An Ensemble Learning Approach for DR Grading

Md Faisal Rahman1, Palak Thakur1

1Department of Computer Engineering, National Institute of Technology, Jamshedpur ,831014, India

|  |  |  |
| --- | --- | --- |
| KEYWORDS |  | ABSTRACT |
| Ensemble Learning  Neural Network  CNN  DR  MobileNet  VGG16  VGG19  IDRiD |  | Diabetic Retinopathy (DR) remains a significant cause of vision impairment and blindness worldwide, especially in individuals with diabetes mellitus. In response to the pressing need for efficient DR screening and diagnosis, this study focuses on developing a robust deep learning model for automated DR detection and classification using retinal images from the IDRiD dataset. By employing advanced computer vision techniques and leveraging ensemble learning, the aim is to enhance the accuracy and reliability of DR diagnosis, enabling early intervention to prevent vision-related complications.  The project methodology involves training and optimizing transfer learning models—MobileNet, VGG16, and VGG19—using the IDRiD dataset exclusively. Ensemble learning is utilized to combine predictions from these models through maximum voting, enhancing the overall classification accuracy. This approach streamlines the process by leveraging different architectures and their learned representations, contributing to a more robust and effective diagnostic tool.  Key findings from the project include the successful development and optimization of the ensemble learning model, demonstrating high accuracy in DR detection and classification specifically with the IDRiD dataset. This research contributes significantly to the advancement of automated DR diagnosis, offering a cost-effective and efficient solution for early detection and management of this sight-threatening condition. |

# Introduction

People with diabetes can have an eye disease called Diabetic Retinopathy (DR). This is when high blood sugar levels cause damage to blood vessels in the [retina](https://www.aao.org/eye-health/anatomy/retina-list). These blood vessels can swell and leak. Or they can close, stopping blood from passing through. Sometimes abnormal, new blood vessels grow on the retina. All these changes can steal your vision.[1] This ocular condition primarily affects the retina's blood vessels, which are responsible for transporting oxygenated blood and essential nutrients to different parts of the retina. DR manifests because of diabetes, leading to various pathological changes such as swelling and leakage of retinal blood vessels, as well as abnormal growth of new blood vessels, resulting in vascular blockage and blood leakage into healthy retinal areas.[2]

The severity of DR can vary widely, ranging from the early stages of Non-Proliferative DR (NPDR) to the more advanced stages known as Proliferative DR (PDR). NPDR encompasses mild, moderate, and severe stages, while PDR signifies advanced disease progression.[3] Clinically, DR is often classified based on the presence of characteristic lesions observed in retinal images, including microaneurysms, haemorrhages, and exudates. This grading system enables clinicians to categorize DR into five distinct grades, ranging from no DR (Grade 0) to severe PDR (Grade 4), facilitating diagnosis and treatment planning.[4]

The urgency of DR (DR) diagnosis in diabetes patients motivates the development of efficient screening tools. Current manual assessment methods by doctors are time-intensive and subject to interpretation variance.[5]

To expedite and standardize this process, we employed ensemble learning techniques to automate grading of eye images for DR severity using the IDRiD dataset. This ensemble approach incorporates multiple deep learning models, including MobileNetV2, VGG16, and VGG19, to collectively enhance classification accuracy.

Our primary objectives are twofold: first, to develop a robust computer program capable of accurately grading eye images for DR across five severity classes; and second, to optimize the learning process by leveraging ensemble learning, reducing the dependency on extensive training data. This innovative strategy aims to equip healthcare providers with a reliable, rapid diagnostic tool for early DR detection and intervention in diabetes patients.

## Lesion and their description

In DR, various lesions or abnormalities can occur in the retina because of diabetes on the blood vessels. Here are some common lesions associated with DR along with brief descriptions:

***Microaneurysms***

***Description***: Small, round dilations of retinal blood vessels.

***Significance***: Early sign of DR; may lead to leakage of fluid into the retina. [6]

***Haemorrhages***

**Description**: Leakage of blood from damaged blood vessels.

**Significance**: Haemorrhages can cause dark spots in the vision and indicate more advanced stages of retinopathy. [6]

***Exudates***

**Description**: Yellowish deposits composed of fats, proteins, and other substances.

**Significance**: Leakage from damaged blood vessels can lead to the accumulation of exudates, affecting vision.[6]

***Cotton Wool Spots***

**Description**: White, fluffy spots on the retina.

**Significance**: Indicate areas of reduced blood supply to the retina, often associated with nerve fibre layer infarcts.[6]

**Table 1 Lesion and their description**

|  |  |
| --- | --- |
| **Lesion** | **Description** |
| Microaneurysms | small pouches in the walls of tiny blood vessels in the eye |
| Haemorrhages | tiny areas where blood has leaked into the retinal tissue |
| Hard exudates | They typically have a yellow or creamy appearance. The colour is due to the lipid content within the exudates. |
| Cotton wool | They appear as white or greyish patches on the retina, resembling cotton balls or wool, hence the name. |

## Different grades of DR

In the world of DR, we categorize the condition into different stages based on what we see in the eyes. If there is no DR (DR), it means everything looks normal with no irregularities.

