Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 8, 2025

DIABETIC RETINOPATHY GRADING USING DATA FUSION AND CNN-BASED ENSEMBLE LEARNING WITH SOFT VOTING FOR ENHANCED TOP-K ACCURACY

Muhammad Talha Jahangir[*1], Amna Shahid[2], Naveed Ahmad[3], Azhar-ud-din[4], Muhammad Rehman[5], Ali Annas[6]

[*1,2,4,5,6]*Department of Computer Science, MNS-University of Engineering & Technology Multan, Multan, Pakistan*
[3]*Department of Computer Science،NCBA&E, Multan Campus*

[*1]mtalhajahangir@mnsuet.edu.pk

DOI:https://doi.org/10.5281/zenodo.17198671

Keywords
Diabetic Retinopathy, Ophthalmologists, Deep Learning, Xception, DenseNet, Inception V3, Fine Tuning, Ensemble Learning with Soft Voting, Data Fusion, Top-k accuracy

Article History
Rece'need: 19 May, 2025
Accepted: 24 July, 2025
Published. 21 August, 2025

Corresponding Author: *
Muhammad Talha Janhangir

Abstract

Diabetic retinopathy (DR) is the most prevalent eye ailment and leading global causes of vision blindness in humans. To detect and classify retinal images can demand specialized expertise. Fortunately, the integration of deep learning and ensemble learning provides essential support to healthcare professionals. We developed a multi-model classification approach that identifies and ranks diabetic retinopathy severity across five stages. The methodology incorporates four key benchmark datasets: APTOS 2019, IDRiD, Messidor-2, and DDR creating a unified collection of fundus images that are subsequently enhanced through pre-processing procedures. To address class imbalance, we applied the duplicate oversampling technique. The dataset was then divided into training (70%), validation (10%), and testing (20%) sets, maintaining sufficient data for effective training. Our methodology employs three state-of-the-art deep learning architectures (Xception, InceptionV3, and DenseNet121) trained independently and combined through soft voting. Our study introduces top-k accuracy metrics to evaluate the clinical utility of the system. The ensemble model achieved superior performance compared to individual models, with a validation accuracy of 97%. The top-k accuracy metrics demonstrated the clinical value of our approach with top-1 accuracy of 96.86%, top-2 accuracy of 99.30%, and top-3 accuracy of 100.00%. Our data fusion and ensemble learning approach significantly improves the reliability and generalizability of DR grading across diverse populations and imaging conditions. The high top-k accuracy metrics demonstrate the clinical potential of this system as an assistive tool for ophthalmologists.

INTRODUCTION

Prolonged diabetes causes diabetic retinopathy (DR), a progressive disease that affects the retina of an eye, which is left untreated can cause blindness or visual loss. Early diagnosis is crucial as it sometimes begins without any clear symptoms [1]. Some of the signs patients may have as the disease develops are blurred vision, floaters, black spots, decreased color vision, and finally vision loss [2]. DR is difficult because it develops slowly and comes in many degrees of severity, therefore correct detection and grading of it is difficult to offer prompt and successful treatment [4][5]. The majority of the good diabetic-retinopathy (DR) examinations are performed after the pupils have been dilated, followed by the observation and/or photography of the retina. Deep learning and ensemble-based models which employ large retinal image datasets and transfer learning to increase classification accuracy have lately shown promise in

enhancing DR diagnosis [6]. Each of diabetic retinopathy's several phases has distinct clinical features. Patients in the No Diabetic Retinopathy (No DR) phase show no apparent indications of retinal injury, preventative efforts focus on maintaining good glycemic control. Micro-aneurysms, small bulges in the retinal capillaries appear during the Mild Non-Proliferative Diabetic Retinopathy (Mild NPDR) stage; however, symptoms might not be obvious at first.

As the disease progresses into Moderate Non-Proliferative Diabetic Retinopathy (Moderate NPDR), the retina shows more micro-aneurysms, hemorrhages therefore causing some vision loss. Severe ischemia results from significant blockage of retinal blood arteries in severe non-proliferative diabetic retinopathy (Severe NPDR), and symptoms including:

- Venous beading
- Intraretinal microvascular abnormalities (IRMA) become apparent.

Characterized by neovascularization, that is, the growth of new, weak blood vessels on the retina and into the vitreous body likely to break the ultimate and most advanced stage, proliferative diabetic retinopathy (PDR), might cause vitreous hemorrhage if left untreated and retinal detachment and significant visual loss as indicated in Figure 1.

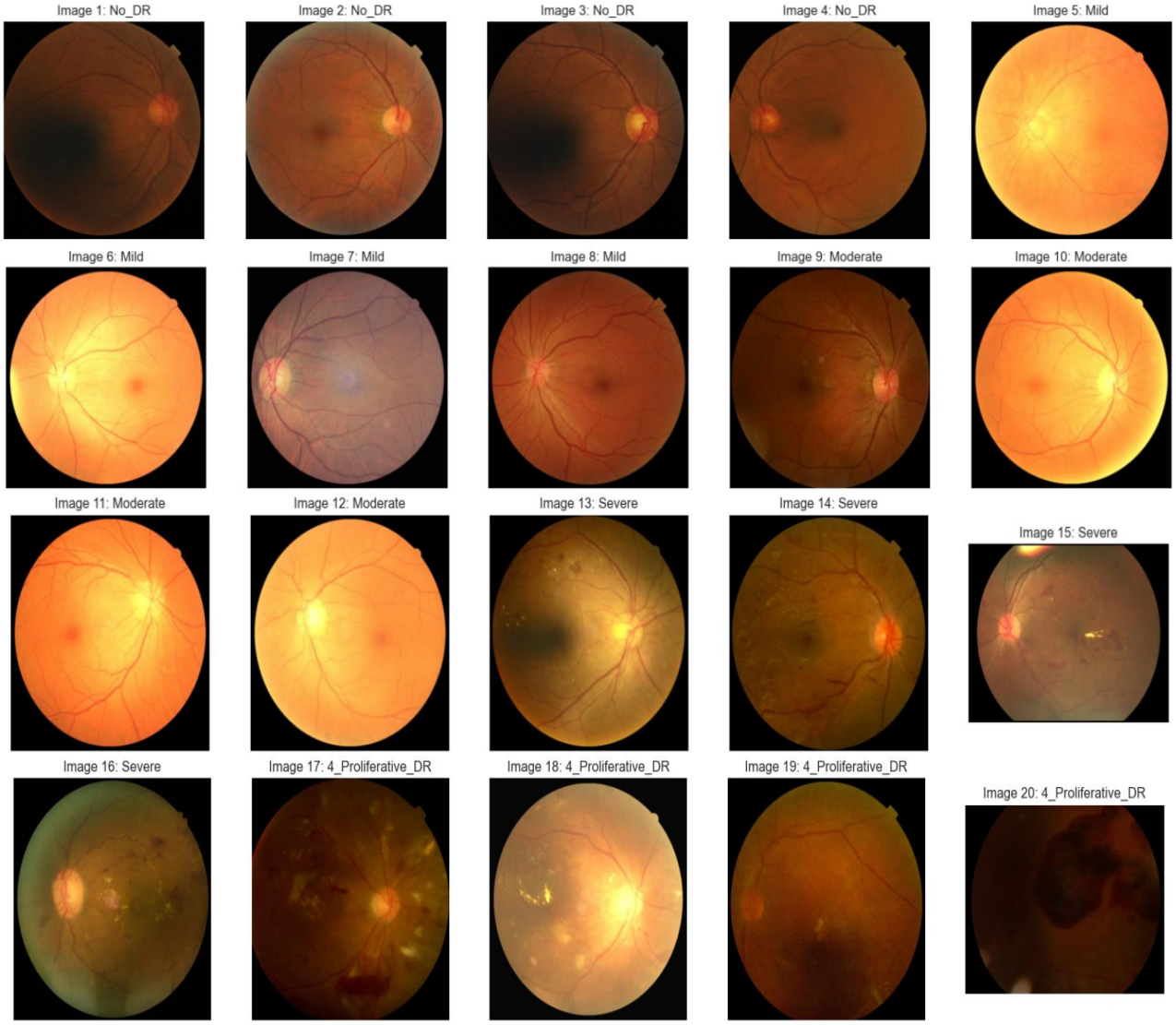


Figure 1: Mapping class labels to their corresponding categories

Our study has brought attention to the use of CNN-based ensemble learning with soft voting techniques, significantly improving Top-k accuracy for multi-class DR classification tasks. The projected global prevalence of diabetic retinopathy (DR) from 2020 to 2045, highlighting the rapid growth from 103 million in 2020 to 161 million in 2045, a 56% increase over 25 years, and over 25% growth in just the first decade. Figure 2 shows the sharp rise of diabetic retinopathy (DR).[3]

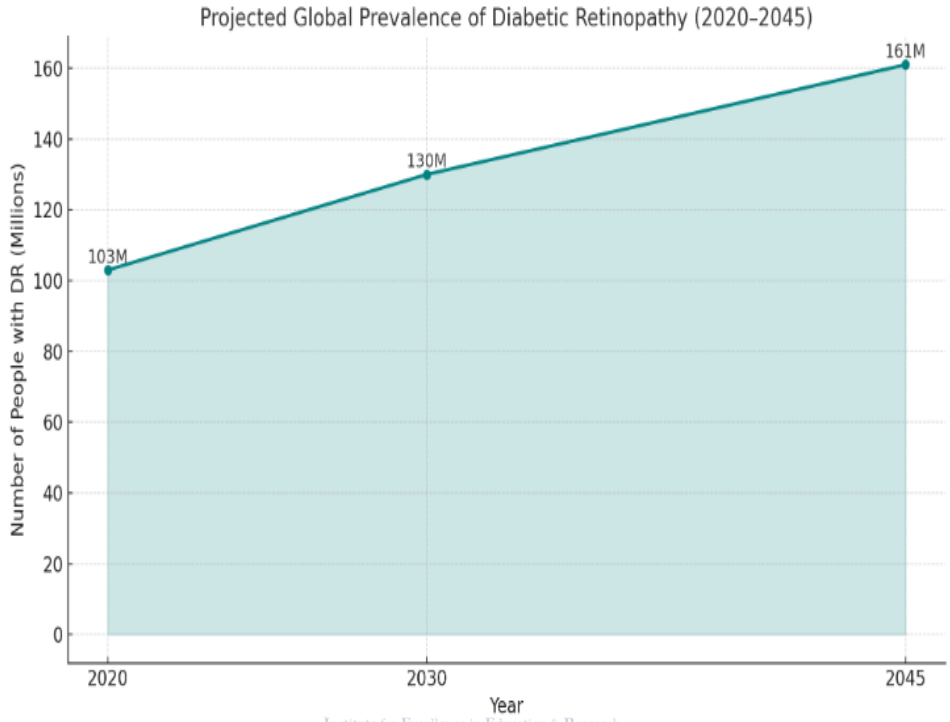


Figure 2: Projected Global Prevalence of Diabetic Retinopathy (2020-2045)

Traditionally, the diagnosis of DR involves manual effort from ophthalmologists through clinical examination and fundus photography analysis, using methods such as optical coherence tomography (OCT). While effective, these techniques are time-consuming, resource-intensive, and highly dependent on trained professionals. In recent years, deep learning (DL), has progressed as a powerful tool for automating medical image analysis. By learning hierarchical features directly from retinal images, DL models have exhibited high accuracy in classifying DR severity levels and detecting subtle pathological features. These models not only alleviate the diagnostic workload on clinicians but also enable large-scale, especially in low-resource settings. Our study proposes a deep learning-based ensemble approach to classify diabetic retinopathy into five severity stages. We combined four standard datasets: APTOS 2019, IDRiD, Messidor-2, and DDR, applied pre-processing and duplicate oversampling to address class imbalance, data augmentation techniques and dataset division. Three models Xception, InceptionV3, and DenseNet121 were separately trained and then merged via soft voting. With a validation accuracy of 97%, ensemble learning showed great clinical use and enhanced generalizability. Research Contributions: By data fusion combined with a CNN-based ensemble learning technique, this study offers a fresh method for diabetic retinopathy (DR) grading. The main contributions are as follows:

- Generalizability: To guarantee diversity, a combined dataset of 18,422 images: APTOS, IDRiD, Messidor-2, DDR is used. Advanced augmentation and oversampling increase generalizability.
- Ensemble Deep Models: Xception, InceptionV3, and DenseNet121 are combined using soft voting, boosting accuracy and reducing model variance.

