Blindness Detection

Project Description

The aim of this project is to detect early signs of blindness caused by Diabetic Retinopathy (DR). Diabetic Retinopathy is an eye disease that can lead to vision loss and blindness in individuals with diabetes. It occurs when high blood sugar levels damage the blood vessels in the retina — the light-sensitive tissue located at the back of the eye.

In its early stages, Diabetic Retinopathy often shows no noticeable symptoms, but as the condition progresses, patients may experience blurry vision, floaters, or even complete vision loss. Diagnosis is typically made through a dilated eye examination, and depending on the severity, treatments may include injections, laser therapy, or surgery. (*Click Here for more information on Diabetic Retinopathy*)

Dataset Description

The dataset used for blindness detection was sourced from a Kaggle competition held in 2019. Each image in the dataset was evaluated by medical professionals and assigned a severity score for Diabetic Retinopathy (DR) on a scale from 0 to 4, where:

- 0 – No DR (Healthy)
- 1 – Mild DR
- 2 – Moderate DR
- 3 – Severe DR
- 4 – Proliferative DR (Most advanced stage)

The dataset includes 3,662 training images and 1,928 testing images. Additionally, it provides:

- A training CSV file containing image filenames with their corresponding DR severity labels.
- A test CSV file containing only the image identifiers for evaluation.

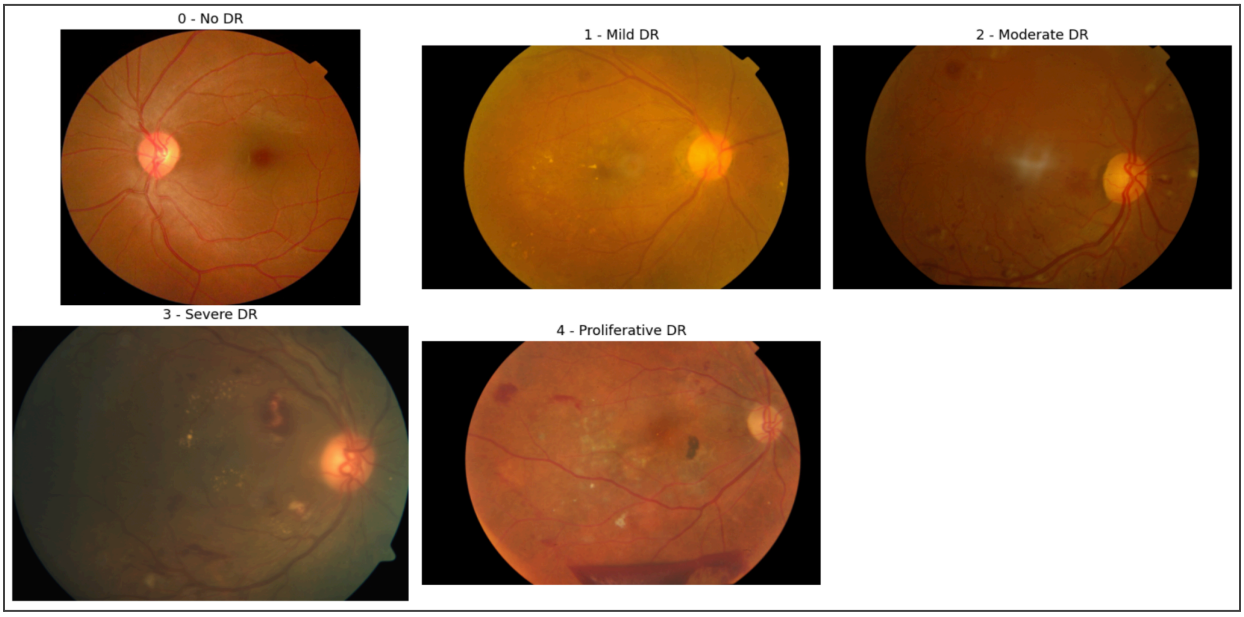
Data Exploration and Preprocessing

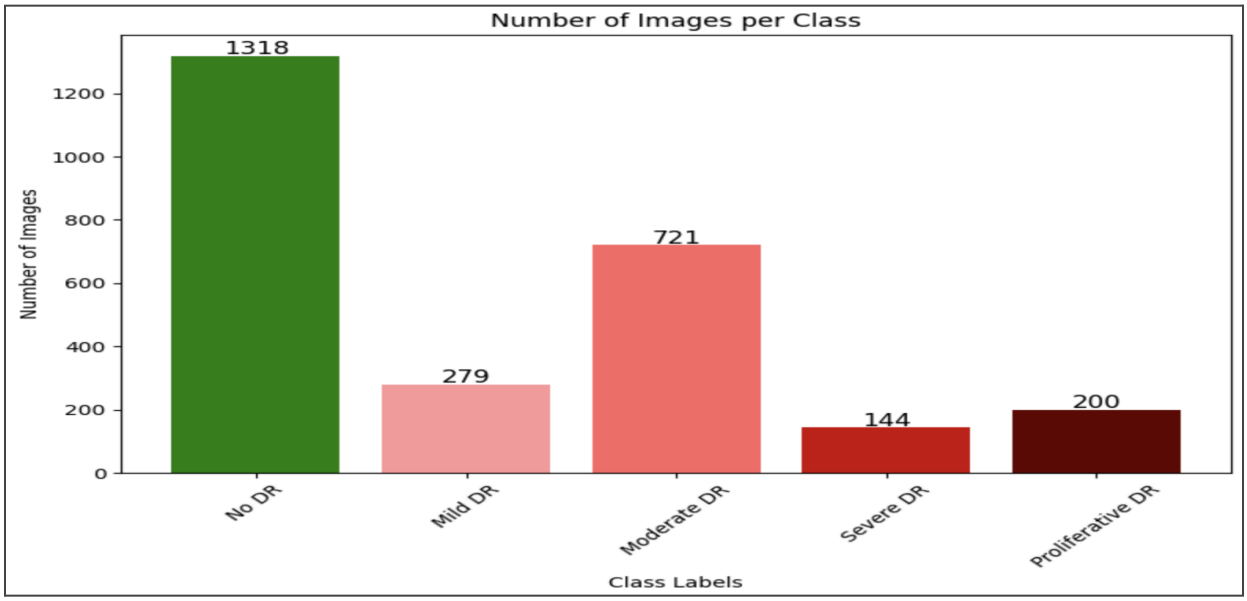
Since this dataset was part of a Kaggle competition, the test labels were not provided by the organizers. Therefore, the original labeled training dataset was divided into three subsets:

- Training set: 2,662 images

- Validation set: 200 images
- Test set: 800 images

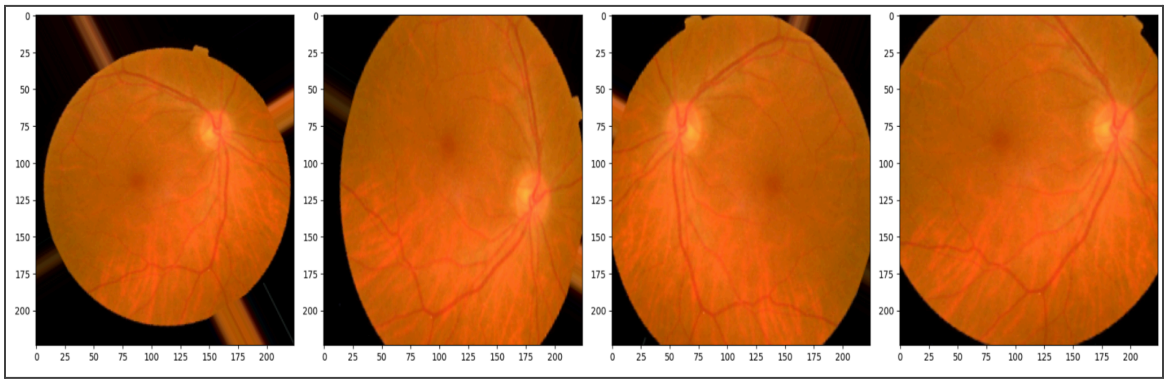
During data exploration, it was observed that the image sizes were inconsistent, varying from 640×480 to 4288×2848 pixels. Additionally, the dataset showed a clear class imbalance — the majority of samples belonged to the Healthy (Label 0) and Moderate (Label 2) categories, while Severe (Label 3) and Proliferative (Label 4) cases were underrepresented.





To address these challenges, data augmentation techniques were applied. All images were resized to 224×224 pixels for uniformity, and various transformations were used to increase data diversity, including rotation, zooming (in/out), horizontal flipping, width and height shifts, and shearing (tilting).

Finally, for all images across the training, validation, and test sets, the pixel values were normalized to a range of 0–1 to ensure faster and more stable model training.