Moving on to mild DR, we observe the presence of tiny things called Microaneurysms. [7]

When it is moderate DR, it is kind of in between Microaneurysms and severe DR, where things get more serious. [7]

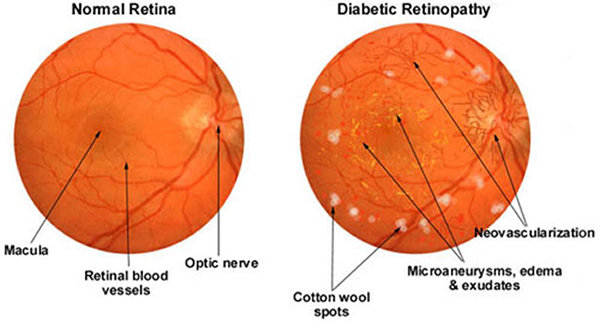
Now, severe DR can show up in different ways. There might be a lot of bleeding inside the eye in all four parts, or more than two parts might have twisted blood vessels. Sometimes, there could be issues with tiny blood vessels in more than one part of the eye. [7][6]

The last stage is called Proliferative DR, and it is even more serious. Here, new blood vessels might start growing where they should not, and there could be bleeding in the front or back part of the eye. [7]. These patients require immediate referral to a retina specialist for further testing and treatment. Peripheral neovascularization is usually treated with laser panretinal photocoagulation. [8]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| a | b | c | d | e |

Figure 1 Retina images with grading levels (a) normal, (b) mild, (c) moderate, (d) severe and (e) proliferative DR

So, in simpler terms, it goes from nothing wrong, to a few tiny irregularities, to more serious signs like bleeding and twisted blood vessels, and finally, to a stage where new blood vessels cause bigger problems. These stages help doctors understand and treat DR in the best way possible.



**Figure 2 Normal vs Effected eye**

Table 2 Intensity grading and Observations

|  |  |
| --- | --- |
| **DR intensity grading** | **Observations** |
| No DR | Absence of any irregularities |
| Mild DR | Only presence of Microaneurysms |
| Moderate DR | Symptoms lying between Microaneurysms and severe DR |
| Severe DR | It could be any of them:  More than 20 intraretinal haemorrhages in each of the 4 quadrants  More than 2 quadrants of venous beading Intraretinal microvascular problem in more than 1 quadrant |
| Proliferative DR | exhibit either neovascularization of the disc or vitreous haemorrhage. |

# Related Works

Early research on DR (DR) initially relied on manual feature extraction techniques to identify and quantify key indicators in retinal images, including veins, the optic disc, and various irregularities such as haemorrhages, microaneurysms, and exudates. Grading of these features was carried out using handcrafted features combined with machine learning algorithms like support vector machines (SVM) and k-nearest neighbour (KNN).

Lachure et al. [9] proposed a pioneering method for detecting microaneurysms and exudates, which are considered early signs of DR, using KNN and SVM classifiers. Their approach involved preprocessing of images followed by feature extraction using the Gray-Level Co-occurrence Matrix (GLCM) for further classification, with SVM demonstrating superior performance compared to KNN.

Patton et al. [10] contributed by exploring and establishing standards for retinal image examination, focusing on the detection of specific landmarks and signs associated with DR (DR). Their work emphasized the importance of accurate image analysis techniques for identifying retinal abnormalities related to DR.

Wejdan et al. [11] proposed two different methods: CNN512 (first scenario) and an adopted YOLOv3 model (second scenario) to classify fundus images into five Grades, no-DR, mild, moderate, severe, and proliferative DR. They used the DDR and the APTOS Kaggle 2019 public datasets. The first CNN model (CNN512) consisted of one zero padding layer with a value of 2, six convolutional layers each followed by max pool layers, eight batch normalization layers, two fully connected layers, and one SoftMax layer for classification. The input image size was 512 × 512 × 3. The second model was utilized to detect and localize the DR lesions, achieving a 0.216 mean average precision (MAP) in lesion localization for the DDR dataset. In classification, CNN512 achieved 88.6% and 84.1% accuracy for the DDR and the APTOS datasets, respectively. YOLOv3 acquired a classification accuracy of 89%. The accuracy could be enhanced by exploring more image processing techniques and model fine-tuning.

Acharya et al. [12] suggested a method for grading diabetic macular Edema (DME) into stages using feature extraction techniques without the need for segmentation. This approach streamlined the grading process and offered a structured method for staging DME based on extracted features. Deep Learning Techniques

In deep learning method, network directly learns low level depiction and high-level parameters from data directly and lowers the need of human intervention for feature engineering. Hacisoftaoglu et al. [13] presented a method based on GoogleNet, AlexNet, ResNet50 and CNN based framework to increase the result of DR recognition in cell-phone based and conventional fundus camera retina pictures. Retraining of these frameworks are performed on various datasets such as Messidor, EyePACS, IDRiD to examine the outcome of deploying images from different group of single, multiple, cross datasets. These techniques achieve high accuracy on independent test datasets. Achieving high accuracy with these models often requires extensive datasets, reflecting the need for comprehensive learning from diverse retinal images. Implementing these techniques efficiently demands significant computational resources, including high-end devices capable of handling the computational demands of training deep learning models on large datasets.