- Top-k Accuracy Results: Top 1 (96.86%), top-2 (99.30%), and top-3 (100%) .

II. LITERATURE REVIEW

Abood and colleagues [1] developed a transfer learning framework utilizing convolutional neural networks for multi-label diabetic retinopathy classification. Their research focused on the APTOS 2019 dataset, implementing comprehensive preprocessing workflows that included dimensional adjustments, standardization, and augmentation strategies to improve model robustness. The investigation leveraged established architectures including VGG16 and InceptionV3 for feature extraction, specifically addressing the hierarchical nature of DR severity classification. Performance evaluation revealed an F1-score of 0.87 alongside 92.4% classification accuracy, validating the efficacy of pre-trained network adaptation for retinal pathology detection. Liu and team [5] introduced an ensemble methodology combining augmented datasets with soft-voting CNN architectures for enhanced DR detection capabilities. The research integrated IDRID and Kaggle DDR datasets, applying Contrast Limited Adaptive Histogram Equalization (CLAHE) and rotational augmentation techniques to expand training data diversity. Their approach merged EfficientNetB0 and DenseNet121 architectures through soft-voting aggregation mechanisms. Results demonstrated 94.7% accuracy with 91.2% sensitivity and 95.6% specificity, emphasizing ensemble methodology advantages in medical imaging applications.

Felorunsho et al. [8] constructed an interpretable deep learning framework prioritizing both accuracy and explainability for DR classification. Their preprocessing pipeline made use of lesion-specific filtering, normalization, and enhancement approaches, all based on the APTOS 2019 dataset. The model integrated CNN ensemble architectures with explainability frameworks including Grad-CAM and SHAP for prediction visualization. The system achieved 93.2% classification accuracy while providing clinical decision transparency, addressing critical requirements for AI deployment in healthcare environments.

Bhulakshmi's research team [14] investigated personalized federated deep learning applications for DR detection across distributed medical datasets. The study combined Kaggle and clinical data sources while maintaining privacy through federated learning protocols. Local preprocessing involved standardization procedures applied at individual client nodes. A modified ResNet18 architecture facilitated global model training through federated averaging with personalized local adaptations. The approach delivered 90.1% global accuracy with 3-4% individual client improvements, demonstrating federated learning potential in privacy-sensitive medical contexts.

Sunkari et al. [16] developed an enhanced ResNet18 framework incorporating Swish activation functions and optimized dropout strategies for improved DR classification. The investigation utilized combining APTOS 2019 and Messidor-2 datasets with 512×512 pixel resizing and histogram equalization enhancement. Their modified ResNet18 architecture with Swish activation demonstrated superior learning characteristics, achieving 93.8% accuracy and 0.91 F1-score, indicating architectural refinement benefits for DR classification tasks.

Ihnaini and colleagues [19] presented a data fusion-based ensemble transfer learning approach. The methodology merged Kaggle DDR and APTOS datasets, implementing preprocessing including normalization and weighted averaging fusion techniques. The ensemble combined DenseNet201 and InceptionV3 architectures for comprehensive feature extraction. Results showed 95.3% accuracy and 0.96 AUC, highlighting multi-source data fusion effectiveness in robust DR detection system development.

Al-Dhubaib et al. [28] explored segmented retinal vessel analysis for DR diagnosis using the IDRID dataset. The research employed U-Net-based vessel segmentation for extracting vascular maps, subsequently utilized as CNN classifier inputs. This segmentation-focused methodology emphasized vascular changes' diagnostic significance in DR progression. Despite relying exclusively on vessel features, the model achieved 88.5% classification accuracy, demonstrating retinal vasculature's informative value for early DR detection.

Rahman's team [35] implemented automated DR grading using DenseNet-169 architecture with the Kaggle APTOS 2019 dataset. Preprocessing included histogram equalization and 224×224-pixel standardization for enhanced image quality. The

methodology involved DenseNet-169 fine-tuning for five-level DR severity classification. Performance evaluation showed 91.2% accuracy and 0.945 AUC, indicating effective discrimination across severity levels.

Thota and Umma Reddy [39] investigated transfer learning effectiveness for DR severity classification using the Messidor-2 dataset. Preprocessing addressed illumination and color inconsistencies through RGB normalization and contrast enhancement. The study adopted pre-trained VGG19 networks for DR classification, addressing limited annotated data challenges. The resulting model achieved 89.4% accuracy, demonstrating significant improvement in DR severity detection with reduced computational requirements.

Bhardwaj et al. [42] put forward an automatic DR grading system utilizing ResNet50-based transfer learning. The research integrated IDRID and Kaggle APTOS datasets. Preprocessing incorporated Z-score normalization and extensive augmentation to address overfitting and class imbalance challenges. The CNN-transfer learning hybrid achieved 92.6% accuracy, demonstrating robustness across imbalanced distributions and dataset heterogeneity Shabib Aftab and colleagues [2] developed an ensemble classification framework for five-class DR detection and grading through data fusion and transfer learning. The methodology combined three datasets: APTOS 2019, IDRID, and Messidor-2 and after fusion totaling 5922 fundus images. Preprocessing included cropping, Gaussian blur denoising, CLAHE contrast enhancement, and 224×224 pixel resizing. Class imbalance was addressed through SMOTE implementation, followed by flipping, rotation, and zooming augmentation techniques, expanding the dataset to 26,180 images with 70:10:20 train-validation-test splits. The classification employed fine-tuned EfficientNetB2, DenseNet121, and ResNet50 models with averaged prediction ensemble. The ensemble framework achieved 96.96% test accuracy for five-class DR grading. Table 1 shows the summary of the literature reviewed.

Table 1: A summary of reviewed literature.

Paper	Datasets	Preprocessing Techniques	Methodology	Results
[1]	Kaggle APTOS 2019	Image resizing, normalization, data augmentation	Multi-label CNN with VGG16 and InceptionV3 backbones (transfer learning)	F1-score: 0.87, Accuracy: 92.4%
[5]	IDRID + Kaggle DDR	CLAHE, resizing, rotation-based augmentation	CNN ensemble: EfficientNetB0 + DenseNet121, soft voting	Accuracy: 94.7%, Sensitivity: 91.2%, Specificity: 95.6%
[8]	APTOS 2019	Normalization, image enhancement, lesion-specific filters	CNN ensemble with Grad-CAM and SHAP for explainability	Accuracy: 93.2%
[14]	Kaggle + Hospital Data	Image normalization on local clients	Federated ResNet18 with personalized layers and federated averaging	Global Accuracy: 90.1%, Local Improvement: +3–4%
[16]	APTOS + Messidor-2	Resize to 512×512, histogram equalization	Modified ResNet18 with Swish activation, dropout optimization	Accuracy: 93.8%, F1-score: 0.91
[19]	Kaggle DDR + APTOS	Normalization, weighted image fusion	Ensemble transfer learning: DenseNet201 + InceptionV3	Accuracy: 95.3%, AUC: 0.96
[28]	IDRID	U-Net-based retinal vessel segmentation	CNN trained on segmented vessel maps	Accuracy: 88.5%
[35]	APTOS 2019	Histogram equalization, resized to 224×224	Fine-tuned DenseNet-169 for 5-class DR classification	Accuracy: 91.2%, AUC: 0.945
[39]	Messidor-2	RGB normalization, contrast enhancement	Pre-trained VGG19 fine-tuned for DR classification	Accuracy: 89.4%

[42]	IDRID + APTOS	Z-score normalization, extensive data augmentation	Transfer learning with ResNet50 backbone	Accuracy: 92.6%
[2]	APTOS 2019, IDRiD, Messidor-2	Cropping, denoising (Gaussian blur), CLAHE, resizing (224×224), SMOTE, augmentation (flip, zoom, rotate)	Ensemble of EfficientNetB2, DenseNet121, and ResNet50	Accuracy: 96.96%
Our Study	APTOS+MESSIDOR-2+DDR+IDRID	Black background removal, Gaussian blur, CLAHE, gamma correction, sharpening, and resizing to 380×380.	Ensemble learning using DenseNet121, InceptionV3, and Xception models trained on the fused dataset.	Accuracy:97% Top 1: 96.86% Top 2: 99.30% Top 3: 100.00%

III. DATASETS INVOLVED:

Four benchmark datasets are used in this study: Messidor-2, DDR, IDRiD stands for: Indian Diabetic Retinopathy Image Dataset, APTOS 2019 stands for: Asia Pacific Tele-Ophthalmology Society 2019, publicly available dataset [23]. These pictures are shot in several clinical situations and provide a wider representation of retinal diseases. On a scale of 0 to 4, hospital clinicians have painstakingly categorized the degree of diabetic retinopathy in every picture. Five different grades of severity make up this grading system for the illness. The first database representing the Indian population is IDRiD [24]. We focus on grading diseases, which includes examining the severity of diabetic retinopathy. IDRiD uses five categories to categorize fundus images, like the APTOS dataset: No DR, Mild, Moderate, Severe, and Proliferative DR. Messidor-2: The benchmark dataset Messidor-2 is also freely accessible at ADCIS [25]. It comprises diabetic retinopathy tests, each of which has two fundus images that are focused on the macula. There are 1,748 images in all, with clinicians assigning a score of 0 to 4 to 1,744 of them. A fundus camera was used to capture these pictures without pharmacological dilation. DDR: The DDR stands for: Dataset for Diabetic Retinopathy is a notable standard dataset comprising 13,673 fundus images from 9,598 patients across 147 hospitals in 23 provinces of China. These high-quality, expertly annotated images are classified into five levels of diabetic retinopathy severity, making it a valueable asset for deep learning models.