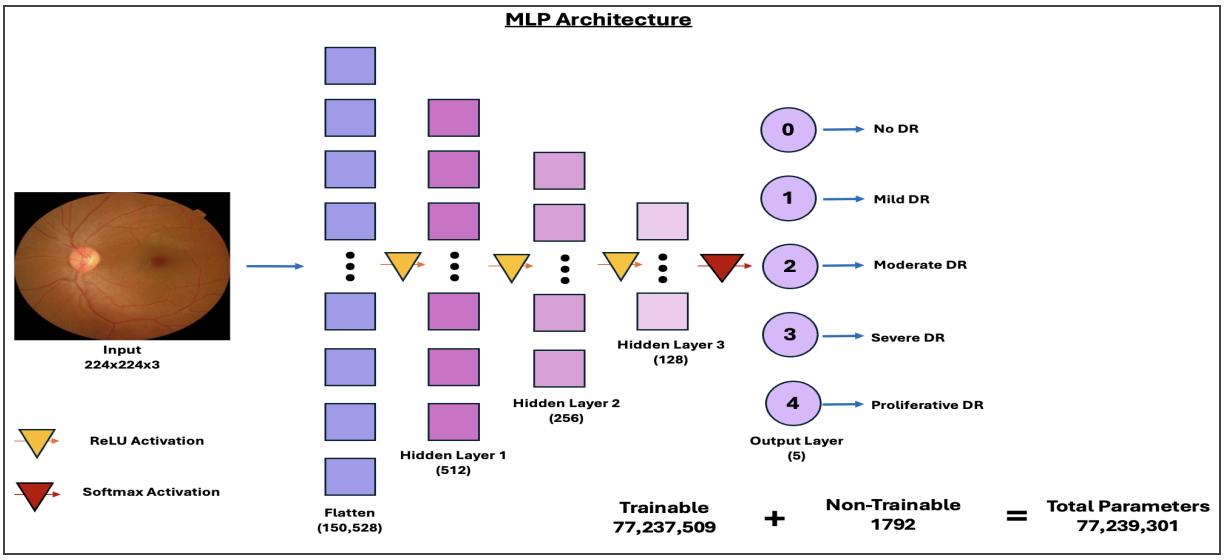
Model Workflow Experiment

To identify the optimal model for blindness detection using retinal images, various Artificial Neural Network (ANN) architectures were examined. The experimentation commenced with a traditional Multi-Layer Perceptron (MLP) model, followed by Convolutional Neural Networks (CNNs), and finally advanced to transfer learning approaches utilizing pre-trained models.

[1] Multi-Layer Perceptron (MLP) Model

Model Architecture:

The Multi-Layer Perceptron (MLP) model was built by flattening each image into a one-dimensional vector, which was then passed to an input layer with 150,528 nodes (calculated as 224 × 224 × 3, where 224×224 represents image height and width, and 3 corresponds to the RGB color channels). The network included three hidden layers with 512, 256, and 128 neurons, respectively, followed by an output layer with 5 nodes, representing the five Diabetic Retinopathy classes.



Using this MLP architecture, approximately 77 million parameters (including weights and biases) were trained. The detailed calculation is shown in the table below:

Layers	Output Shape	Param #	Calculation
Flatten_1 (Flatten)	(None, 150528)	0	224*224*3 = 150528
Dense_1 (Dense)	(None, 512)	**77,070,848**	512 * (150,528 + 1)
batch_normalization	(None, 512)	2048	**(2*512)** + **(2*512)**
dropout (Dropout)	(None, 512)	0	
dense_2 (Dense)	(None, 256)	**131,328**	256 * (512 + 1)
Batch_normalization_1 (BatchNormalization)	(None, 256)	1024	**(2*256)** + **(2*256)**
dropout_1 (Dropout)	(None, 256)	0	
dense_3 (Dense)	(None, 128)	**32,896**	128 * (256 + 1)
Batch_normalization_2 (BatchNormalization)	(None, 128)	512	**(2*128)** + **(2*128)**
dropout_2 (Dropout)	(None, 128)	0	
dense_4 (Dense)	(None, 5)	**645**	5 * (128 + 1)

Note: Batch Normalization: 2 * (gamma & beta) → **Trainable Params** | 2 * (moving average & variance) → **Non-trainable Params**

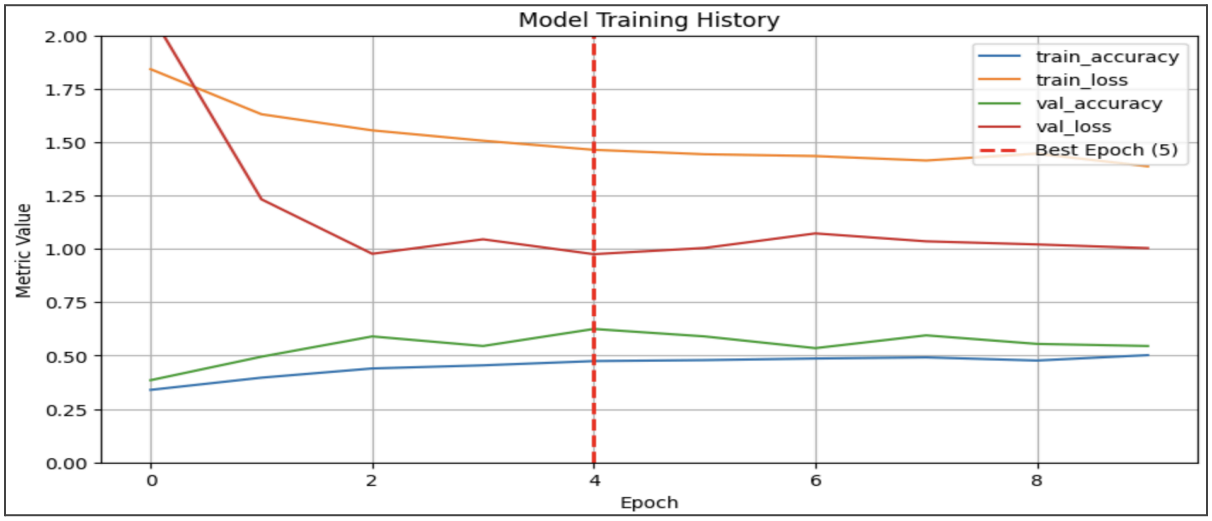
77,237,509 (Trainable params) + 1792 (Non-trainable params) = 77,239,301 (Total Parameters)

To improve training stability and reduce overfitting, *Batch Normalization* and a 20% *Dropout* were applied after each hidden layer. The *ReLU* activation function was used in all hidden layers, while the *Softmax* activation was applied to the output layer to handle the multi-class classification task.

The model was trained using the *Adam* optimizer with a *learning rate* of 0.001, and *Categorical Crossentropy* was chosen as the loss function. To address the class imbalance, *balanced* class weights were applied during training. Additionally, *early stopping* was implemented to prevent overfitting and ensure optimal model performance.

Model Performance:

The below chart compares the Training Vs Validation Loss and Accuracy curve analysis of MLP model.

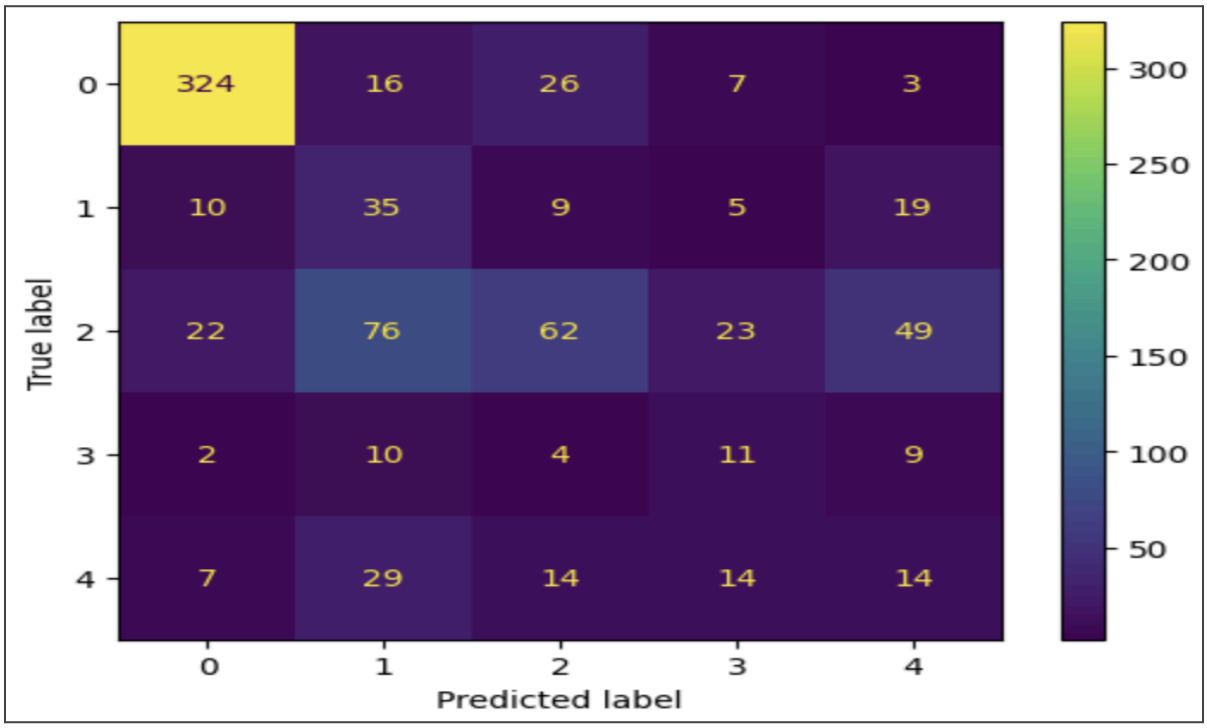


The X-axis represents the number of epochs, while the Y-axis shows the corresponding loss and accuracy values. The chart shows that the MLP model learns during the first few epochs, with both training and validation accuracy improving and validation loss reaching its best point around *Epoch 5*, where the model generalizes the most effectively. After this point, validation loss begins to rise and validation accuracy fluctuates, indicating early overfitting, even though training accuracy continues to increase. This pattern is expected because an MLP on raw 224×224×3 images has very high dimensional input and cannot capture spatial features well. Overall, the model performs best around Epoch 5 but struggles afterward, suggesting that CNNs or transfer learning would be more suitable for this image classification task.