Alyoubi et al. [14] presented the new cutting-edge strategies for DR color fundus pictures location and deep learning procedures for classification. Moreover, the DR datasets for the color fundus retina surveyed. Contrast testing issues that need more examination are additionally taken by them. Li et al. [15] presented a new method for grading DR and DME jointly using cross disease attention network. It investigates the inward connections between the diseases with just picture level oversight. They designed two attention modules separately to learn disease dependent and disease specific features and then integrate these for grading DR and DME to increase the outcome of grading. They use Messidor and IDRiD dataset for evaluation. Reguant et al. [16] proposed a method where they initially perform visualization using CNN to find the innate image features engaged with the CNN's accountability process. Then, they fundamentally break down those provisions regarding generally known pathologies in particular haemorrhages, microaneurysms and exudates, and other visual segments. And additionally examine different CNNs by thinking about what picture highlights they get at learning time for predicting their clinical importance. The analyses are done on openly accessible fundus EyePACS and DIARETDB1 datasets. While different CNNs bring out predictable characterization results, the pace of getting image features conflict among models can be pretty much higher as 70%.

Here, we opted to classify the most recent papers based on whether they used deep learning in the ambience of DR. Researchers can improve the performance of the DR referral system by using artificial intelligence to create more robust diagnosis system that can offer quantitative information for many diseases that meets international clinical relevance requirements. Some of the shortcoming of traditional technique is overcome by machine learning specifically deep learning techniques because of its automatic feature extraction technique. Discussed recent literature concludes that deep learning can be efficiently used for retinal lesion detection and grading

# Methodology

Our study employed a comprehensive approach consisting of preprocessing, data augmentation, and ensemble learning to optimize the classification of retinal fundus images for DR (DR) using the IDRiD dataset.

## Preprocessing

Techniques included artifact removal using Otsu thresholding and contour detection, noise reduction using morphological opening and Non-Local Means Denoising (NLMD), contrast enhancement using Contrast Limited Adaptive Histogram Equalization (CLAHE), and image resizing for computational efficiency.

**Artifact Removal:**

Employed Otsu thresholding followed by contour detection and sorting to isolate the region of interest (ROI) and remove undesirable artifacts like black backgrounds.

The Otsu method utilizes an image histogram to determine the optimal global threshold value for image segmentation [17]. In the context of retinal fundus images, this approach is applied to differentiate between the image background and the Region of Interest (ROI). This method transforms a grayscale image into a binary image through a nonlinear operation. The algorithm takes a grayscale image as input and produces a binary image based on the pixel intensities of the original image. If a pixel's intensity exceeds the calculated threshold value, the corresponding output pixel is set to white; otherwise, it is set to black (0). The threshold value (T) is determined by the formula:

Here, represents the mean intensity of the pixels in the foreground (ROI), and is the mean intensity of all pixels in the image. This threshold value effectively separates the pixels into two classes (foreground and background) based on their intensities, facilitating the creation of a binary image for further analysis or processing.Top of Form

**Noise Removal:**

Utilized morphological opening to smooth optic discs and bright lesions, aiding in the detection of microaneurysms and exudates.   
Denoising retinal fundus images is crucial while preserving important features like lesions and exudates for classification purposes. The Non-Local Means Denoising (NLMD) method [18] is employed to effectively remove noise without compromising essential image features. The denoising of an image x = (x1; x2; x3) in channel i to the pixel j is executed according to (1) and (2) [19]

Here, B (j, r) denotes a neighbourhood of radius r surrounding pixel j. The weight w is determined by the squared Frobenius norm distance (or another induced norm distance) between color patches with centres at j and k which decay under a Gaussian kernel (j, k).

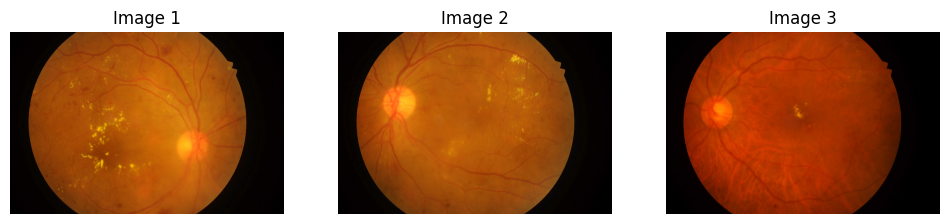
**Image Enhancement (CLAHE):**

Following denoising, CLAHE is applied to enhance contrast. CLAHE is a variant of adaptive histogram equalization designed to address issues of darkness, brightness, and uneven illumination. To leverage the YUV color space, the NLMD output images are converted, isolating the Y channel which represents illumination components [20][21]

**Image Resizing:**

Resized pre-processed images to a uniform size (512x512) to reduce computational complexity during model training.