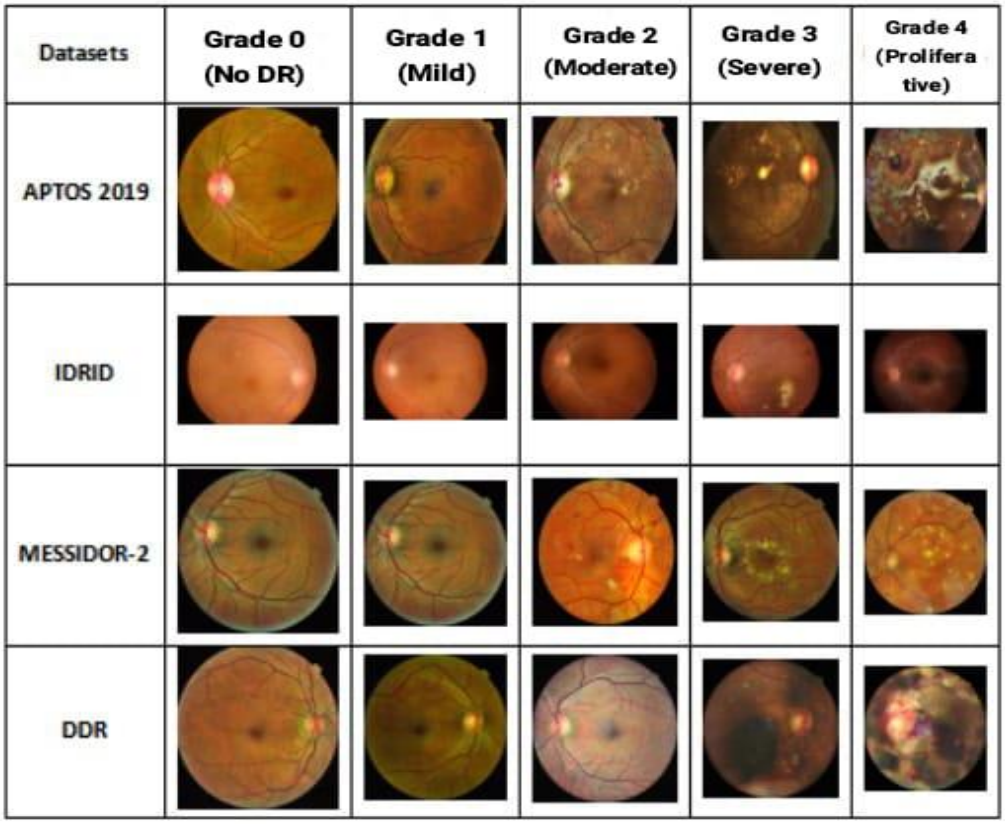


Figure 3: A fundus image from each grade of four

DR datasets

Data fusion's benefit is that it improves the model's ability. The overfitting to a single dataset is decreased by training the model on a wider range of data. By increasing the number of images, fusion also gives the model, a large number of data to learn from. DDR enabled models to learn from various retinal images in our instance of diabetic retinopathy (DR) detection, which combined datasets such as IDRiD, APTOS, and Messidor-2. Figure 4 shows the class distribution after data fusion.

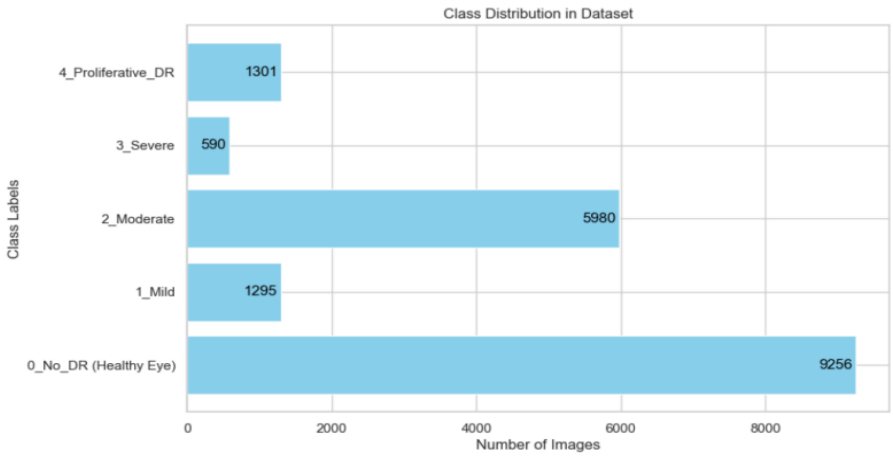


Figure 4: Class Distribution in Dataset after Data Fusion

The horizontal bar chart shows an imbalanced dataset for diabetic retinopathy (DR) classification. There are significantly more images of healthy eyes (9256) compared to the different stages of DR: Mild (1295), Moderate (5980), Severe (590), and Proliferative (1301). This imbalance highlights a challenge for training accurate DR detection models.

Pre-Processing:

Data preprocessing upholds data quality and is done in the following steps:

Removing Black Background: This step aims to isolate the relevant image content by addressing the presence of a black background. It converts the image to grayscale and applies a binary threshold to create a mask. This mask is then used to set the alpha channel of the image, effectively making the black (or near-black) pixels transparent. This can improve the focus of subsequent processing and model training.

Applying Gaussian Blur: Gaussian blur is applied to reduce high-frequency noise within the images. This smoothing operation helps to minimize the impact of minor variations and artifacts that might not be relevant to the underlying structures. The degree of the blur is determined by the kernel size of (5,5), which averages pixel values across a 5x5 neighborhood. A standard deviation of 0 indicates that the kernel will determine the sigma based on its size.

Applying CLAHE: CLAHE stands for: Contrast Limited Adaptive Histogram Equalization improves the visibility of minute details in both bright and dark areas by enhancing the local contrast of the fundus images. The images is divided into tiny 8x8 tiles and using histogram equalization on each tile, it functions. By restricting the improvement of contrast within each tile before redistributing the clipped values, the 0.5 clipLimit stops excessive noise amplification. This adaptive approach helps to reveal subtle features without over-saturating uniform areas.

Applying Gamma Correction: Gamma correction alters the overall brightness and tonal range of the images using a gamma value of 1.1. Since the gamma value is greater than 1, this operation will slightly darken the image. This adjustment can help to normalize the intensity distribution and potentially improve the discrimination of features, especially in images that might be consistently too bright.

Applying Sharpening: The sharpening step enhances the edges and fine details in the images, making them appear more distinct. It utilizes a predetermined 3x3 sharpening kernel that is convolved with the picture. This kernel emphasizes the difference in pixel intensities between neighboring pixels, thereby accentuating edges and making the structures within the image more pronounced.

Resizing Image: All processed images are scaled to a consistent dimension of 299*299 pixels. This standardization is pivotal for many deep learning models that entail a fixed input size. The cv2.resize function is used for this purpose, and by default, it likely uses bilinear interpolation to estimate the pixel values in the resized image, providing a good balance between speed and quality.

All pre-processing steps are shown in Figure 5.

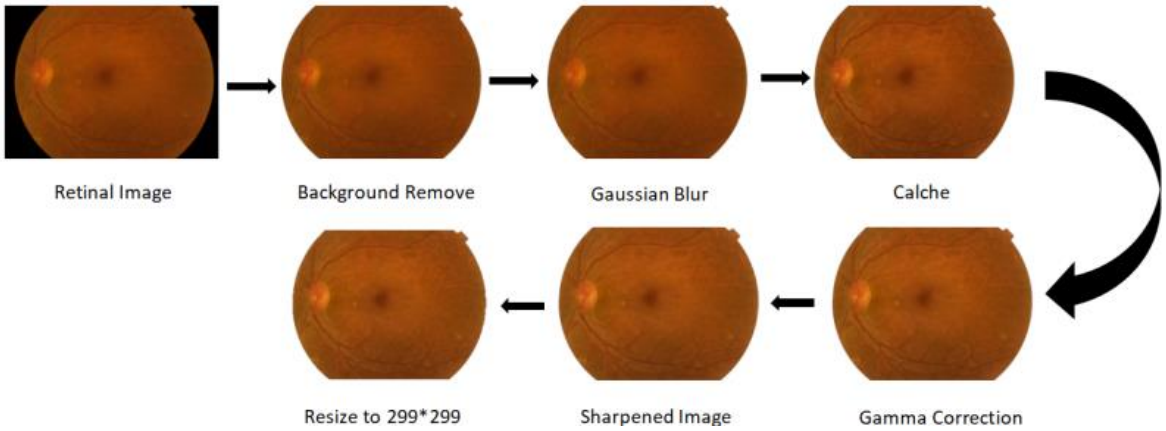


Figure 5: Pre-processing Steps

Class Balancing:
After the application of duplicate oversampling technique, the number of images grew to 46,280. Figure 6 shows the class distribution after balancing.

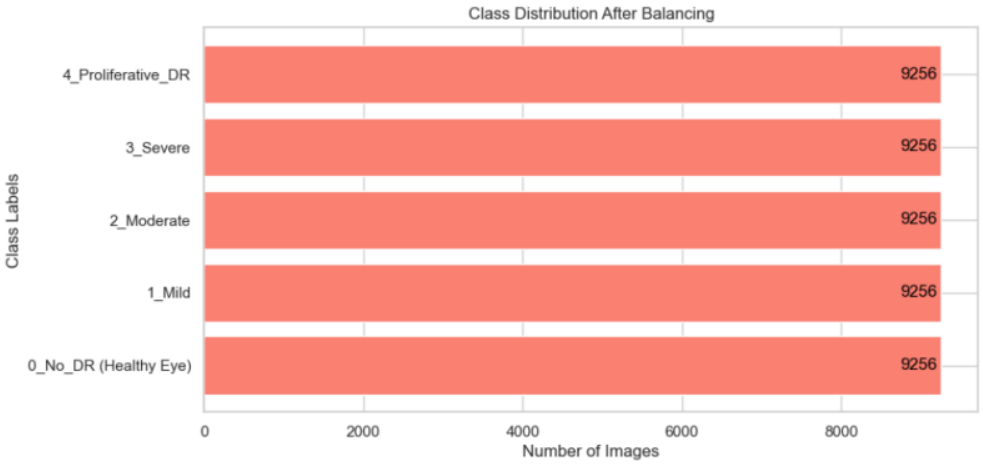


Figure 6: Class Distribution After Balancing.

Data Augmentation
Data augmentation is performed on the fundus images using a variety of methods following pre-processing. By enabling, this enhancement procedure enhances the model's capacity to generalize. These changes include, among other things, zooming, scaling, flipping, rotating, shearing, and moving. Among the improvements utilized in this study are zooming, rotation, and horizontal and vertical flipping. The dataset is subjected to these transformations at random, with a rotation range of 15 degrees and a zooming range of 0.1. The ImageDataGenerator setup provided by a library like Keras describes several data augmentation methods used throughout the pre-processing phase of neural network training. Firstly, dividing each pixel value by 255 renormalizes the images' pixel values range between 0 and 1. This is a typical pre-processing procedure that facilitates quicker and more reliable training. Validation split reserves 10% of the image data for the purpose of generating a validation set. Horizontal flipping randomly flips images along the vertical axis to introduce variability and increase the model's robustness to different orientations. The brightness range randomly varies the brightness of the images within a given range of 0.9 to 1.1. The zoom range enables on the images by a factor of up to 0.1. The rotation range allows for minor rotations of up to 15 degrees in either direction. The width shift range and height shift range, respectively, introduce small random shifts in the horizontal and vertical directions by a factor of up to 0.05 of the total width or height. Finally, fill mode determines how newly created pixels which may result from transformations such as rotation or translation are filled, the 'reflect' mode here fills them by mirroring the values from the image's borders. Table 2 presents the augmentation techniques used.

Table 2: Augmentation techniques used on the dataset.

Name	Value
Rescale	1/255
Validation Split	0.1
Horizontal Flip	True
Rotation Range	15 degrees
Zoom Range	0.1
Brightness Range	[0.9,1.1]

Width Shift Range	0.05
Height Shift Range	0.05
Fill Mode	'reflect'

Dataset Division: Before the model training phase, the enlarged dataset is divided into three categories: training, validation, and testing. The data is split in a 70:10:20 ratio with 70% given to the training set, 10% given to the validation set, and 20% allocated to the testing set. The models are trained and their performance assessed using the testing set. The validation set, on the other hand, helps to assess how well the model behaves during the training phase. Mostly employed for visual image processing, a deep learning architecture called convolutional neural network (CNN) consists of nested layers that learn hierarchical features from data. Convolutional layers using learnable filters often come next with ReLU to add non-linearity and pooling layers reducing the feature maps in dimensionality while maintaining important information. As shown in figure 7, [55], which performs the final classification or regression based on the learned high-level representations, the extracted and compressed features are passed through one or more fully linked layers at the finish.