Train Accuracy	Validation Accuracy	Test Accuracy
47.45%	62.50%	55.75%

From the above table, the MLP model achieves 47.45% training accuracy, which indicates that it is not fitting the training data very well — a sign of underfitting. Interestingly, the validation accuracy is higher at 62.50%, which can happen when dropout, batch normalization, and class weights act as regularizers, making the model behave more conservatively on the training data. The test accuracy (55.75%) falls between the two, indicating moderate generalization but confirming that the MLP struggles to learn strong image features.

Confusion Matrix:

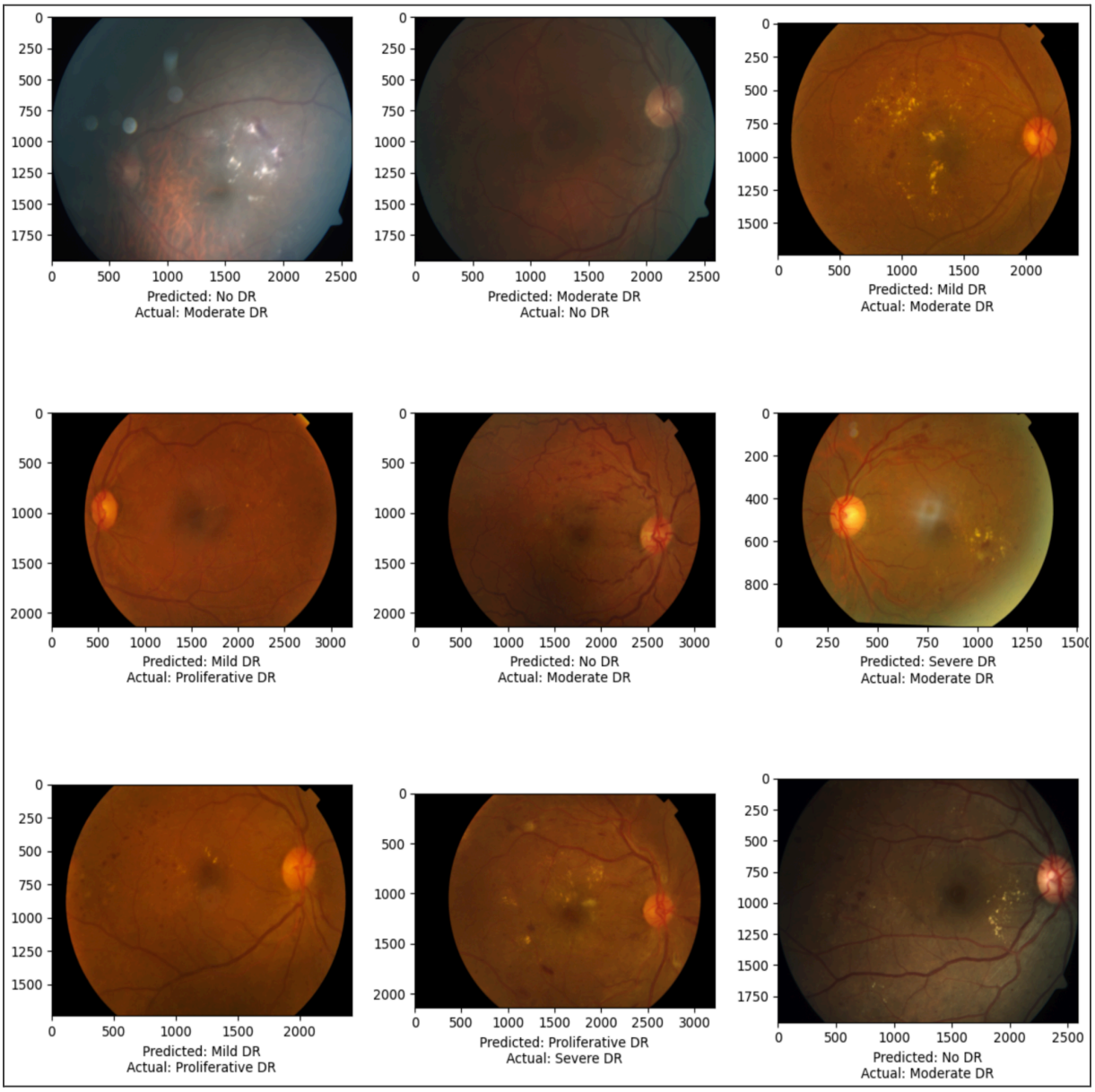


The confusion matrix provides a summary of how many retinal images the MLP model classified correctly or incorrectly for each category. Here's a detailed interpretation for each class:

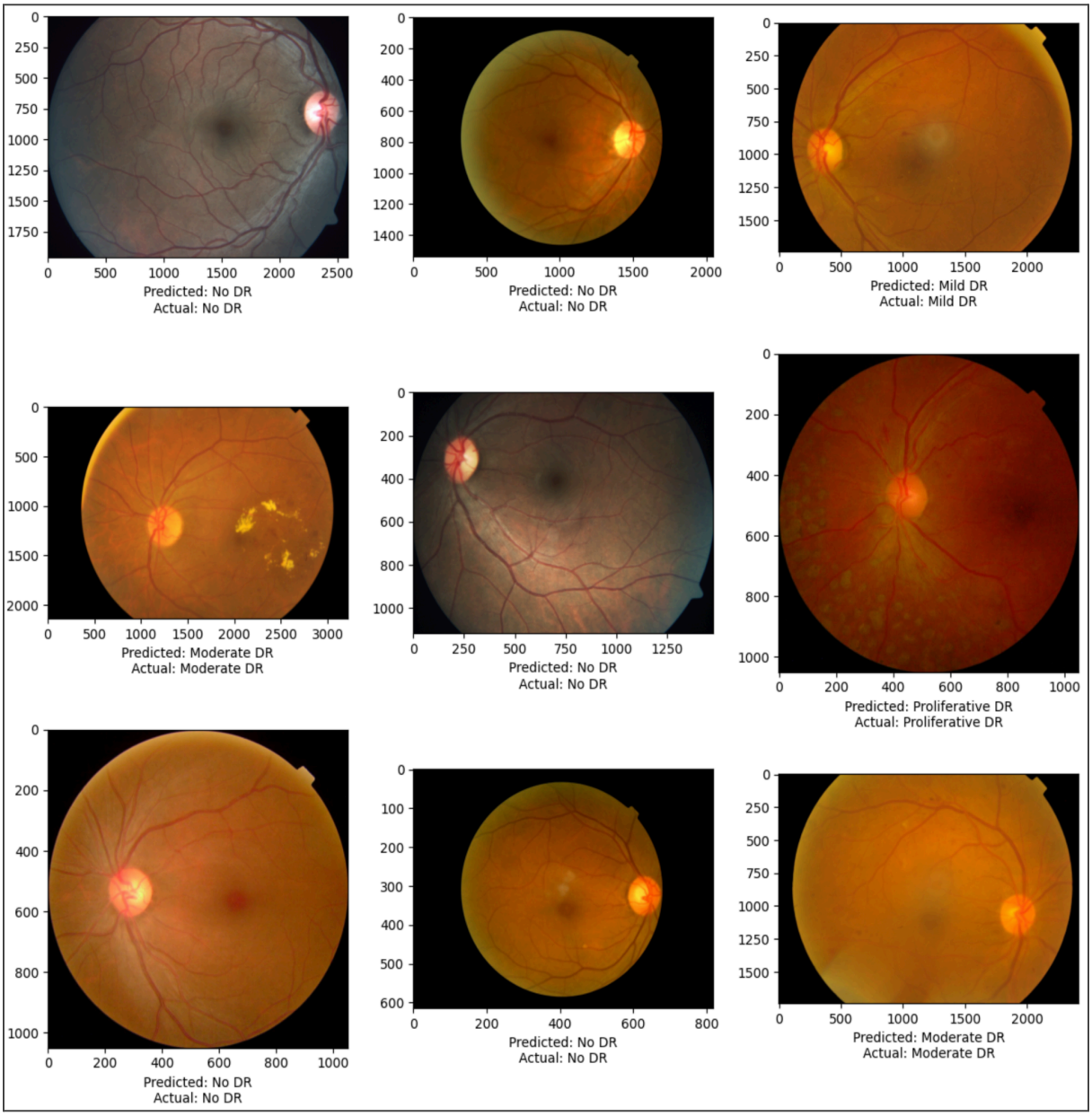
- **Class 0 (No DR)**: The model correctly identified 324 healthy images, with 52 incorrectly predicted as one of the DR classes.
- **Class 1 (Mild DR)**: It accurately classified 35 Mild DR cases, while 43 were assigned to other categories.