After these preprocessing steps, the images are free from artifacts and noise, and features relevant to DR diagnosis are enhanced and made more visible. The original images often contain unnecessary background regions, which are removed during preprocessing. The enhanced images provide better visualization of blood lesions and other attributes critical for accurate classification.



A. Before Preprocessing

***B. After Preprocessing***

Figure 3 Result of Image Before(A) and After(B) applying preprocessing step

## Data Augmentation:

We used Data Augmentation to address dataset imbalance and mitigate overfitting risks. This technique aims to efficiently increase data diversity without compromising image quality, which is crucial in medical imaging tasks.

Specifically, we utilized photometric augmentation, a method that modifies RGB channels by shifting pixel values (r, g, b) to new values (r', g', b') based on predefined heuristics. This technique focuses on altering color and lighting aspects while preserving image geometry [22]. Key methods that we used within photometric augmentation include:

**Brightness Alteration:** Adjusts the overall lightness or darkness of an image by adding a factor value to each pixel's original intensity. Increasing the factor brightens the image, while decreasing it darkens the image.

Brightness(x) = Source(x) + factor

Here, "Source (x)" refers to the input pixels, and "Brightness (x)" refers to the output pixels after adjusting the brightness level using specified factor values. A factor value less than 1 results in a darker image, while a factor value greater than 1 produces a brighter image.

**Contrast Adjustment:** Enhances the difference in brightness between pixels, making light areas lighter and dark areas darker. This adjustment involves adding a factor value to each pixel's original intensity.

Contrast(x) = Source(x) + factor

Factor values of more than 1 increase the contrast, and values less than 1 decrease it.

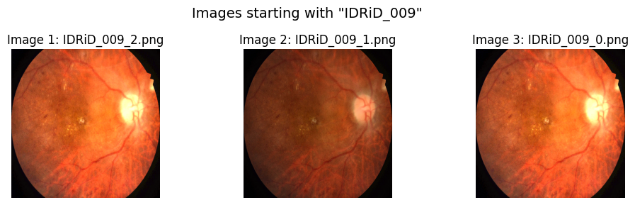
**Color Modification:** Alters the color balance of an image to make color more vivid or less saturated. This is achieved by adding a factor value to each pixel's original color intensity.

Color(x) = Source(x) + factor

**Sharpness Tuning:** Enhances image clarity and detail by adjusting sharpness, which improves edge definition between tonal zones. This is done by adding a factor value to each pixel's original intensity.

Sharpness(x) = Source(x) + factor

Data augmentation played a crucial role in expanding the diversity of the training dataset. By applying augmentation techniques, we generated additional variations of each image. This process aimed to expose the model to different perspectives of the same image, improving its ability to generalize and perform well on unseen data.

A. Before Augmentation B. After Augmentation

Figure 4 Result of Image Before(A) and After(B) applying Augmentation step

# Implementation Details:

The IDRiD dataset was partitioned into three subsets: training, validation, and test, using a stratified split strategy to ensure a representative distribution of samples across these subsets. The training set comprised 80% of the dataset, while both the validation and test sets each comprised 10% of the dataset.

Each model was trained on the training dataset for 50 epochs. The final layer of each model utilized a sigmoid activation function with 5 neurons, suitable for multi-class classification. We employed the Adam optimizer with a categorical cross-entropy loss function and accuracy as the evaluation metric.

## Ensemble Learning:

Ensemble learning is a powerful technique employed in our study to enhance the robustness and accuracy of our DR (DR) classification model.

This approach involves combining predictions from multiple individual models to produce a final prediction that is often more reliable than any single model. In our research, we utilized ensemble learning by integrating predictions from diverse deep learning architectures, including MobileNetV2, VGG16, and

VGG19. This customization involved extending the pre-trained models with tailored classification layers to suit our specific task.

Each of these models possesses unique strengths and weaknesses due to variations in network architectures, feature extraction capabilities, and learning capacities.

The ensemble learning process leverages the principle of diversity among models, where each model may excel in different aspects of DR classification. By combining predictions using a majority voting mechanism, we aim to capitalize on the collective wisdom of multiple models, resulting in a more accurate and robust classification outcome.

We separately trained each model and then combined each of them for prediction by adding weights to each model based on their accuracy achieved on test set. Below is a summary of the architecture for each model utilized in our study:

**Individual Model Architectures**

**VGG16:**

The VGG16 architecture comprises multiple convolutional layers, each followed by max-pooling layers for down sampling. It consists of 13 convolutional layers and 3 fully connected layers, culminating in an output layer for classification. Most unique thing about VGG16 is that instead of having many hyper-parameters they focused on having convolution layers of 3x3 filter with stride 1 and always used the same padding and max pool layer of 2x2 filter of stride 2.

VGG16 is object detection and classification algorithm which can classify 1000 images of 1000 different categories with 92.7% accuracy. It is one of the popular algorithms for image classification and is easy to use with transfer learning.