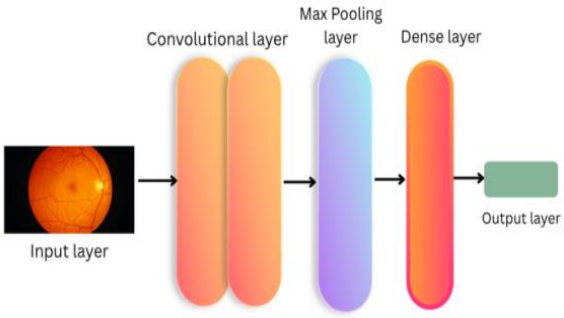


Figure 7: Simple CNN Architecture

Xception, is a deep convolutional neural network architecture that notably advances the Inception concept by completely replacing traditional inception modules with depthwise separable convolutions. This novel approach effectively separates the spatial convolution and channel-wise convolution steps into distinct operations, leading to a highly efficient model. By decoupling these processes, Xception achieves a marked drop in computational cost and the number of parameters, often while delivering competitive or even superior performance compared to conventional CNNs, as shown in Figure 8.[53]

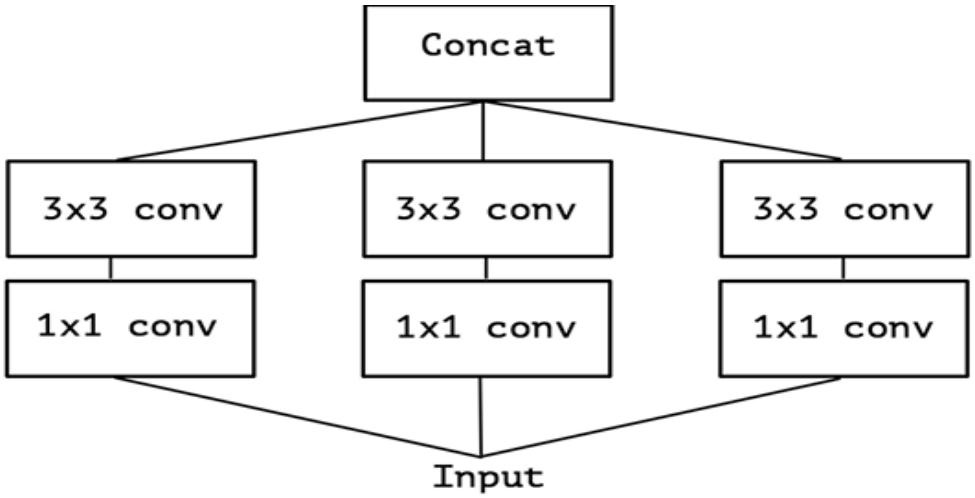


Figure 8: Xception Model Architecture

InceptionV3 is a widely known Convolutional Neural Network (CNN) architecture, part of Google's Inception family, designed for high performance in image recognition tasks. Its core innovation lies in the "Inception modules," which parallelize convolutions with different pooling operations and filter sizes e.g., 1x1, 3x3, 5x5, concatenating their outputs to capture features at various scales simultaneously. A significant

improvement in InceptionV3 is the extensive use of factorization, breaking down larger 2D convolutions into smaller, asymmetric ones to reduce computational cost and parameter count while maintaining expressiveness, making it an efficient and highly accurate model often used for transfer learning, as shown in Figure 9.[54]

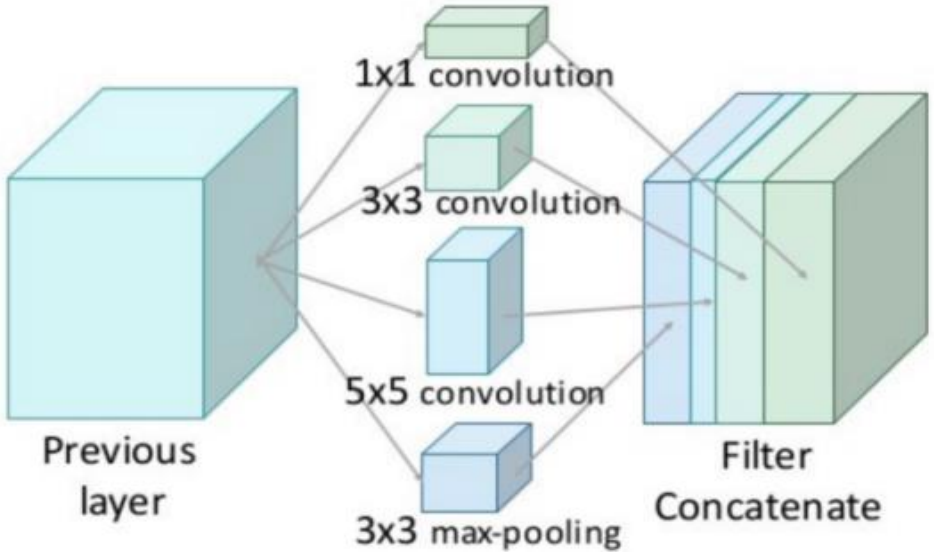


Figure 9: InceptionV3 Model Architecture

DenseNet, or Dense Convolutional Network, is a deep learning architecture for CNNs introduced "Densely Connected Convolutional Networks". This revolutionary architecture promotes extensive feature reuse by ensuring each layer processes feature maps from all former layers within its dense block, thereby strengthening feature propagation and concatenation while leading to more compact models that efficiently address challenges like vanishing gradients and parameter efficiency, as shown in Figure 10.

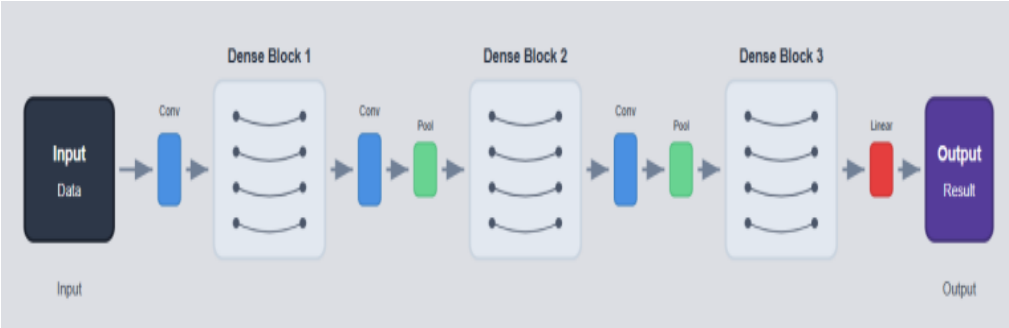


Figure 10: Densenet121 Model Architecture

IV. METHODOLOGY

Our research suggested a binary and multi-class classification system for five separate grades. Three pre-trained models of Convolutional Neural Network (CNN):

- Xception,
- InceptionV3,
- DenseNet121 defines this framework.

These models are trained independently using transfer learning on a fused dataset comprising four benchmark datasets:

- APTOS 2019,
- IDRiD,
- Messidor-2,
- DDR.

It presents a multi-class classification framework meant to categorize diabetic retinopathy into five different levels as well as a binary classification. One can reach this goal by merging our baseline datasets and producing an ensemble of three already trained CNN models.

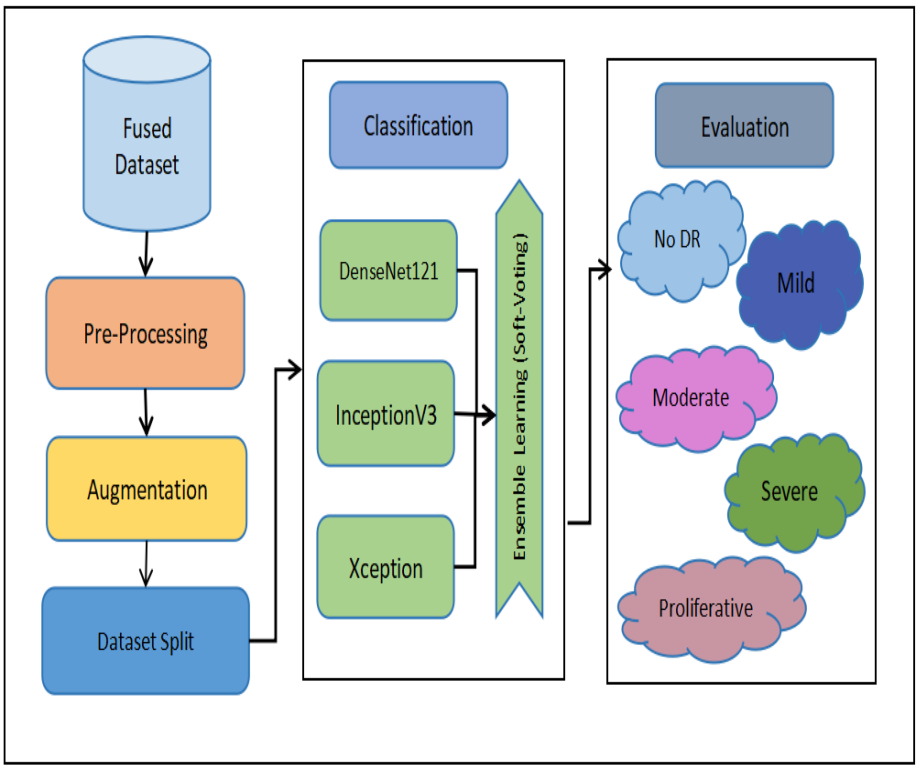


Figure 11: Proposed ensemble framework.

The models leverage transfer learning by importing ImageNet weights, with the last 100 layers unfrozen to fine-tune on the DR-specific dataset. To enhance model generalization and reduce overfitting, the input images are resized to 299 by 299 pixels and normalized using a custom ImageDataGenerator pipeline combining random horizontal flips, brightness corrections, rotations, and spatial shifts. Each model follows a unified classification head architecture added to the base model. This head has a Global Average Pooling layer (GAP Layer) followed by dense layers with 512 and 256 neurons respectively. Corresponding to the five DR severity categories, the last classification layer is a dense layer with softmax activation and output neurons, as shown in Figure 12.

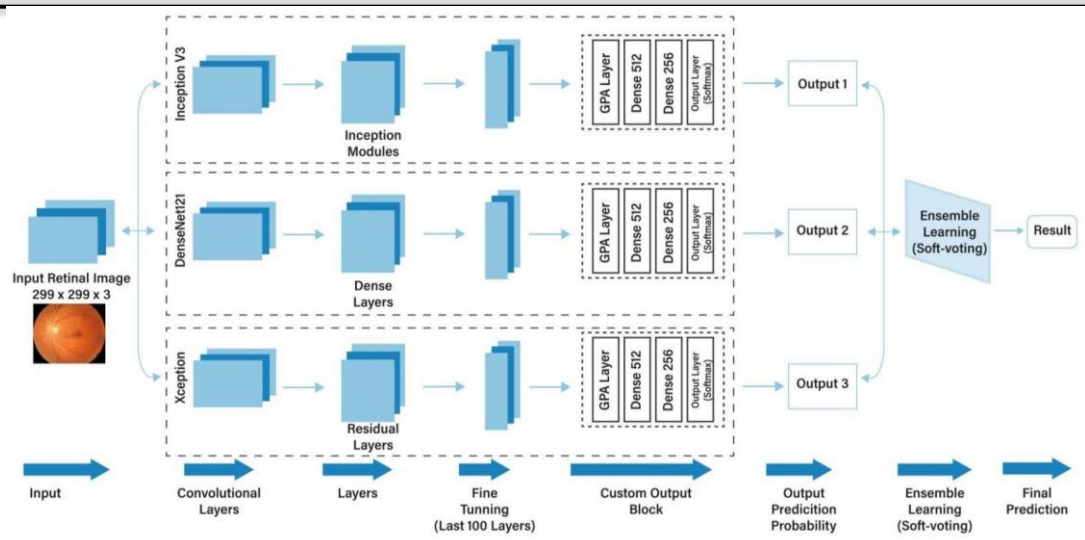


Figure 12: Architecture of the pre-trained models along with fine-tuned layers.