- **Class 2 (Moderate DR)**: The model detected 62 Moderate DR images, but misclassified 22 as No DR and 76 as Mild DR, indicating a tendency to underestimate disease severity.
- **Class 3 (Severe DR)**: Only 11 Severe DR cases were correctly identified, and 16 were misclassified as milder stages, which is problematic for clinical use.
- **Class 4 (Proliferative DR)**: The model correctly predicted 14 cases, while 64 were misclassified across all remaining classes—including 7 incorrectly labeled as healthy—creating a substantial risk of missing advanced disease.

Visuals of Misclassification by MLP:



Visuals of Correct classification by MLP:



Conclusion:

Overall, the model's training history and confusion matrix show that the MLP is not effective for diabetic retinopathy classification. It underfits the training data, overfits early, and fails to learn strong image features. While it performs reasonably on healthy images, it frequently misclassifies all DR stages—especially Moderate, Severe, and Proliferative cases—often

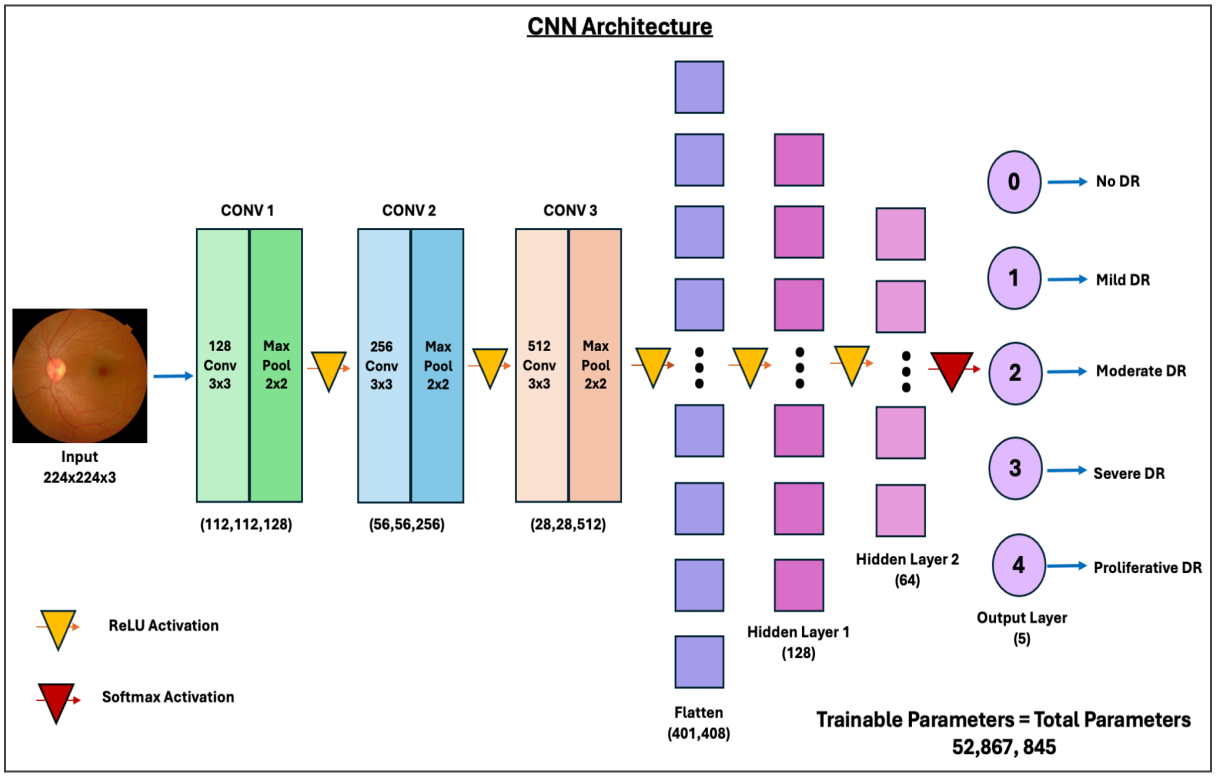
predicting them as milder or even healthy. This makes the model unreliable for clinical use and confirms that a CNN approach is needed for better accuracy and safer predictions.

[2] Convolutional Neural Network (CNN) Model

Following the MLP, the next step was to implement a CNN model, which is widely regarded as the most effective architecture for image classification tasks.

Model Architecture:

The CNN model was built from scratch with three convolutional blocks of 128, 256, and 512 filters of size 3 x 3, each followed by max pooling of size 2 x 2 to progressively extract spatial features. The flattened feature maps are passed through two fully connected layers of 128 and 64 neurons with ReLU activation, and an output layer with 5 nodes with softmax activation to predict class probabilities.



In contrast to the traditional MLP model, this architecture involves roughly 53 million trainable parameters (Weights and Bias), which is about 24 million fewer parameters. A breakdown of these calculations is provided in the table below.

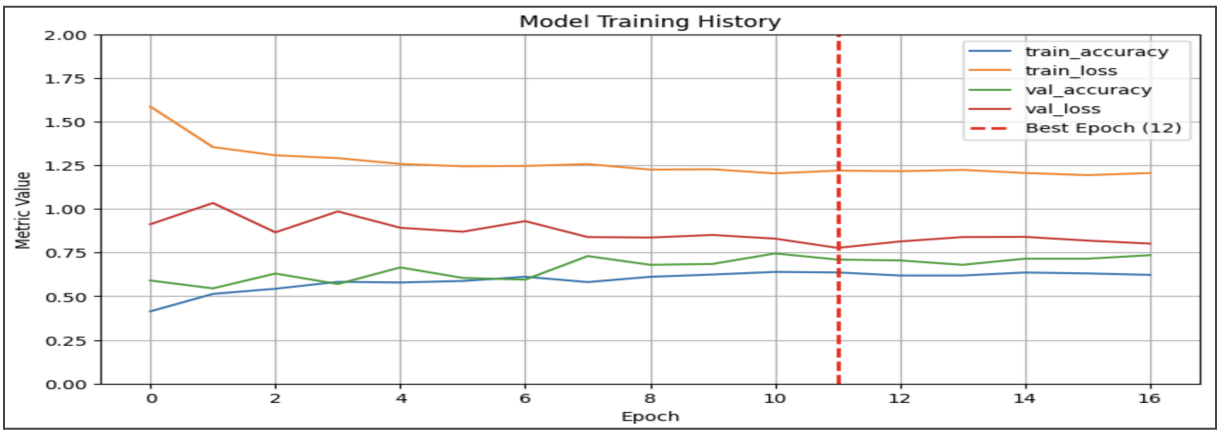
Layers	Output Shape	Param #	Calculation
Conv_1 (Conv2D)	(None, 224,224,128)	**3,584**	128 * [(3*3*3) + 1]
Pool_1 (MaxPooling2D)	(None, 112,112,128)	**0**	
Conv_2 (Conv2D)	(None, 112,112,256)	**295,168**	256 * [(3*3*128) + 1]
Pool_2 (MaxPooling2D)	(None, 56,56,256)	**0**	
Conv_3 (Conv2D)	(None, 56,56,512)	**1,180,160**	512 * [(3*3*256) + 1]
Pool_3 (MaxPooling2D)	(None, 28,28, 512)	**0**	
Flatten (Flatten)	(None, 401408)	**0**	
Dense (Dense)	(None, 128)	**51,380,352**	128 * (401408 + 1)
Dense_1 (Dense)	(None, 64)	**8,256**	64 * (128 + 1)
Dense_2 (Dense)	(None, 5)	**325**	5 * (64 + 1)

52,867,845 (Trainable params) + 0 (Non-trainable params) = 52,867,845 (Total Parameters)

The ReLU activation function was used throughout the network—from the first convolutional block up to the second fully connected layer—while Softmax was applied only at the output layer. All other components, including the optimizer, loss function, and balanced class weights, were kept consistent with the MLP model.