**MobileNetV2**

MobileNetV2 is a lightweight and efficient convolutional neural network (CNN) architecture designed for mobile and edge devices. It significantly reduces computational complexity while maintaining high performance in various computer vision tasks. MobileNetV2 achieves a balance between model size and accuracy, making it well-suited for resource-constrained environments. It features depth wise separable convolutions, which separate the spatial and channel-wise transformations, resulting in a substantial reduction in computational cost and model size.

**VGG19**

VGG19 is a deep convolutional neural network (CNN) architecture known for its depth and effectiveness in image classification tasks. It is an extension of the VGG16 model, featuring additional convolutional layers that contribute to increased model complexity and representational capacity

VGG19 consists of 19 layers (hence the name), including 16 convolutional layers and 3 fully connected layers. The deeper architecture allows for more complex feature extraction and learning. It maintains a uniform architecture with small (3x3) convolutional filters and max-pooling layers (2x2) used for spatial reduction. This consistent structure contributes to the model's effectiveness and interpretability.

**Model Customization:**

**Flatten Layer:**

The output of the last convolutional layer of the all-base models are flattened into a one-dimensional vector. Flattening is a crucial step in transitioning from the convolutional layers, which handle spatial hierarchies of features, to the fully connected layers, which interpret these features to make final predictions. The convolutional layers output a multi-dimensional tensor that represents the activation maps, reflecting the learned features from the input image.

**Dense Layer:**

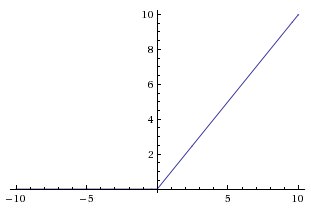
The dense layer, also known as a fully connected layer, plays a crucial role in a neural network by connecting every neuron in the layer to every neuron in the previous layer. This dense connectivity allows the model to learn complex representations of the data. In this context, we have added a dense layer with 512 neurons to our model.

The ReLU (Rectified Linear Unit) activation function is applied to each unit in the dense layer. The ReLU function is defined as:

**f(x) = max (0, x)**

This means that for any input value 𝑥, the output is 𝑥 if x is positive, and 0 if 𝑥 is negative. This simple function introduces non-linearity into the model, which is essential for learning complex patterns. Non-linearity allows the network to approximate more complicated functions than a linear model.

Graphically it looks like this,



**Figure 5 RELU function Graph**

**Dropout Layer:**

To mitigate overfitting, a Dropout layer with a rate of 0.5 is added. This layer randomly sets 50% of the input units to zero at each update during training, helping to prevent co-adaptation of neurons.

**Output Layer:**

Finally, a Dense layer with 5 units and SoftMax activation function is added. This layer provides the output probabilities for the five classes.

Mathematically, SoftMax is defined as,

Here,

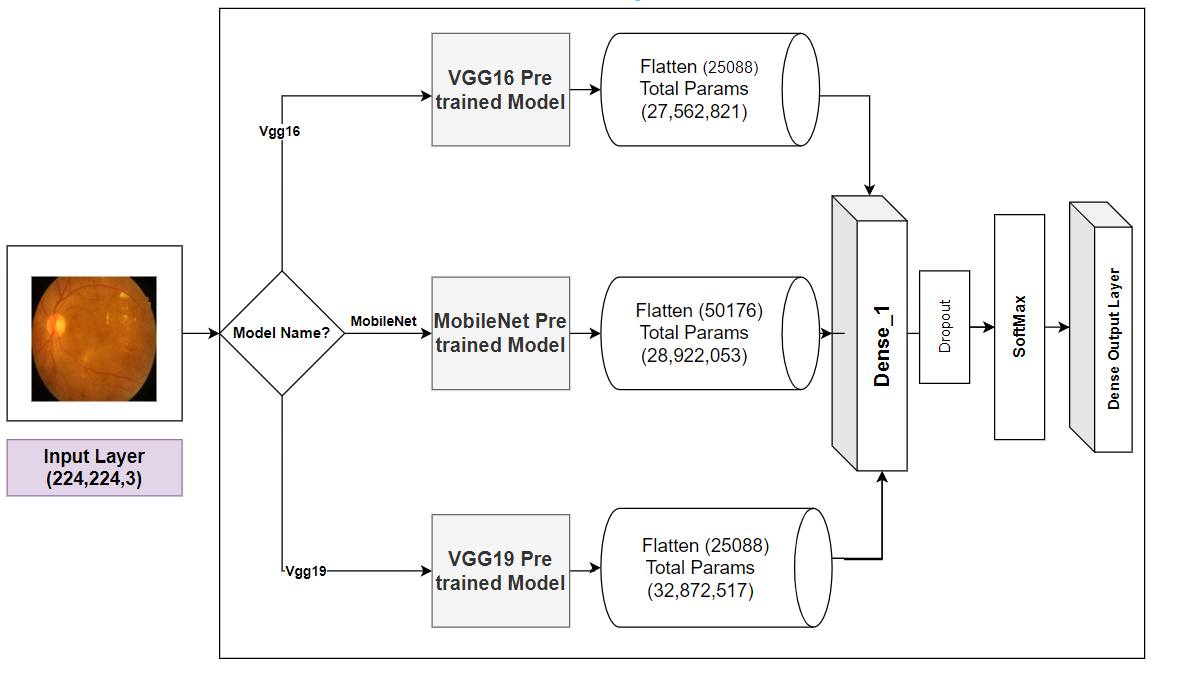
y = input vector to a SoftMax function. It consists of n elements for n classes

yi = i-th element of input vector,

= standard exponential function applied on y,

n = number of classes

SoftMax is an activation function that scales numbers/logits into probabilities. The output of a SoftMax is a vector (say v) with probabilities of each possible outcome. The probabilities in vector v sums to one for all possible outcomes or classes.