The reason behind using an ensemble approach with soft voting is to leverage the shared advantage of different CNN architectures, leading to more accurate and robust DR grading compared to relying on a single model. The use of soft voting allows the ensemble to potentially give more importance to the predictions of models that have demonstrated better individual performance. The ensemble model's effectiveness is evaluated using top-k accuracy metrics, which assess the clinical utility of the system by considering whether the true DR severity is among the top-k ranked predictions. Measurement of Performance: Quantitative indicators used to evaluate the correctness and effectiveness of a model's forecasts [20], performance metrics. We used a variety of metrics appropriate for both binary (Harmful vs. Non-Harmful) and multi-class classification to assess how well our diabetic retinopathy classification models performed. For the binary classification task, we determined accuracy. We evaluated recall, which is the ratio of real damaged cases properly discovered out of all predicted harmful cases, as well as accuracy, the proportion of properly identified harmful cases. Precision and recall, the F1-score, was used to offer a balanced indicator of the performance of the model. Furthermore, computed specificity for the proportion of correctly identified non-harmful cases. For the multi-class categorization job, we expanded our assessment to include per-class precision, recall, and F1-score to grasp the model's result on each particular severity level of diabetic retinopathy. To show the distribution of predicted and true classes, we also prefer calculating the confusion matrix. To determine how well the model could differentiate between several degrees of severity, we determined the Area Under the Receiver Operating Characteristic Curve (AUC). Lastly, acknowledging the ordinal character of the illness severity, we additionally provide Top-k accuracy including Top 1, Top 2, and Top-3, which consider a forecast as accurate if the actual class lies within the top one, two, or three predicted classes, respectively. Considering the clinical relevance of both lowering false negatives and knowing performance across all disease stages, these measures were selected to offer a detailed analysis of the models' ability to correctly detect the presence and severity of diabetic retinopathy. To evaluate the trained model's generalizability, all metrics were evaluated on the validation dataset. The measurements employed in this study are as follows from equation (1)-(9).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (5)$$

$$\text{Top-k Accuracy} = \frac{\text{Number of samples where true label is in top k predictions}}{\text{Total number of samples}} \quad (6)$$

$$\text{Top-1 Accuracy} = \frac{\text{Number of samples where true label is the most probable predicted class}}{\text{Total number of samples}} \quad (7)$$

$$\text{Top-2 Accuracy} = \frac{\text{Number of samples where true label is among the top 2 most probable predicted classes}}{\text{Total number of samples}} \quad (8)$$

$$\text{Top-3 Accuracy} = \frac{\text{Number of samples where true label is among the top 3 most probable predicted classes}}{\text{Total number of samples}} \quad (9)$$

V. RESULTS

Experiment Setup: The experiments were performed on a high-performance setup with the following specifications:

- CPU: AMD RYZEN 9 5900X
- GPU: NVIDIA GEFORCE RTX 4080 SUPER 16G VENTUS 3X OC
- Memory: 32 GB RAM

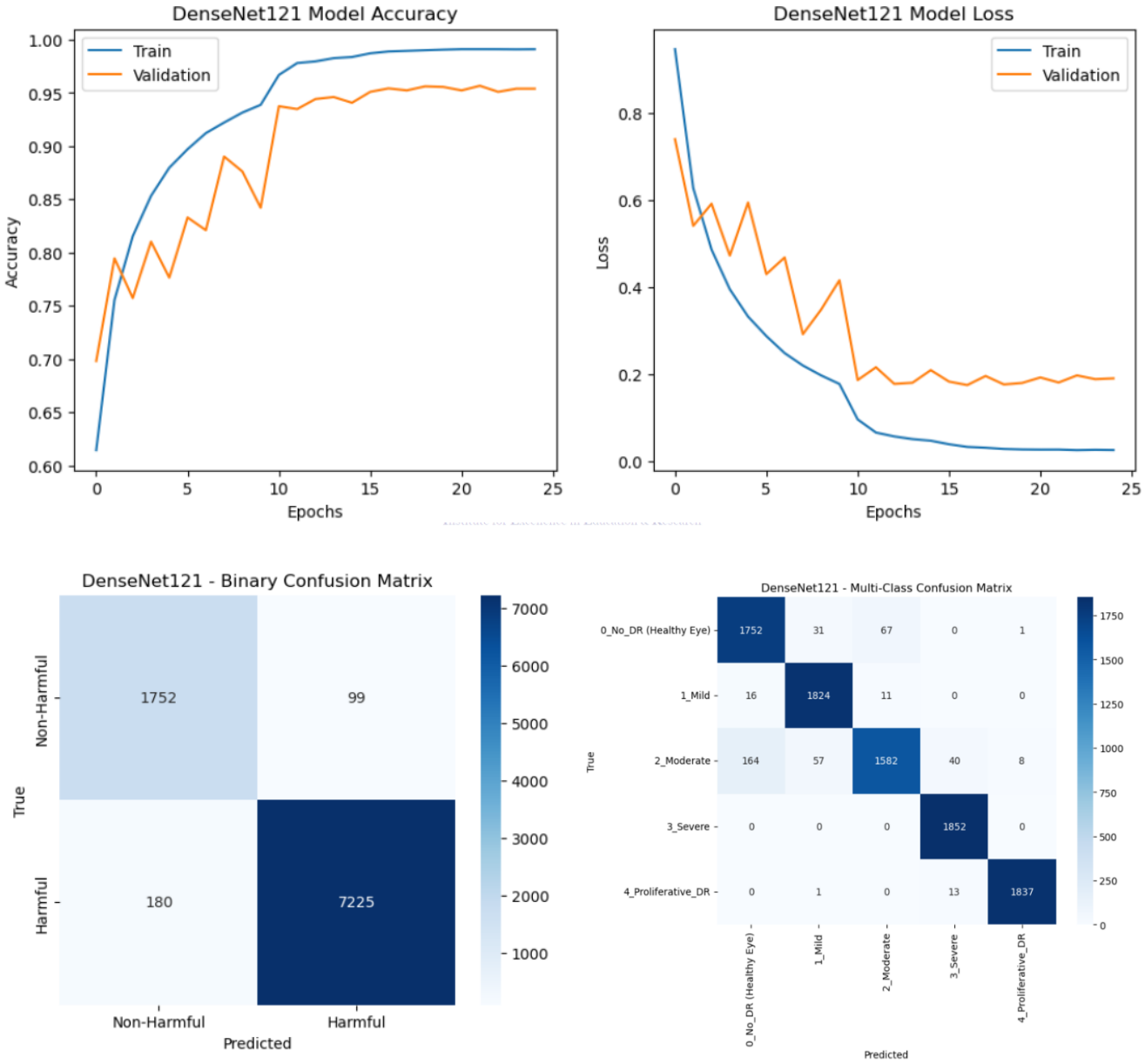
Our study tackles Diabetic Retinopathy detection using an advanced AI system that combines over 18,422 eye images from four different datasets. These images undergo thorough pre-processing, are balanced using oversampling, and are further expanded through augmentation to ensure robust model learning. The system's core is an ensemble of three powerful CNN models Xception, InceptionV3, and DenseNet121, whose individual predictions are intelligently combined using soft voting for a final, highly accurate diagnosis. Developed with Python on Kaggle Notebooks using optimized settings like a 0.0001 learning rate and 32 batch size over 40 epochs as shown in Table 3.

Table 3: Configuration of Training Parameters

Hyperparameter	**Value**
Optimizer	Adam
Learning Rate	0.0001
Batch Size	32
Training Epochs	40
Early Stopping Patience	8
ReduceLROnPlateau Monitor	'val_loss'
ReduceLROnPlateau Factor	0.2
ReduceLROnPlateau Patience	2

Confusion metrices of testing set for each model:
DenseNet121: An efficacy assessment of the DenseNet121 model. The top-left plot depicts that over 25 epochs of the training and validation accuracy, showing consistent progress with validation accuracy reaching approximately 94%, indicating good

generalization. The top-right plot shows the corresponding loss curves, where both training and validation losses steadily decline, suggesting effective learning with minimal overfitting. The bottom-left confusion matrix summarizes the binary classification (non-harmful vs. harmful), showing strong performance with 7,225 correct harmful and 1,752 correct non-harmful predictions. The bottom-center multi-class confusion matrix demonstrates the model's competence to classify all five DR stages accurately, particularly for the more severe stages (e.g., 1,837 correct predictions for proliferative DR). Finally, the bottom-right ROC curve confirms high discriminative power across all classes, with AUC values ranging from 0.99 to 1.00, reflecting the model's exceptional sensitivity and specificity for multi-class DR detection, as shown in Figure 13.



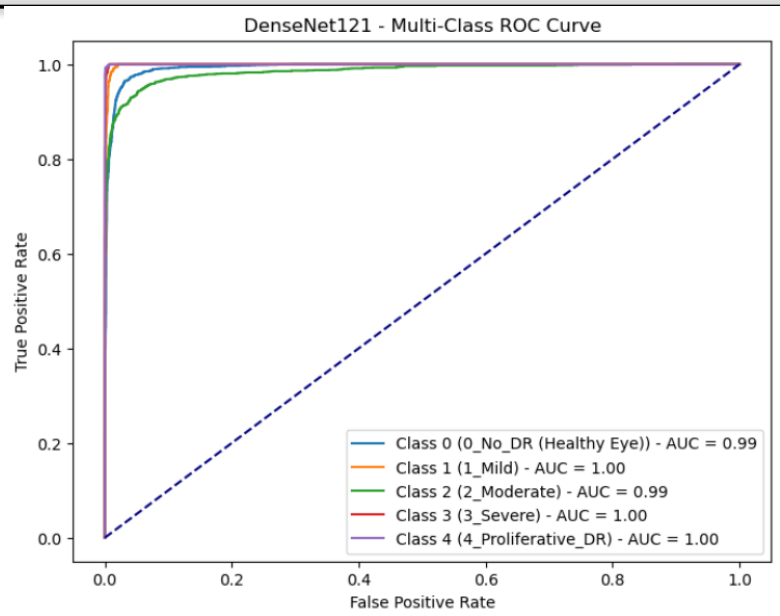
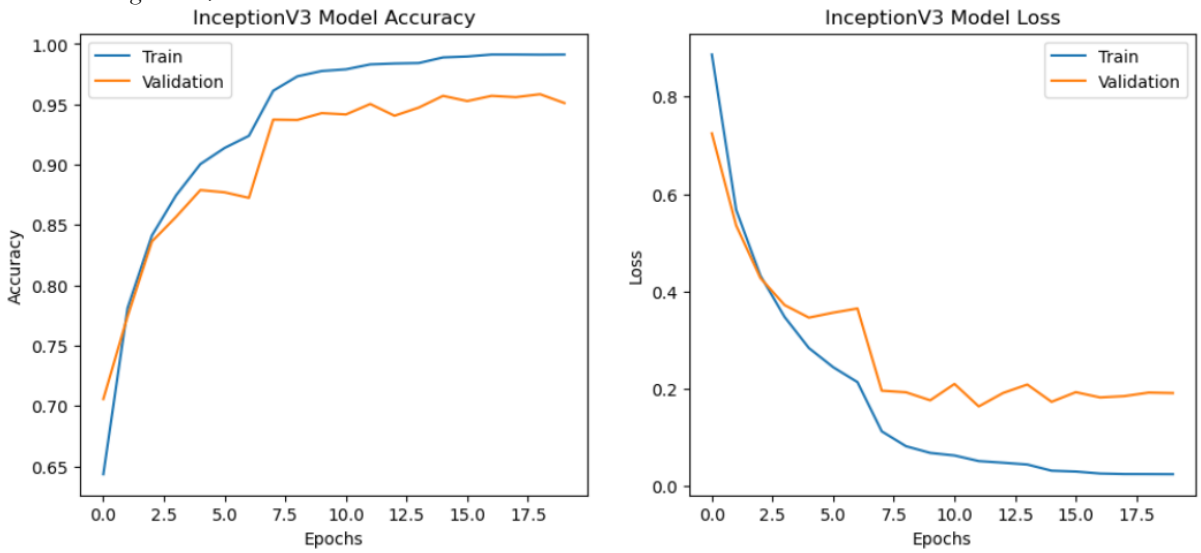