Model Performance:

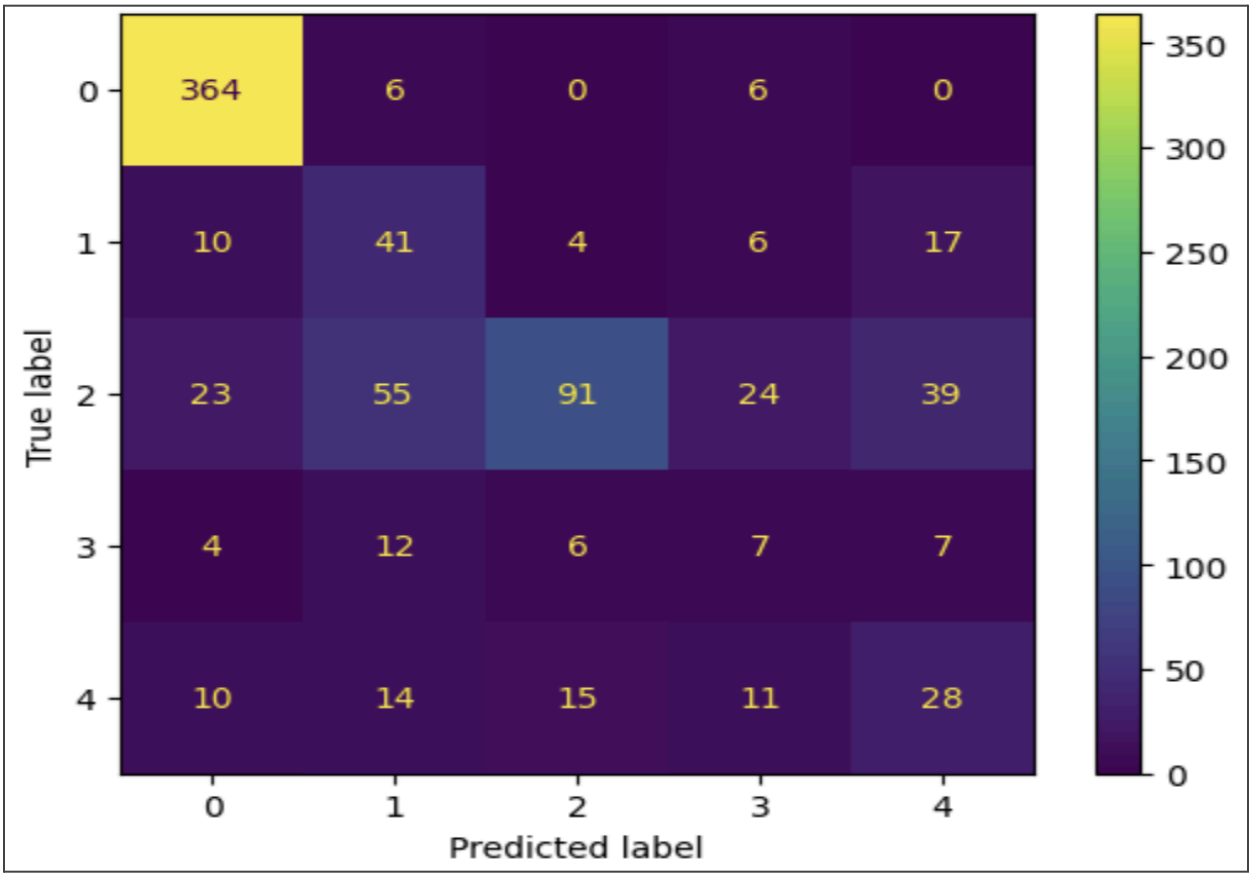
The below chart compares the Training Vs Validation Loss and Accuracy curve analysis of CNN model.



Train Accuracy	Validation Accuracy	Test Accuracy
63.64%	71.00%	66.37%

The CNN model, with 52.87 million parameters—fewer than the 77 million in the previous MLP—demonstrated effective learning and generalization. Training ran for 17 epochs and was stopped by Early Stopping (patience=5, monitoring val_loss). Optimal performance occurred at Epoch 12 with the lowest Validation Loss of 0.7769. Final accuracies were 63.64% (train), 71.00% (validation), and 66.37% (test). Training Accuracy remained below Validation Accuracy, as expected due to data augmentation introducing varied inputs each epoch, supporting better generalization. Compared to the MLP, the CNN achieved higher validation performance with fewer parameters, highlighting its efficiency and superior ability to generalize on unseen data.

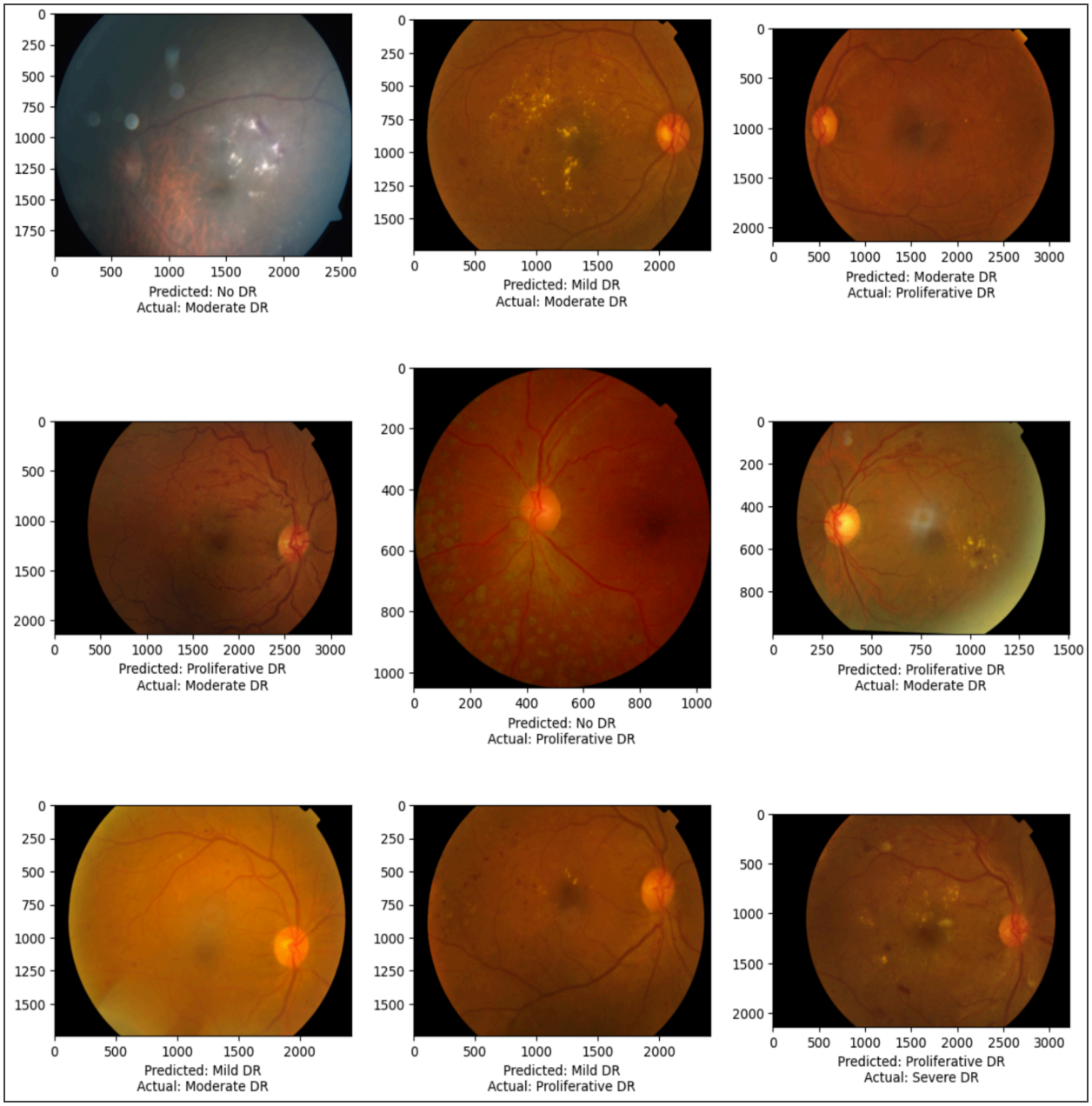
Confusion Matrix:



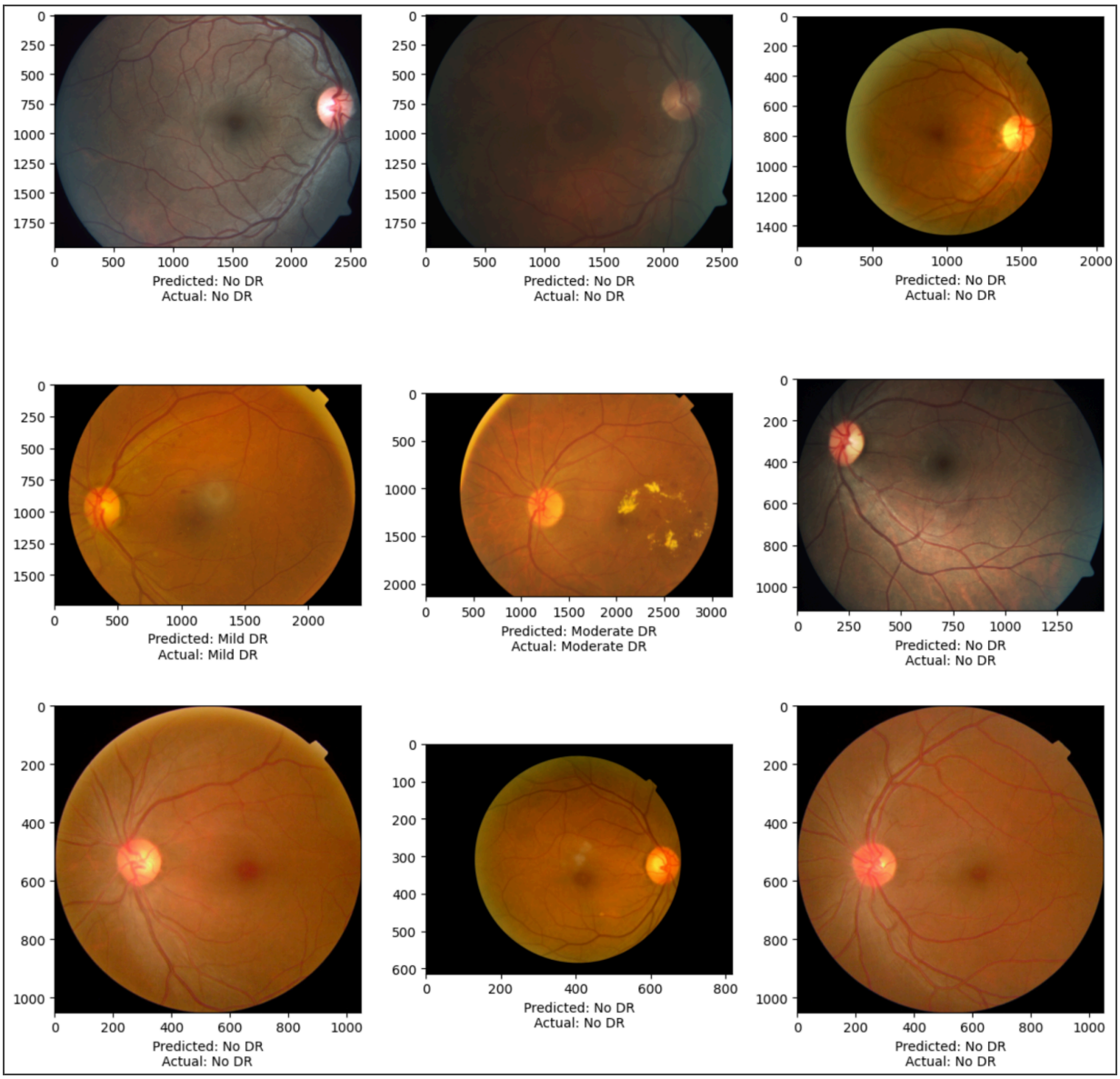
An interpretation of the CNN model's confusion matrix, compared with the previous MLP model, is provided below:

- Class 0 (No DR):
    - Compared to the MLP, the CNN model correctly classified 40 additional images.
    - Only 12 retinal images were incorrectly labeled as Mild DR or Severe DR, indicating high specificity for healthy cases.
- Class 1 (Mild DR):
    - The CNN classified 41 Mild DR images correctly, which is an improvement of 6 images over the previous MLP model.
    - The total number of misclassifications for other classes dropped by 6, decreasing from 43 to 37, demonstrating a noticeable improvement in accuracy for this class.
- Class 2 (Moderate DR):
    - The model's performance improved, correctly identifying approximately 47% more Moderate DR cases compared to the earlier model.
    - There was also a 20% reduction in the tendency to misclassify Moderate DR patients as either healthy or Mild DR, indicating better separation between these classes.
- Class 3 (Severe DR):
    - The number of correctly classified Severe DR images dropped from 11 (in the previous model) to 7 with the CNN.
    - Misclassifications of Severe DR as healthy, Mild, or Moderate cases rose from 16 to 22, underlining an increased risk of missing this severe condition.
- Class 4 (Proliferative DR):
    - The CNN model correctly recognized twice as many Proliferative DR cases as the previous model, indicating substantial improvement.
    - While the overall number of misclassifications fell from 64 to 50, there was an increase of 3 cases where patients were incorrectly classified as healthy, which deserves careful consideration.

Visuals of Misclassification by CNN:



Visuals of Correct classification by CNN:



Conclusion:

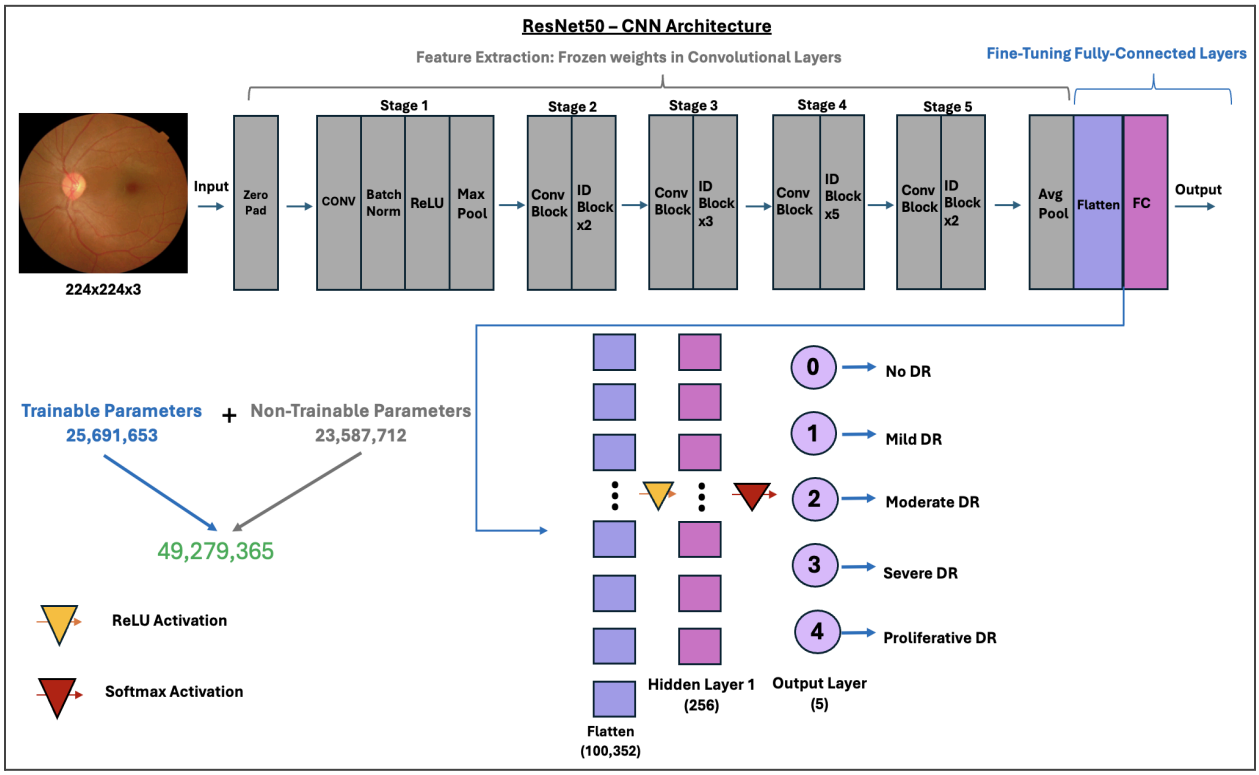
Compared to the previous MLP model, the CNN achieved better generalization with fewer parameters (52.87M vs 77.24M). The final Test Accuracy improved to 66.37%, higher than that of the MLP, indicating better performance on unseen data. Class-wise analysis revealed notable improvements in detecting No DR, Mild DR, Moderate DR, and Proliferative DR cases, with reduced misclassifications compared to the MLP. However, performance for Severe DR declined, highlighting a higher risk of misclassification for this class. To further enhance performance, especially for challenging cases like Severe DR, a pre-trained model, ResNet50, was explored next.

[3] ResNet50-Based Convolutional Neural Network with Transfer Learning

The final step was to implement a Convolutional Neural Network using Transfer Learning. Transfer Learning allows a model to apply knowledge learned from a large, diverse dataset—such as ImageNet—to a smaller, domain-specific dataset like retinal images. Given the limited number of available samples in this project, using a pre-trained network is highly advantageous, as it provides strong feature representations from the outset and reduces the risk of overfitting. Among the many CNN architectures trained on ImageNet, the ResNet50 model was selected and fine-tuned for the diabetic retinopathy detection task.

It should be noted that when using any pre-trained model, the input images must be preprocessed to match the format and scaling applied during the model's original training.

Model Architecture:



The ResNet50 architecture consists of 50 layers that incorporate skip (shortcut) connections, allowing the network to bypass certain layers and pass the output of one layer directly to a deeper layer. These residual connections effectively resolve the degradation problem that typically occurs in very deep networks.

