pp. 1074–1082, Jul. 2026, doi: 10.70917/IJCISIM-2026-3980.

[19] M. Al-Smadi, M. Hammad, Q. B. Baker, and S. A. Al-Zboon, "A transfer learning with deep neural network approach for diabetic retinopathy classification," *Int. J. Electr. Comput. Eng.*, vol. 11, no. 4, pp. 3492–3501, Aug. 2021, doi: 10.11591/IJECE.V11I4.PP3492-3501.

[20] N. Sikder, M. Masud, A. K. Bairagi, A. S. M. Arif, A. Al Nahid, and H. A. Alhumyani, "Severity Classification of Diabetic Retinopathy Using an Ensemble Learning Algorithm through Analyzing Retinal Images," *Symmetry 2021, Vol. 13, Page 670*, vol. 13, no. 4, p. 670, Apr. 2021, doi: 10.3390/SYM13040670.

[21] B. Kotiyal and H. Pathak, "Diabetic Retinopathy Binary Image Classification Using Pyspark," *Int. J. Math. Eng. Manag. Sci.*, vol. 7, no. 5, pp. 624–642, 2022, doi: 10.33889/IJMEMS.2022.7.5.041.

[22] Padmanayana and D. A. Anoop, "Binary Classification of DR-Diabetic Retinopathy using CNN with Fundus Colour Images," *Mater. Today Proc.*, vol. 58, pp. 212–216, Jan. 2022, doi: 10.1016/J.MATPR.2022.01.466.

[23] C. L. Lin and K. C. Wu, "Development of revised ResNet-50 for diabetic retinopathy detection," *BMC Bioinforma. 2023 241*, vol. 24, no. 1, pp. 157-, Apr. 2023, doi: 10.1186/S12859-023-05293-1.

[24] A. Skouta, A. Elmoufidi, S. Jai-Andaloussi, and O. Ochetto, "Automated Binary Classification of Diabetic Retinopathy by Convolutional Neural Networks," *Adv. Intell. Syst. Comput.*, vol. 1188, pp. 177–187, 2021, doi: 10.1007/978-981-15-6048-4_16/SAVE-

RESEARCH.
[25] H. Liu, K. Yue, S. Cheng, C. Pan, J. Sun, and W. Li, "Hybrid Model Structure for Diabetic Retinopathy Classification," *J. Healthc. Eng.*, vol. 2020, 2020, doi: 10.1155/2020/8840174.
[26] X. Li, T. Pang, B. Xiong, W. Liu, P. Liang, and T. Wang, "Convolutional neural networks based transfer learning for diabetic retinopathy fundus image classification," *Proc. - 2017 10th Int. Congr. Image Signal Process. Biomed. Eng. Informatics, CISP-BMEI 2017*, vol. 2018-January, Jul. 2017, doi: 10.1109/CISP-BMEI.2017.8301998.
[27] Z. Rafie, M. S. Talab, B. E. Z. Koor, A. Garavand, C. Salehnasab, and M. Ghaderzadeh, "Leveraging XGBoost and explainable AI for accurate prediction of type 2 diabetes," *BMC Public Health*, vol. 25, no. 1, Dec. 2025, doi: 10.1186/S12889-025-24953-W.
[28] D. B. Soomro *et al.*, "Automated dual CNN-based feature extraction with SMOTE for imbalanced diabetic retinopathy classification," *Image Vis. Comput.*, vol. 159, p. 105537, Jun. 2025, doi: 10.1016/J.IMAVIS.2025.105537.
[29] S. Yasmin, M. Hasan, F. Jahan, and M. F. Mridha, "DR-RetinaNet: A Deep Learning Approach for Early Screening of Diabetic Retinopathy," *Human-Centric Intell. Syst. 2025 54*, vol. 5, no. 4, pp. 497–513, Nov. 2025, doi: 10.1007/S44230-025-00114-5.
[30] G. P. Reddy, P. D. Javali, R. Ramyashree, S. Raghavendra, C. B. Chandrakala, and P. S. Venugopala, "A Hybrid Deep Learning Approach With Explainable AI for Diabetic Retinopathy Classification," *IEEE Access*, vol. 14, pp. 28819–28839, 2026, doi: 10.1109/ACCESS.2026.3665564.
[31] B. Menaouer, Z. Dermane, N. El Houda Kebir, and N. Matta, "Diabetic Retinopathy Classification Using Hybrid Deep Learning Approach," *SN Comput. Sci. 2022 35*, vol. 3, no. 5, pp. 357-, Jul. 2022, doi: 10.1007/S42979-022-01240-8.
[32] P. Porwal *et al.*, "Indian Diabetic Retinopathy Image Dataset (IDRiD)," *IEEE Dataport*, 2018, doi: 10.21227/H25W98.
[33] "APTOS 2019 Blindness Detection | Kaggle." Accessed: Oct. 21, 2026. [Online]. Available: https://www.kaggle.com/c/aptos2019-blindness-detection
[34] D. Das, S. K. Biswas, and S. Bandyopadhyay, "Detection of Diabetic Retinopathy using Convolutional Neural Networks for Feature Extraction and Classification (DRFEC)," *Multimed. Tools Appl.*, vol. 82, no. 19, p. 1, Aug. 2022, doi: 10.1007/s11042-022-14165-4.
[35] M. S. Khandakar, M. Samsuddoha, S. Jahan, and R. H. Faisal, "An automated analytics approach for diabetic retinopathy detection with ensemble deep learning models in healthcare," *Healthc. Anal.*, vol. 9, p. 100449, Jun. 2026, doi: 10.1016/J.HEALTH.2026.100449.
[36] S. Saleem, R. Talib, M. K. Hanif, and M. U. Sarwar, "Clinical-guided deep learning framework for diabetic retinopathy: integrating lesion-aware attention, adversarial augmentation, and uncertainty quantification," *Sci. Reports 2026 161*, vol. 16, no. 1, pp. 24161-, Aug. 2026, doi: 10.1038/s41598-026-58690-w.
[37] S. Akhtar, S. Aftab, S. Kousar, A. Rehman, M. Ahmad, and A. Q. Saeed, "Performance Analysis of Training Optimizers on Diabetic Retinopathy Detection Using SP-Net CNN," *2024 Int. Conf. Decis. Aid Sci. Appl. DASA 2024*, 2024, doi: 10.1109/DASA63652.2024.10836543.
[38] Y. Fu, Y. Wei, S. Chen, G. Mushtaq, and F. Siddiqui, "Detection of diabetic retinopathy using deep learning methodology," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1070, no. 1, p. 012049, Feb. 2021, doi: 10.1088/1757-899X/1070/1/012049.
[39] W. L. Alyoubi, M. F. Abulkhair, and W. M. Shalash, "Diabetic Retinopathy Fundus Image Classification and Lesions Localization System Using Deep Learning," *Sensors (Basel).*, vol. 21, no. 11, Jun. 2021, doi: 10.3390/S21113704.
[40] M. Shaban *et al.*, "A convolutional neural network for the screening and staging of diabetic retinopathy," *PLoS One*, vol. 15, no. 6, Jun. 2020, doi: 10.1371/JOURNAL.PONE.0233514.