Figure 13: DenseNet121: Model Accuracy, Model Loss, Binary Confusion Matrix, Multi-Class Confusion Matrix, Multiclass ROC Curve

InceptionV3: The performance evaluation of the InceptionV3 model for diabetic retinopathy classification. The top-left plot shows the model's accuracy improving steadily over 18 epochs, with training accuracy nearing 98% and validation accuracy around 94%, indicating strong learning and generalization. The top-right plot illustrates a consistent decline in both training and validation loss, confirming stable model convergence. The binary confusion matrix (bottom-left) reflects excellent discrimination between harmful and non-harmful cases, with 7,212 harmful and 1,728 non-harmful instances correctly classified. The multi-class confusion matrix (bottom-center) demonstrates strong classification performance across all five DR stages, particularly in identifying severe and proliferative DR cases with minimal misclassification. The bottom-right ROC curve further supports this, with AUC values of 0.99 to 1.00 across all classes, underscoring the model's high sensitivity and robustness in multi-class diabetic retinopathy detection, as shown in Figure 14.



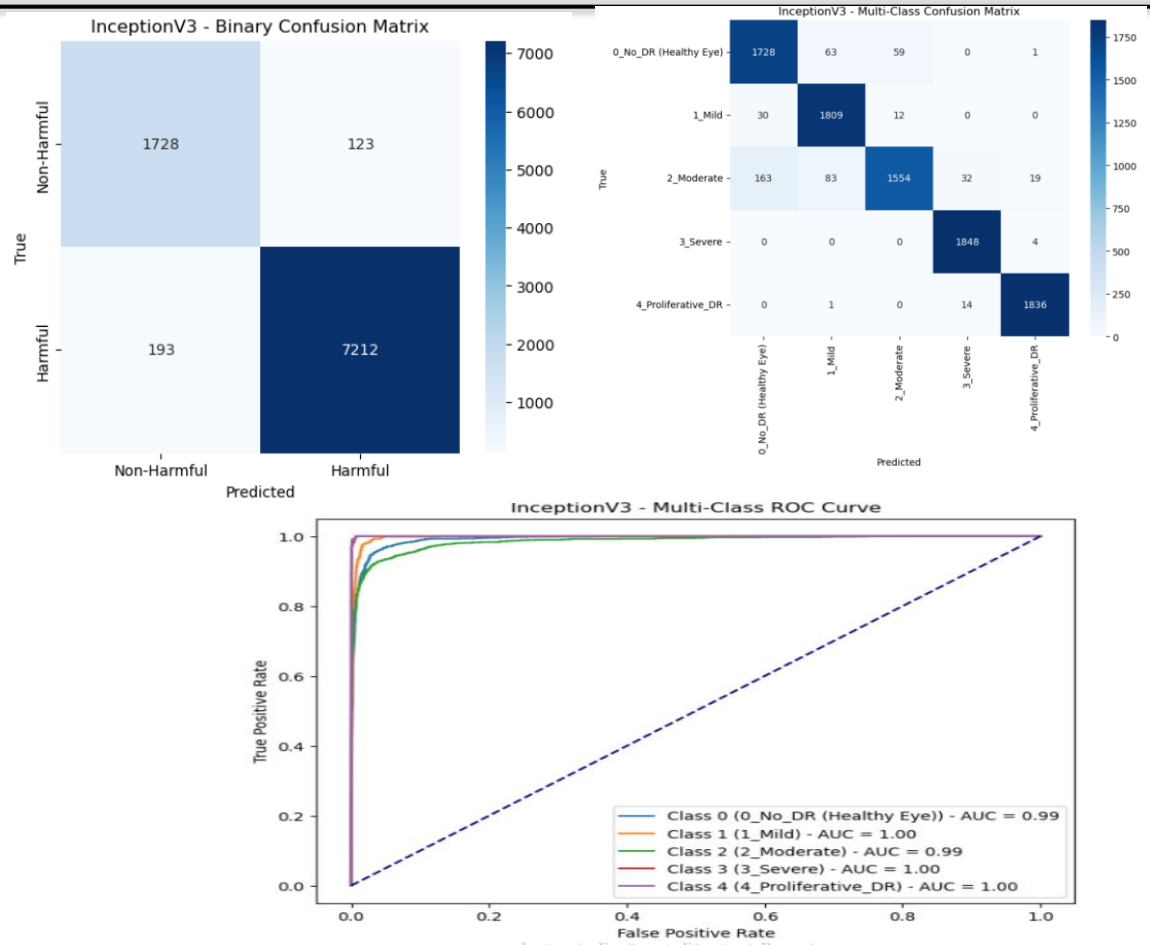
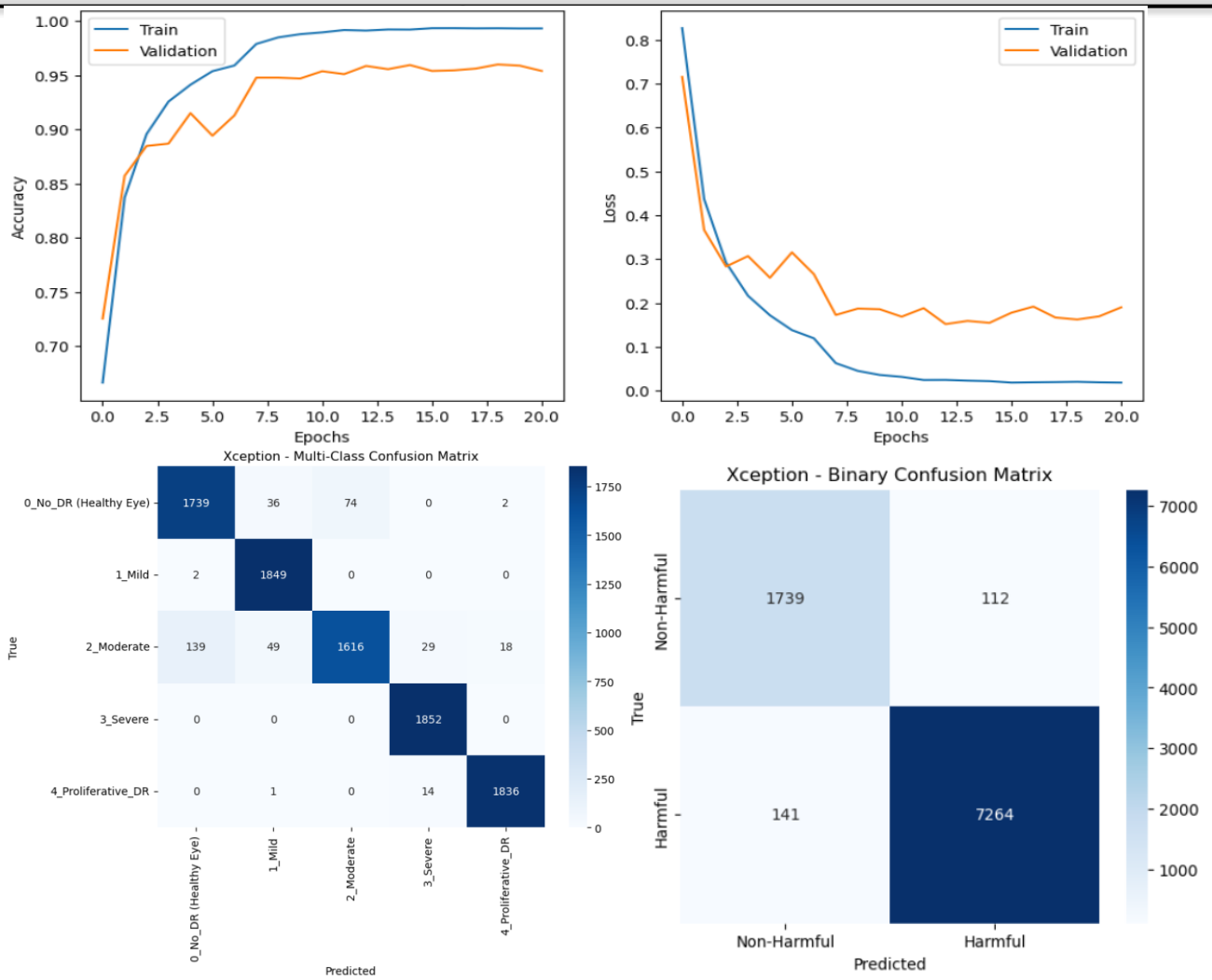


Figure 14: InceptionV3: Model Accuracy, Model Loss, Binary Confusion Matrix, Multi-Class Confusion Matrix, Multiclass ROC Curve

Xception: The training and validation accuracy graphs show steady learning, with the training accuracy reaching nearly 99% and validation accuracy around 96%, while the corresponding loss curves demonstrate effective convergence with decreasing loss across 20 epochs. The binary confusion matrix confirms accurate differentiation between harmful and non-harmful classes, with high true positive (7,264) and true negative (1,739) counts. The multi-class confusion matrix highlights strong predictive performance across all DR severity levels, particularly in correctly identifying moderate, severe, and proliferative DR with minimal misclassification. The ROC curve for multi-class classification further supports model efficacy, with AUC values of 0.99–1.00 for all five classes, validating Xception's high sensitivity and reliability in DR grading, as shown in Figure 15.