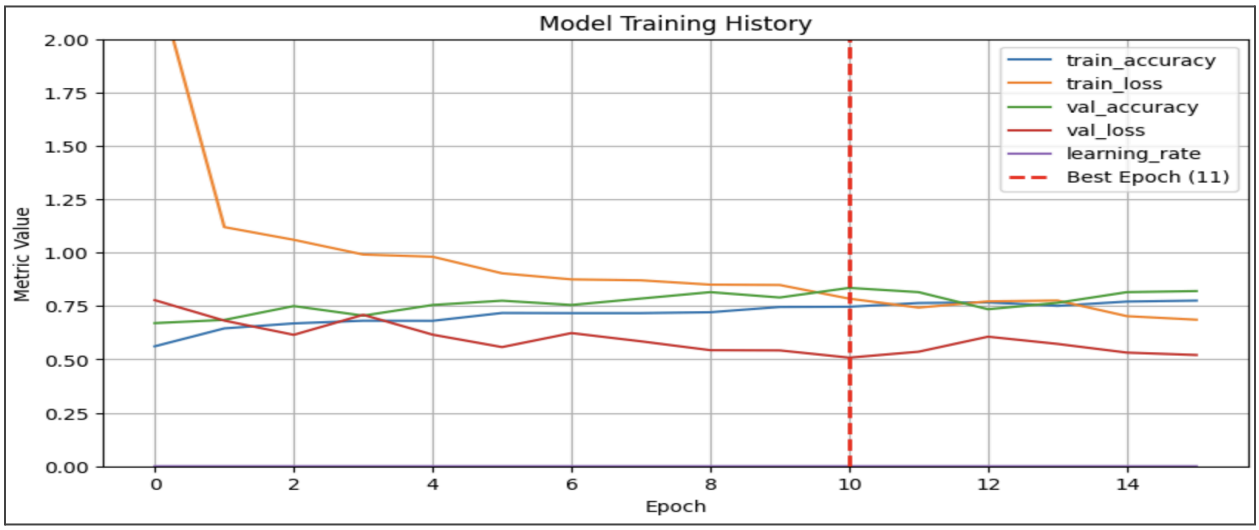
In this project, ResNet50 was used in a ***feature extraction*** setup, where all convolutional layers were frozen and only the fully connected layers were trained. After flattening the extracted feature maps, the model produced 100,352 features, which were passed through a single hidden layer with 256 units, followed by a 5-class output layer.

Under this configuration, the model contained ***25.69 million trainable parameters*** (from the newly added dense layers) and ***23.59 million non-trainable parameters*** (from the frozen ResNet50 base), resulting in a total of ***49.28 million parameters.*** The total trainable parameters is significantly lower than both the custom CNN and the earlier MLP models.

The ResNet50 feature-extraction model was compiled using the Adam optimizer, categorical cross-entropy loss, and accuracy as the performance metric. Class imbalance was addressed using class weights, and training stability was ensured through EarlyStopping and ReduceLROnPlateau callbacks. The model was trained for up to 20 epochs using augmented training data and validated on the validation set, allowing it to learn effectively while preventing overfitting and optimizing learning rate adjustments.

Model Performance:

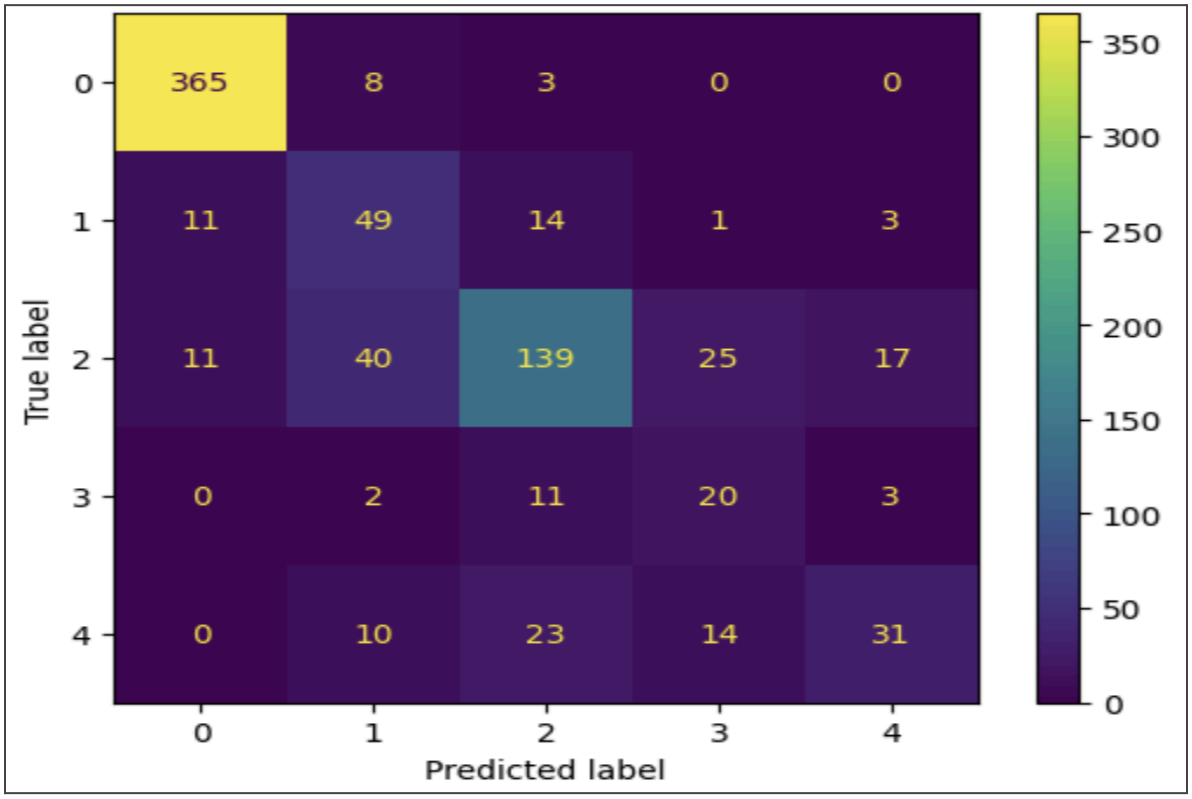
The below chart compares the Training Vs Validation Loss and Accuracy curve analysis of ResNet50-CNN model.



Train Accuracy	Validation Accuracy	Test Accuracy
74.68%	83.50%	75.50%

The mode's training history shows a solid learning and generalization performance, supported by a careful setup designed to address class imbalance and overfitting concerns. The training process demonstrates a consistent rise in accuracy and drop in loss across both training and validation phases, with the best epoch selected around epoch 11, and a learning rate schedule and early stopping applied for stability. The architecture uses a pre-trained ResNet50 backbone, followed by flattening, a dense layer, dropout, and a final output layer tailored to your classification task. Across all evaluations, the model achieved a train accuracy of 74.68%, a validation accuracy of 83.50%, and a test accuracy of 75.50%. This indicates that the model is not only able to learn effectively from the training data but also generalizes well to unseen data, as reflected by the strong validation and test performance.

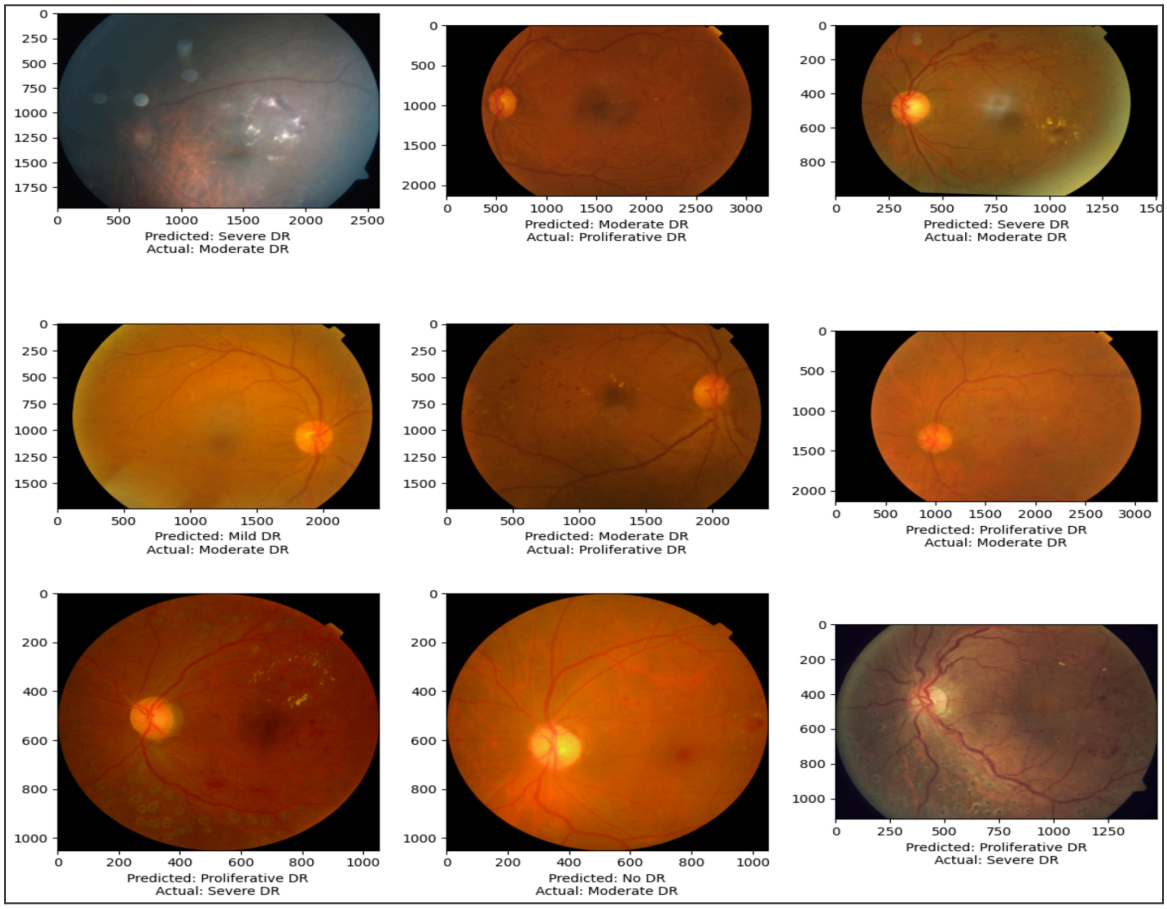
Confusion Matrix:



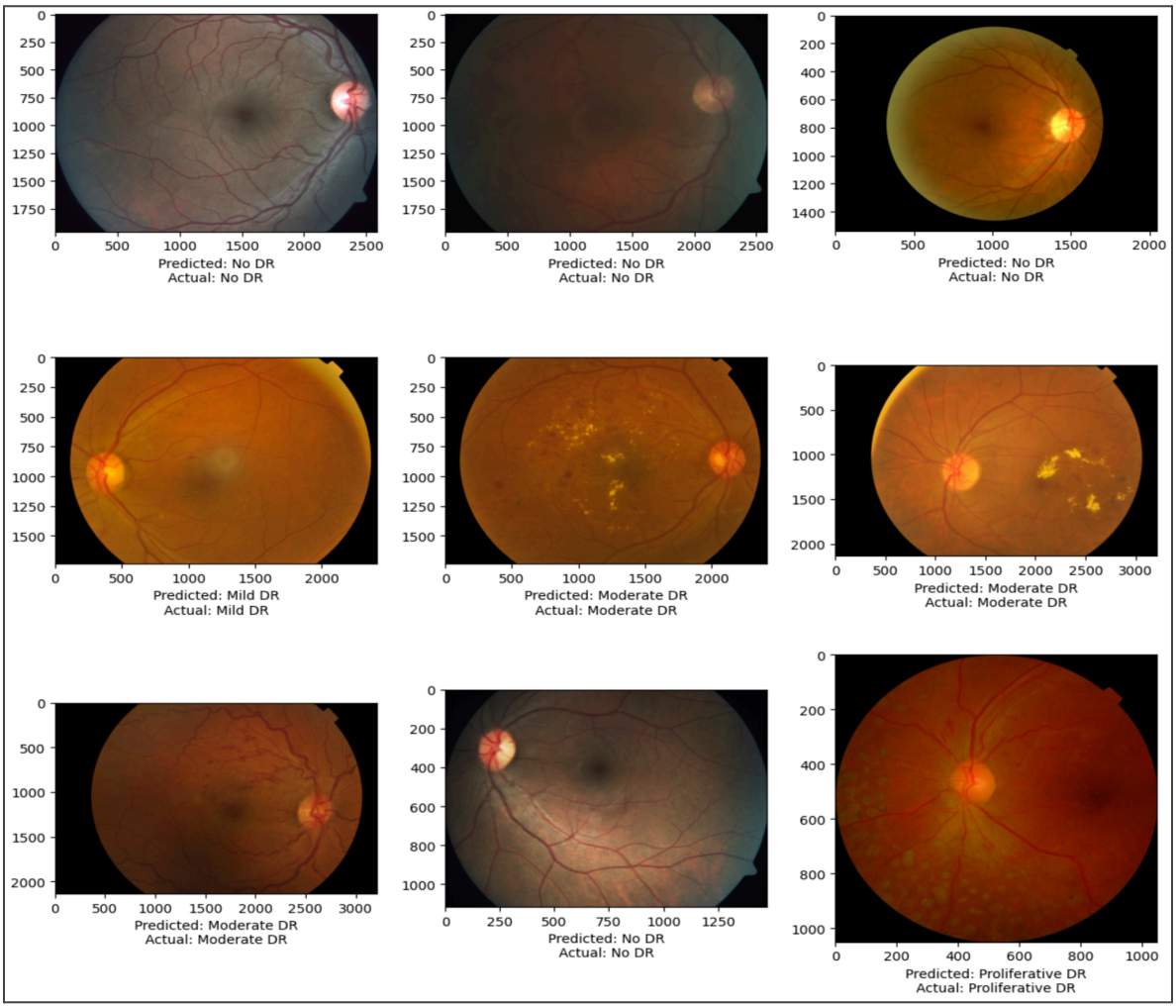
An interpretation of the ResNet50-CNN model's confusion matrix, compared with the previous custom CNN model, is provided below:

- **Class 0 (No DR)**: The true positive count is very similar to the Custom CNN model, showing just a slight increase by one, with a reduction of one in misclassifications.
- **Class 1 (Mild DR)**: The ResNet50-CNN model demonstrates a modest improvement by correctly identifying eight more Mild DR cases compared to the previous model.
- **Class 2 (Moderate DR)**: Classification improved considerably, with over a 50% increase in correct predictions. Additionally, there was about a 35% decrease in the misclassification of Moderate DR cases as healthy or Mild DR, indicating improved differentiation for this class.
- **Class 3 (Severe DR)**: The model correctly identified Severe DR patients at nearly three times the rate of the previous model, reflecting a substantial advancement. Furthermore, no Severe DR cases were misclassified as healthy.
- **Class 4 (Proliferative DR)**: A slight improvement was observed, with three more cases classified correctly compared to the earlier model, and critically, no Proliferative DR patients were misclassified as healthy.

Visuals of Misclassification by ResNet50-CNN:



Visuals of Correct classification by ResNet50-CNN:



Conclusion:

In conclusion, implementing transfer learning with the ResNet50 architecture led to a meaningful boost in diabetic retinopathy classification performance compared to the custom CNN. ResNet50's pre-trained feature extraction and skip connections enabled the model to achieve higher accuracy, more reliable generalization, and stronger separation among the five DR classes—especially for moderate, severe, and proliferative cases. The model achieved 75.50% test accuracy, outperforming the custom CNN in correctly identifying Mild, Moderate, Severe, and Proliferative DR, and significantly reducing misclassifications of severe and proliferative cases as healthy. This demonstrates the effectiveness of transfer learning with a well-designed architecture, especially in settings with limited medical image data.

Best Model Selection

Comparative Analysis

	MLP	Custom CNN	ResNet50 CNN
Structure	Feature Extraction ✗ | No. of Hidden Layers: 3	Feature Extraction ✓ | No. of Hidden Layers: 2	Feature Extraction ✓ | No. of Hidden Layers: 1
Trainable Parameters	77,237,509	52,867,845 ▼	25,691,653 ▼
Train Accuracy	47.45%	63.64% ▲	74.68% ▲
Val Accuracy	62.50%	71.00% ▲	83.50% ▲
Test Accuracy	55.75%	66.37% ▲	75.50% ▲
Train Loss	1.464	1.219 ▼	0.785 ▼
Val Loss	0.975	0.777 ▼	0.508 ▼
Test Loss	1.047	0.901 ▼	0.625 ▼
Critical Recall	Severe DR: 30.56% | Proliferative DR: 17.95%	Severe DR: 19.44% | Proliferative DR: 35.90%	Severe DR: 55.56% | Proliferative DR: 39.74%

Best Model

Based on the comparative analysis of MLP, Custom CNN, and ResNet50 CNN architectures for diabetic retinopathy classification, the optimal choice for final model selection is ResNet50 CNN. Below is the reasoned justification for this selection, addressing key evaluation metrics and aligning with clinical priorities.

1. **Feature Extraction and Structural Suitability**

   Convolutional models are inherently better suited for image classification tasks due to their ability to automatically extract hierarchical features. Both Custom CNN and ResNet50 CNN leverage feature extraction, while MLP does not, relying on fully connected layers and missing out on spatial information present in retinal images. As a result, MLP delivers significantly poorer results.

2. **Model Complexity and Trainable Parameters**

   ResNet50 CNN is the most parameter-efficient of the three, with only 25 million trainable weights versus 52 million in Custom CNN and 77 million in MLP. A lower parameter count in ResNet50 makes it less prone to overfitting and more computationally efficient for deployment, without sacrificing performance.

3. **Accuracy Across Data Splits**

   ResNet50 consistently outperforms competing models across train, validation, and test sets. This demonstrates ResNet50's superior ability to generalize, detecting disease reliably on unseen clinical data.

4. **Loss Metrics**

   The ResNet50 model also achieves the lowest loss values across all splits, signaling better calibration and less prediction error. Lower loss correlates with improved model confidence—a critical property for clinical screening systems.

5. **Critical Recall for Risky Cases**

   Of utmost priority for clinical use is the recall on high-risk classes—the detection rate for Severe and Proliferative DR. ResNet50 achieves 55.56% recall for Severe DR and 39.74% for Proliferative DR. This is significantly higher than both Custom CNN and MLP, meaning the model catches more dangerous cases (minimizing false negatives), which is essential for safe clinical deployment.

The ResNet50 model is the best model because of its convolutional feature extraction, robust accuracy, efficiency, low loss values, and outstanding recall for severe DR stages make it the safest choice for real-world clinical screening.