**Figure 6 Framework of Our proposed Model**

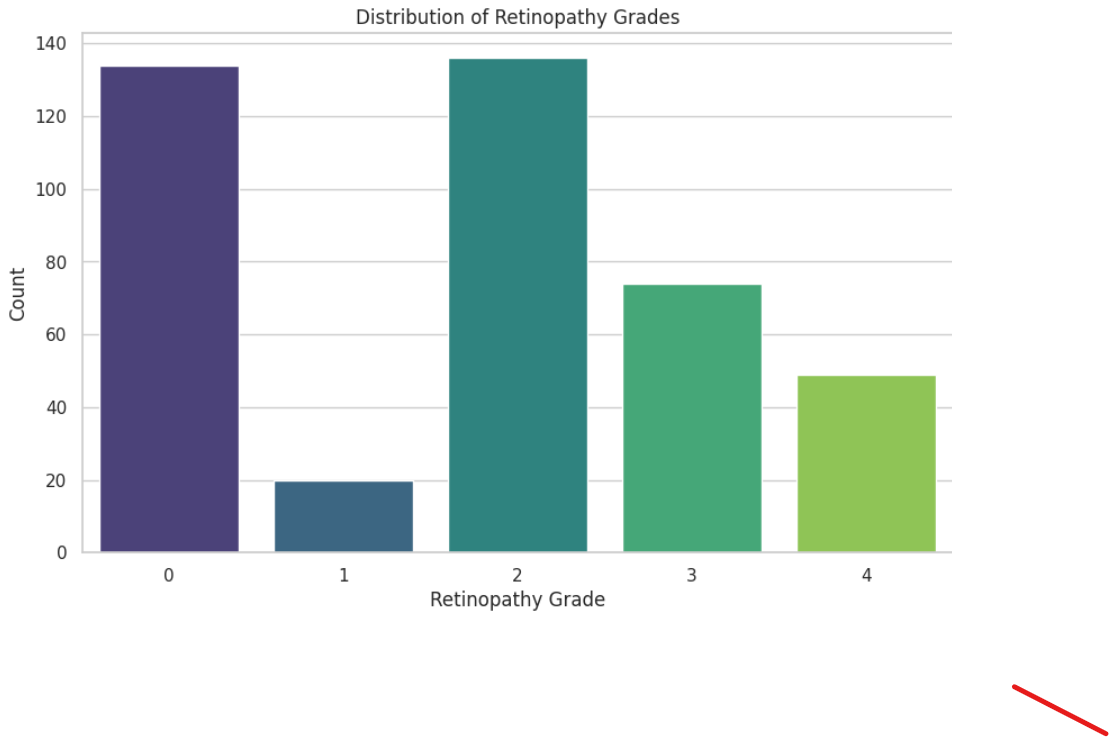
# Experiment, Result and Comparisons

## Dataset Description

The IDRID (Indian DR image dataset) comes from real eye clinic tests in Nanded, Maharashtra, India. it is the first dataset showing what eye conditions are like in India [23]. Indian DR Image Dataset (IDRiD) dataset consists of typical DR lesions and normal retinal structures annotated at a pixel level. This dataset also provides information on the disease severity of DR and diabetic macular Edema for each image. This dataset is perfect for the development and evaluation of image analysis algorithms for early detection of DR.

This dataset is free and has 516 pictures. they are divided into two groups: training set and testing set. training set has total of 413 images and testing set has total of 103 images. the severity of DR is classified into five groups based on a scale called the international clinical DR.

The data distribution in different grades are not balanced. Given below is graph of data distribution of images in IDRiD dataset



**Figure 7 Distribution of dataset**

To make it balanced we have applied augmentation techniques on the dataset. We selected an augmentation factor for each class such that each class consists almost 250 sample. Total sample after augmentation increased from 413 to 1267. Augmentation techniques that were applied has been discussion earlier in augmentation section. Below is the graph depicting image distribution in different classes after augmentation