Train
Validation
Accuracy
Epochs
Loss
Xception - Multi-Class Confusion Matrix
0_No_DR (Healthy Eye)
1_Mild
2_Moderate
3_Severe
4_Proliferative_DR
True
Predicted
Xception - Binary Confusion Matrix
Non-Harmful
Harmful
1739
112
141
7264

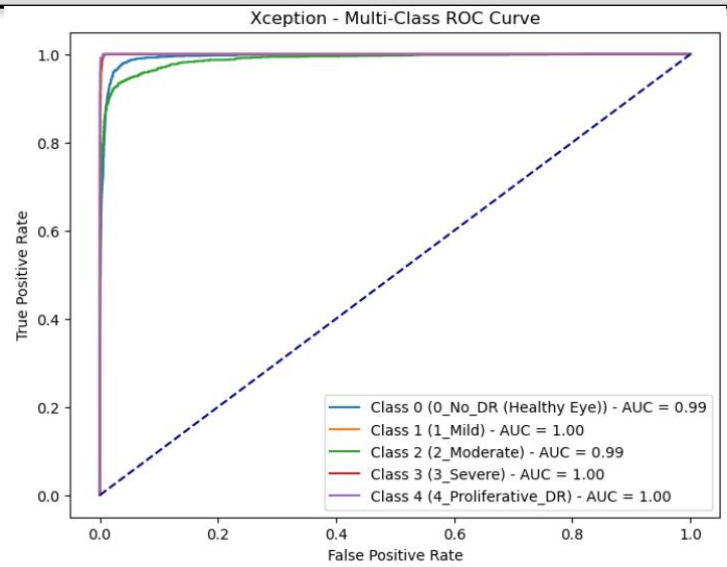
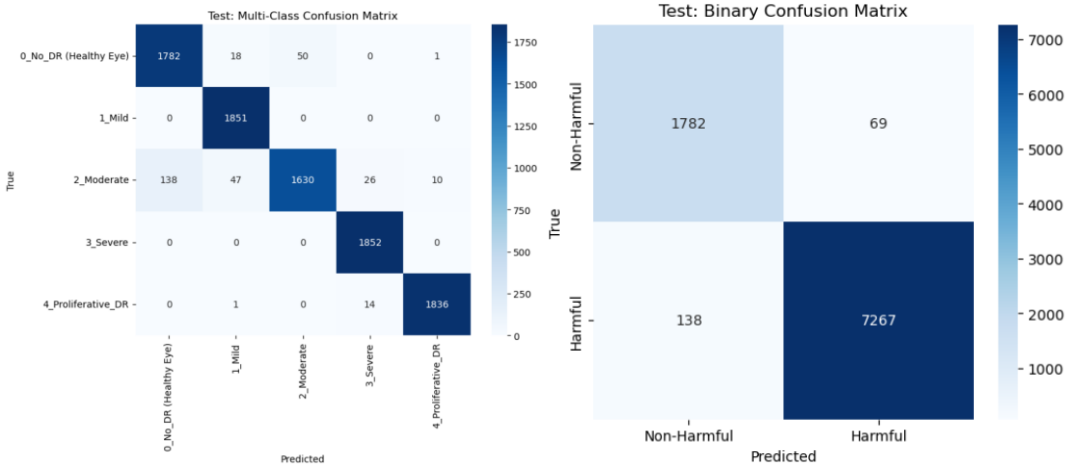


Figure 15: Xception: Model Accuracy, Model Loss, Binary Confusion Matrix, Multi-Class Confusion Matrix, Multiclass ROC Curve

Confusion metrices of testing set for ensemble learning: The testing performance of the ensemble learning model for both binary and multi-class classification. In the binary confusion matrix, the model accurately classified 1782 non-harmful and 7267 Harmful instances, with only 69 False Positives and 138 False Negatives, demonstrating excellent classification capability, particularly for Harmful cases. The multi-class confusion matrix further reflects robust results, with high correct classification counts along the diagonal 1782 for healthy eyes, 1851 for mild DR, 1630 for moderate DR, 1852 for severe DR, and 1836 for proliferative DR while off-diagonal entries represent limited misclassifications, such as 138 moderate DR cases predicted as mild. Finally, the multi-class ROC curves highlight the model's exceptional discriminative power, with AUC scores of 1.00 for four classes (No_DR, Mild, Severe, Proliferative_DR) and 0.99 for Moderate DR, confirming near-perfect classification ability across all stages, as shown in Figure 16.



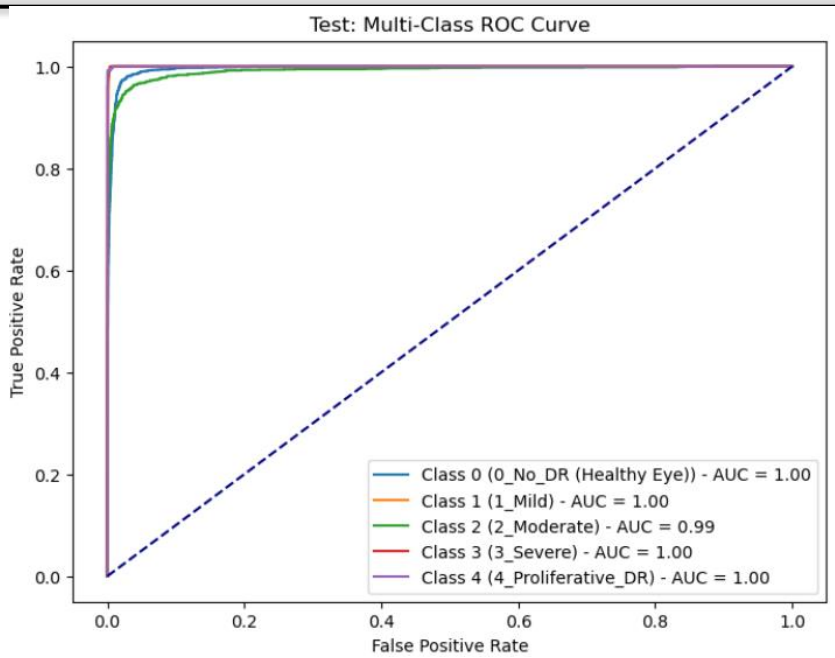


Figure 16: Binary Confusion Matrix, Multi-Class Confusion Matrix, Multi-Class ROC Curve of ensemble learning.

Ensemble Learning Performance: The results clearly validate the effectiveness of the proposed ensemble learning approach. The ensemble model secured a validation accuracy of 97.00%, which is a 1% improvement over the individual best-performing models (Xception and DenseNet121, both at 96.00%). This improvement highlights the benefit of combining the diverse feature representations learned by different CNN architectures. By leveraging the weighted averaging of their prediction probabilities (soft voting), the ensemble model effectively mitigates the individual weaknesses of the constituent models and capitalizes on their strengths, leading to more accurate and robust predictions. Furthermore, the top-k accuracy metrics underscore the clinical utility of the ensemble model. The ensemble achieved a top 1 accuracy of 96.86%, indicating that in nearly 97% of the cases, the model's highest prediction matches the true diabetic retinopathy grade. This high top 1 accuracy is crucial for direct diagnostic assistance to ophthalmologists. The top 2 accuracy of 99.30% further emphasizes the system's reliability. In over 99% of the cases, the correct DR grade is within the top two predictions made by the ensemble. This is particularly valuable in borderline cases where distinguishing between adjacent severity levels can be challenging. Providing the top two likely diagnoses allows ophthalmologists to focus their attention and potentially order further examination if needed, even if the top-1 prediction is not definitively conclusive. Notably, the ensemble model attained a top 3 accuracy of 100.00%. This perfect top 3 accuracy signifies that for every single case in the validation set; the true diabetic retinopathy grade was among the top three predictions made by the ensemble. This exceptionally high top 3 accuracy strongly suggests that the system can reliably narrow down the diagnostic possibilities to a very small set, significantly aiding ophthalmologists, as shown in Figure 17.

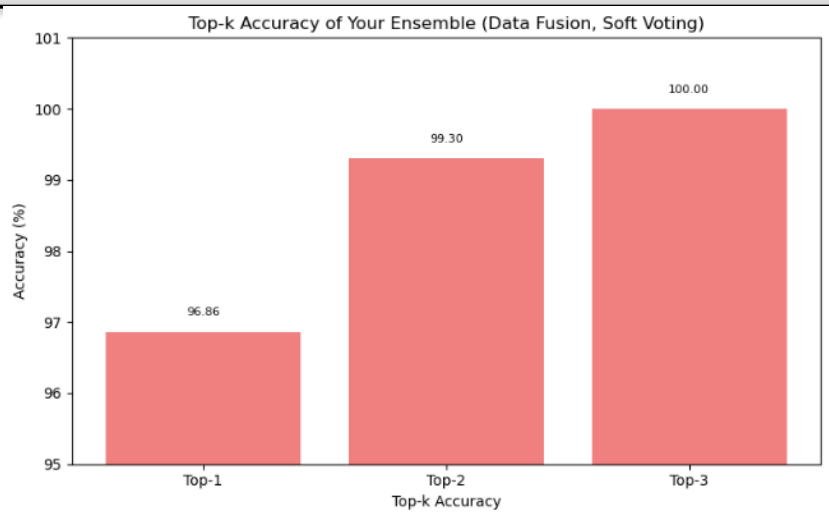


Figure 17: Top-k Accuracy of Ensemble Learning with Soft-Voting.

The outcome evaluation of the proposed ensemble learning model compares it with the performance of its individual constituent models: Xception, DenseNet, and Inception. The model's performance was assessed on the validation dataset, and the key performance metrics, including validation accuracy and top-k accuracies, are in Table 4.

Table 4: Comparative Analysis

Model	Senstivity(%)	Validation Accuracy (%)	Top 1 Accuracy (%)	Top 2 Accuracy (%)	Top 3 Accuracy (%)
Xception	98.10	96.00	95.68	99.27	100.00
DenseNet	97.57	96.00	95.70	99.16	100.00
Inception	97.39	95.00	95.27	99.16	99.95
Ensemble	**98.14**	**97.00**	**96.86**	**99.30**	**100.00**

VI. DISCUSSION

The comparative analysis between Aftab & Akhtar's ensemble approach [2] and our proposed methodology reveals significant insights into the evolution of diabetic retinopathy (DR) classification systems. Aftab & Akhtar [2] achieved remarkable results with their ensemble of EfficientNetB2, DenseNet121, and ResNet50, reporting a test accuracy of 96.96% using three datasets: APTOS 2019, IDRiD, and Messidor-2 with SMOTE for class imbalance and sophisticated preprocessing including CLAHE for contrast enhancement.

Our proposed ensemble, combining Xception, InceptionV3, and DenseNet121, achieved a marginally higher accuracy of 97% using four datasets named as: APTOS, IDRiD, Messidor-2, and DDR with enhanced preprocessing techniques including gamma correction and black background removal. The significance of our approach extends beyond the modest numerical improvement, as the introduction of top-k accuracy metrics provides clinically relevant evaluation with 96.86%, 99.30% and 100.00%, demonstrating exceptional clinical utility where the correct DR severity grade appears among the top two predictions in 99.3% cases. The architectural diversity in our ensemble, combining Xception's depthwise separable convolutions, InceptionV3's multi-scale assessment, and DenseNet121's dense connections, creates a more

diverse feature representation space that potentially provides better generalizability across different imaging conditions and populations. Both studies demonstrate the superiority of ensemble learning strategies over individual model architectures and employ soft voting for prediction aggregation, which preserves confidence levels and enables better uncertainty estimation for clinical decision-making. As Table 5 shows, the convergence toward ensemble learning methods in both studies strengthens the case for architectural variety and multi-model approaches in medical image analysis, so indicating that automatic DR classification has matured enough for clinical assistance uses tackling the global burden.

Table 5: Accuracy of different Deep Learning Models.

Model	Accuracy
EfficientNetB2, DenseNet121, and ResNet50 (ensemble) [2]	96.96%
Proposed Method:Xception + InceptionV3 + DenseNet121 (Ensemble)	**97.0%** **Top 1: 96.86%** **Top 2: 99.30%** **Top 3: 100.00%**

VII. CONCLUSION

Our study presents a extremely effective ensemble deep learning system for grading diabetic retinopathy (DR). By tactically combining four benchmark datasets, applying rigorous preprocessing and data augmentation techniques, and ensembling three powerful CNN architectures (Xception, InceptionV3, and DenseNet121) using soft voting, the system achieves a remarkable validation accuracy of 97%. Furthermore, the distinguished top-k accuracy metrics underscore the clinical utility of this approach in reliably ranking DR severity. The findings strongly suggest that this data fusion and ensemble learning methodology holds real promise as a helpful tool for ophthalmologists, amplifying diagnostic accuracy and facilitating timely interventions to resist vision loss caused by DR across diverse patient populations and image variation.

Funding: No funding was received.