[41] S. H. Abbood, H. N. A. Hamed, M. S. M. Rahim, A. Rehman, T. Saba, and S. A. Bahaj, "Hybrid Retinal Image Enhancement Algorithm for Diabetic Retinopathy Diagnostic Using Deep Learning Model," *IEEE Access*, vol. 10, pp. 73079–73086, 2022, doi: 10.1109/ACCESS.2022.3189374.

[42] N. Nasir, P. Oswald, O. Alshaltone, F. Barneih, M. Al Shabi, and A. Al-Shammaa, "Deep DR: Detection of Diabetic Retinopathy using a Convolutional Neural Network," *2022 Adv. Sci. Eng. Technol. Int. Conf. ASET 2022*, 2022, doi: 10.1109/ASET53988.2022.9734314.

[43] R. Adriman, K. Muchtar, and N. Maulina, "Performance Evaluation of Binary Classification of Diabetic Retinopathy through Deep Learning Techniques using Texture Feature," *Procedia Comput. Sci.*, vol. 179, pp. 88–94, Jan. 2021, doi: 10.1016/J.PROCS.2020.12.012.

[44] M. Rao, M. Zhu, and T. Wang, "Conversion and Implementation of State-of-the-Art Deep Learning Algorithms for the Classification of Diabetic Retinopathy," Oct. 2020, Accessed: Oct. 21, 2026. [Online]. Available: https://arxiv.org/pdf/2010.11692

[45] S. S. Mondal, N. Mandal, K. K. Singh, A. Singh, and I. Izonin, "EDLDR: An Ensemble Deep Learning Technique for Detection and Classification of Diabetic Retinopathy," *Diagnostics*, vol. 13, no. 1, p. 124, Jan. 2022, doi: 10.3390/DIAGNOSTICS13010124.

[46] Z. Salehnasab, A. Mousavizadeh, G. Ghalamfarsa, A. Garavand, and C. Salehnasab, "Predictive modeling of COVID-19 hospitalization using twenty machine learning classification algorithms on cohort data," *Front. Heal. Informatics*, vol. 12, 2023, doi: 10.30699/FHI.V12I0.473.

[47] D. S. S. Raja, S. Kumarganesh, K. M. Sagayam, and H. Dang, "Diabetic retinopathy detection and grading system using deep learning approach," *Digit. Heal.*, vol. 12, p. 20552076251410984, Jan. 2026, doi: 10.1177/20552076251410982.

[48] G. Ali *et al.*, "Hierarchical deep learning with stability-driven normalization and regularization for accurate multi-class diabetic retinopathy diagnosis," *Ain Shams Eng. J.*, vol. 17, no. 12, p. 104453, Dec. 2026, doi: 10.1016/J.ASEJ.2026.104453.

[49] M. T. A. Tashrif *et al.*, "Privacy-aware diabetic retinopathy grading and visual lesion-focused interpretability through mixture-of-experts federated deep learning with explainable AI," *Sci. Reports 2026*, Jun. 2026, doi: 10.1038/S41598-026-58292-6.

[50] M. Ghaderzadeh and C. Salehnasab, "Filter-Based Feature Selection for Type II Diabetes Prediction," *J. Clin. Care Ski.*, vol. 6, no. 3, pp. 121–128, 2025, doi: 10.58209/JCCS.6.3.121.

[51] M. Ghaderzadeh, A. Garavand, and C. Salehnasab, "Artificial intelligence in polycystic ovary syndrome: a systematic review of diagnostic and predictive applications," *BMC Med. Inform. Decis. Mak.*, vol. 25, no. 1, p. 427, Dec. 2025, doi: 10.1186/S12911-025-03255-6.



Copyright © by authors and 50Sea. This work is licensed under Creative Commons Attribution 4.0 International License.