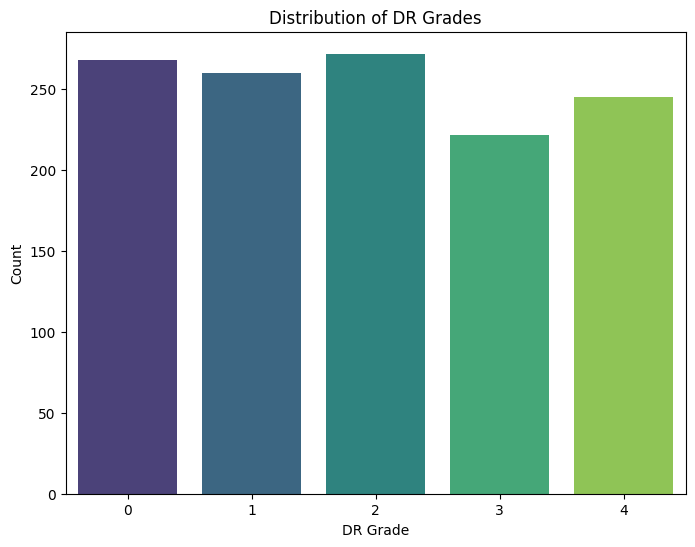


Figure 8 Distribution of dataset after augmentation

## Experimental Setup:

For this study, an Intel Core i9-13900K 24-core desktop CPU was utilized. This CPU features a base frequency of 3.0 GHz for the Performance-cores (P-cores) and 2.2 GHz for the Efficient-cores (E-cores), with a maximum clock speed of 5.8 GHz. It supports up to 32 threads and operates at a power consumption of 125W. The processor architecture includes 8 Performance-cores and 16 Efficient-cores, providing a balance between high-performance computing and energy efficiency.

Additionally, the Intel Core i9-13900K incorporates Intel UHD Graphics 770 and is compatible with the LGA 1700 socket. It boasts 36MB of cache memory, enhancing data access speed and efficiency.

The experimental desktop system was configured with 128GB of RAM, although any system with at least 12GB of RAM can be used for conducting the experiments outlined in this study. The operating system used was Windows 11 Version 21H1 or later, optimized to leverage the capabilities of the 13th generation Intel Core i9 processor.

For software development and experimentation, Visual Studio Code version 1.89.1 was employed as the integrated development environment (IDE), providing a streamlined environment for coding, testing, and analysing experimental results.

## Performance metrics

The overall accuracy of the ensemble model, which combines predictions from multiple individual models using weighted averaging was recorded to be 95.27%. The ensemble approach leverages the strengths of each model to achieve improved performance in DR classification, as demonstrated by the final accuracy results.

Performance metrics are crucial tools used to evaluate the effectiveness and accuracy of machine learning models. Some common performance metrics used in classification tasks:

**Precision:** Precision is a fundamental metric used to evaluate the performance of a classification model, particularly in scenarios where correctly identifying positive instances is critical. Precision measures the accuracy of positive predictions made by the model and answers the question: "Of all instances predicted as positive, how many are actually positive?"

Precision =

**Recall (Sensitivity or True Positive Rate):** Recall measures the ability of the model to correctly identify positive instances. It answers the question: "Of all actual positives, how many did we predict as positive?". Recall provides insights into the model's ability to identify all positive instances in the dataset. A high recall indicates that the model is successfully capturing most of the positive cases, with few positive instances being missed (i.e., low false negative rate).

Recall =

**F1 Score:** The F1 score is a metric used to evaluate the performance of a classification model, particularly in cases where there is a need to balance precision and recall. It is the harmonic mean of precision and recall, providing a single metric that captures both aspects of a model’s accuracy.

𝐹 1 = 2 ×

**Specificity (True Negative Rate):** Specificity, also known as the true negative rate, is a performance metric used in classification tasks, particularly in binary classification. It measures the proportion of actual negatives that are correctly identified by the model, answering the question: "Of all the actual negative instances, how many were correctly predicted as negative?"

Specificity =

**Table 3 Metrics achieved by individual class**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Class 0 | Class 1 | Class 2 | Class 3 | Class 4 |
| Precision | 88.8% | 100% | 88.8% | 100% | 100% |
| Recall | 96.0% | 100% | 88.8% | 95.5% | 96.0% |
| F1-Score | 92.3% | 100% | 88.8% | 97.6% | 97.9% |
| Specificity | 97.0% | 100% | 97.0% | 100% | 100% |

**Confusion Matrix**

A confusion matrix is a fundamental tool used to evaluate the performance of a classification model. It provides a comprehensive summary of the predictions made by the model compared to the actual ground truth labels. The confusion matrix is particularly useful for understanding the types and frequencies of prediction errors made by the model across different classes.