VIII. REFERENCES

Abood, R. H., & Hamad, A. H. (2025). Multi-Label Diabetic Retinopathy Detection Using Transfer Learning Based CNN. Fusion: Practice and Applications, 17(2), 279–293. https://doi.org/10.54216/FPA.170221

Aftab, S., & Akhtar, S. (2025). Diabetic Retinopathy Severity Classification Using Data Fusion and Ensemble Transfer Learning. Journal of Software Engineering and Applications, 18(1), 1–23. https://doi.org/10.4236/jsea.2025.181001

Teo, Z. L., Tham, Y. C., Yu, M., et al. (2022). The global burden of diabetic retinopathy and its projection to 2045. Diabetes Research and Clinical Practice, 183, 109145. https://doi.org/10.1016/j.diabres.2021.109145

Ghosh, S., & Prasad, P. (2025). Multi-Source Retinal Image Fusion for Robust DR Grading Using Deep Ensemble Learning. IEEE Transactions on Medical Imaging. https://doi.org/10.1109/TMI.2025.3259876

Liu, F., Yang, J., & Chen, Z. (2025). Enhancing DR Detection with Augmented Data and Soft-Voting CNN Ensembles. Computers in Biology and Medicine, 160, 106057. https://doi.org/10.1016/j.compbiomed.2025.106057

Li, Y., & Zhang, H. (2025). A DL-Based Model for DR Grading. Scientific Reports, 15, 87171. https://doi.org/10.1038/s41598-025-87171-9

Rajalakshmi, R., et al. (2025). Creating a Retinal Image Database for DR Screening. Scientific Reports, 15, 7853. https://doi.org/10.1038/s41598-025-91941-w

Felorunsho, O., et al. (2025). Explainable Ensemble DL Model for DR. LAUTECH Journal of Engineering and Technology, 19(1), 1–14.

https://laujet.com/index.php/laujet/article/view/774

Khalaf, A. A., & Al-Qurashi, M. A. (2025). DL Empowered Diagnosis of DR. Intelligent Automation and Soft Computing, 40(1), 59314. https://www.techscience.com/iasc/v40n1/59314/html

Husein, M. A., & Pratama, D. (2025). CNN and Oversampling for DR Classification. Journal of Robotics and Control, 6(1), 160–170. https://journal.umy.ac.id/index.php/jrc/article/view/25331

Chen, L., & Wang, X. (2024). Ensemble Deep Learning and EfficientNet for DR. Nature Communications, 15, 81132. https://www.nature.com/articles/s41598-024-81132-4

Liu, H., Yue, K., Cheng, S., Pan, C., Sun, J., & Li, W. (2020). Hybrid Model Structure for Diabetic Retinopathy Classification. Journal of Healthcare Engineering, 2020, Article ID 8840174. https://doi.org/10.1155/2020/8840174

Moussa, S. S., & Nguyen, Q. D. (2024). DR Diagnosis Using ChatGPT and AutoML. Ophthalmology 1 Science, 4, 100495. https://doi.org/10.1016/j.ophsci.2024.100495

Bhulakshmi, D., & Rajput, D. S. (2024). FedDL: Personalized Federated DL for DR. PeerJ Computer Science, 10, e1584. https://doi.org/10.7717/peerj-cs.1584

Arora, L., et al. (2024). Ensemble DL and EfficientNet for DR. (PMC). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11655640/

Sunkari, S., et al. (2024). Refined ResNet18 with Swish for DR Classification. Biomedical Signal Processing and Control, 88, 105630. https://doi.org/10.1016/j.bspc.2023.105630

Al-Dhubaib, M., & Albezzawy, M. (2024). Early DR Detection with DL. Big Data and Cognitive Computing, 5(4), 125. https://www.mdpi.com/2673-2688/5/4/125

Patel, H., & Bhatt, N. (2024). Enhanced DR Detection via Deep Ensemble Models. IJISAE, 12(1s), 4706. https://ijisae.org/index.php/IJISAE/article/view/4706

Ihnaini, B., Akhtar, S., Ahmad, M., & Aftab, S. (2024). Data Fusion Based Ensemble Transfer Learning Approach to Detect Diabetic Retinopathy. 2024 2nd International Conference on Cyber Resilience (ICCR), Dubai, 26–28 February 2024, 1–5. https://doi.org/10.1109/iccr61006.2024.10532998

Akhtar, S., Aftab, S., Ahmad, M., & Ihnaini, B. (2024). A Transfer Learning Based Framework for Diabetic Retinopathy Detection Using Data Fusion. 2024 2nd International Conference on Cyber Resilience (ICCR), Dubai, 26–28 February 2024, 1–5. https://doi.org/10.1109/iccr61006.2024.10533112

Akhtar, S., & Aftab, S. (2024). A Classification Framework for Diabetic Retinopathy Detection Using Transfer Learning. 2024 ICCR, Dubai, 26–28 February 2024, 1–5. https://doi.org/10.1109/iccr61006.2024.10533135

Kaggle DDR Dataset: https://www.kaggle.com/datasets/mariaherrerot/ddrdataset

Kaggle APTOS 2019 Dataset: https://www.kaggle.com/competitions/aptos2019-blindness-detection/data

IDRID: https://ieee-dataport.org/open-access/indian-diabetic-retinopathy-image-dataset-idrid

Messsidor-2:https://www.adcis.net/en/third-party/messidor2/

For Messidor-2 Label: https://www.kaggle.com/datasets/google-brain/messidor2-dr-grades

Akhtar, S., Aftab, S., Ahmad, M., & Akhtar, A. (2024). Diabetic Retinopathy Severity Grading Using Transfer Learning Techniques. International Journal of Engineering and Manufacturing, 14, 41–53.

https://doi.org/10.5815/ijem.2024.06.04
Al-Dhubaib, M., & Albezzawy, M. (2023). Segmented Vessel Images for DR Diagnosis. Diagnostics, 13(9), 1616. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10184939/
Singh, D., & Verma, P. (2024). Grading DR Using an Ensemble of Self-Supervised CNNs. (ResearchGate). https://www.researchgate.net/publication/379505953
Zhang, X., et al. (2023). Wrapped Approach Using Unlabeled Data for DR Diagnosis. Sensors. https://doi.org/10.3390/app13031901
Jawed, R., et al. (2024). DL-Based Identification and Categorization of DR. IJIST, 6(2), 772–784. https://www.researchgate.net/publication/382085768
Kotiyal, B., & Pathak, H. (2022). DR Binary Classification Using Pyspark. IJMEMS, 7, 624–642. https://doi.org/10.33889/ijmems.2022.7.5.041
Abbas, S., Latif, M., & Ashraf, S. (2022). Survey on DL-Based DR Classification. PMC. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9914068/
Al-Ahmadi, I. G. (2023). Effect of Data Balancing on Ensemble Networks for DR. SIGMA, 41, 10–22. https://sigma.yildiz.edu.tr/storage/upload/pdfs/1727182233-en.pdf
Rahman, M. T., & Dola, A. (2021). Automated DR Grading using DenseNet-169. EICT 2021. https://ieeexplore.ieee.org/abstract/document/9733431
Martinez-Murcia, F. J., Ortiz, A., Ramírez, J., Górriz, J. M., & Cruz, R. (2021). Deep Residual Transfer Learning for Automatic Diagnosis and Grading of Diabetic Retinopathy. Neurocomputing, 452, 424–434. https://doi.org/10.1016/j.neucom.2020.04.148
Le, D., Alam, M., Yao, C. K., Lim, J. I., Hsieh, Y., Chan, R. V. P., et al. (2020). Transfer Learning for Automated OCTA Detection of Diabetic Retinopathy. TVST, 9, 35. https://doi.org/10.1167/tvst.9.2.35
Kassani, S. H., Kassani, P. H., Khazaeinezhad, R., Wesolowski, M. J., Schneider, K. A., & Deters, R. (2019). Diabetic Retinopathy Classification Using a Modified Xception Architecture. 2019 IEEE ISSPIT. https://doi.org/10.1109/isspit47144.2019.9001846
Thota, N. B., & Umma Reddy, D. (2020). Improving DR Severity Classification with Transfer Learning. 2020 IEEE MWSCAS. https://doi.org/10.1109/mwscas48704.2020.9184473
Gangwar, A. K., & Ravi, V. (2020). DR Detection Using Transfer Learning and Deep Learning. Evolution in Computational Intelligence, Springer. https://doi.org/10.1007/978-981-15-5788-0_64
Hagos, M. T., & Kant, S. (2019). Transfer Learning Based Detection of DR from Small Dataset. arXiv:1905.07203. https://doi.org/10.48550/arXiv.1905.07203
Bhardwaj, C., Jain, S., & Sood, M. (2021). Transfer Learning Based Robust Automatic Detection System for DR Grading. Neural Computing and Applications, 33, 13999–14019. https://doi.org/10.1007/s00521-021-06042-2
Vaibhavi, P. M., & Manjesh, R. (2021). Binary Classification of DR Detection and Web Application. IJRESM, 4, 142–145. https://journal.ijresm.com/index.php/ijresm/article/view/1000
El Houby, E. M. F. (2021). Using Transfer Learning for DR Stage Classification. Applied Computing and Informatics. https://doi.org/10.1108/aci-07-2021-0191
Aswathi, T., Swapna, T. R., & Padmavathi, S. (2021). Transfer Learning Approach for Grading of DR. Journal of Physics: Conference Series, 1767, 012033. https://doi.org/10.1088/1742-6596/1767/1/012033
Ghazal, M., Ali, S. S., Mahmoud, A. H., Shalaby, A. M., & El-Baz, A. (2020). Accurate Detection of Non-Proliferative DR in OCT Images Using CNNs. IEEE Access, 8,

34387–34397. https://doi.org/10.1109/access.2020.2974158

Bora, A., Balasubramanian, S., Babenko, B., Virmani, S., Venugopalan, S., Mitani, A., et al. (2021). Predicting the Risk of Developing DR Using Deep Learning. The Lancet Digital Health, 3, e10–e19. https://doi.org/10.1016/S2589-7500(20)30250-8

Gao, Z., Pan, X., Shao, J., Jiang, X., Su, Z., Jin, K., et al. (2022). Automatic Interpretation and Clinical Evaluation for DR Using Deep Learning. British Journal of Ophthalmology, 107, 1852–1858. https://doi.org/10.1136/bjo-2022-321472

Elsharkawy, M., Sharafeldeen, A., Soliman, A., Khalifa, F., Ghazal, M., El-Daydamony, E., et al. (2022). A Novel CAD System for Early Detection of DR Using 3D-OCT. Diagnostics, 12, 461. https://doi.org/10.3390/diagnostics12020461

Alyoubi, W. L., Abulkhair, M. F., & Shalash, W. M. (2021). DR Fundus Image Classification and Lesions Localization Using DL. Sensors, 21, 3704. https://doi.org/10.3390/s21113704

Chilukoti, S. V., Maida, A. S., & Hei, X. (2022). DR Detection Using Transfer Learning from Pre-Trained CNN Models. TechRxiv. https://dx.doi.org/10.36227/techrxiv.18515357.v1

Shaban, M., Ogur, Z., Mahmoud, A., Switala, A., Shalaby, A., Abu Khalifeh, H., et al. (2020). A CNN for Screening and Staging of DR. PLOS ONE, 15, e0233514. https://doi.org/10.1371/journal.pone.0233514

https://medium.com/@kdk199604/xception-deep-learnings-leap-beyond-inception-05a708c205f9

Xception model: https://iq.opengenus.org/xception-model/

https://www.geeksforgeeks.org/machine-learning/introduction-convolution-neural-network/