The confusion matrix obtained from our model is presented below:

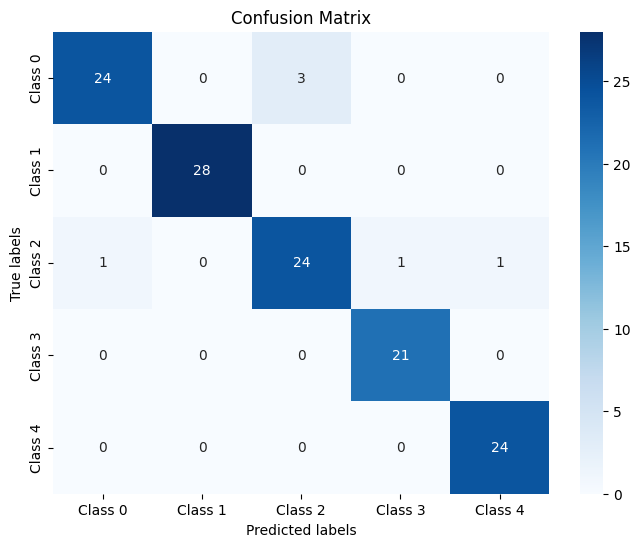


Figure 9 Confusion matrix

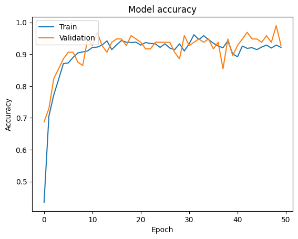
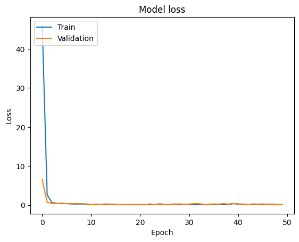
## Experimental Results

Below are the achieved accuracies and loss for the three different models:

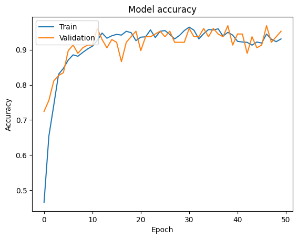
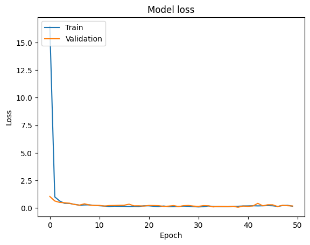
Table 4 Accuracy and Loss of different models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Training**  **Accuracy** | **Validation**  **Accuracy** | **Testing Accuracy** | **Loss** |
| VGG16 | 87.76% | 85.04% | 89.55% | 0.2231 |
| MobileNet | 92.98% | 95.28% | 94.04% | 0.2615 |
| VGG19 | 82.24% | 84.25% | 88.50% | 0.3949 |

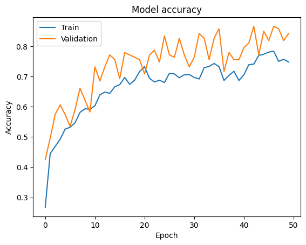
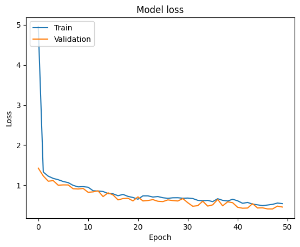
Here are the accuracy and loss graphs for each model that depict the training progress of each model, showing how accuracy improved and loss decreased as training epochs advanced. The curves provide insights into the performance and convergence behaviour of the models during the training process.

*A.Train and validation accuracy (VGG16) B.Train and validation Loss (VGG16)*

*C. Train and validation accuracy (MobileNet) D. Train and validation Loss (MobileNet)*

*E Train and validation accuracy (VGG19) F. Train and validation Loss (VGG19)*

*Fig 10 Visualization of Accuracy curve and Loss curve of VGG16 (A and B), MobileNet (C and D) and VGG19 model (E and F) after training over 50 epochs.*

To leverage the strengths of individual models and optimize the ensemble prediction, weights were assigned to each model based on its performance metrics. The assigned weights reflect the confidence in each model's predictions:

**Table 5 Weights assigned to different models**

|  |  |
| --- | --- |
| **Model** | **Weights** |
| VGG16 | 1.1 |
| MobileNet | 1.2 |
| VGG19 | 1.0 |

The rationale behind weight assignment is to prioritize models with higher accuracy, such as MobileNet, by assigning them higher weights in the ensemble. This approach aims to enhance the predictive performance of the ensemble system by aggregating predictions based on individual model strengths.

Now, to compute the Final prediction we used the given formula:

= () + () + ( )

In the above expression:

, , and ​ represent the predicted outputs (class probabilities) from the VGG16, MobileNet, and VGG19 models, respectively.

​, and denote the assigned weights for each model based on their performance.

The operation computes the weighted sum of predictions from each model to generate the .

The is determined by applying the argmax function to select the class index with the highest value in , representing the final predicted class label for the ensemble.

## COMPARISON WITH EXISTING WORKS

Table 10 provides an overview of the comparison between our proposed model and the existing related work. The proposed ensembled model was compared to some recent studies in DR classification. Table x compares these previous studies and our proposed methodology based on accuracy

As previously stated, the suggested ensemble model was trained on the merged dataset and reached a test accuracy of 95.27% We performed a multi-class classification of fundus images, which contains five classes. Table 10 shows a comparison of our method with other studies and an overview of their limitations. Several researchers performed a binary classification utilizing a single public fundus image dataset, obtaining classification accuracies ranging from 80% to 99%. Other researchers employed single datasets for multi-class classification, resulting in classification accuracies in the range of 80% to 94.59%

**Table 6 Accuracy comparison with existing literature**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Paper** | **Name of Dataset** | **Model** | **Classification Type** | **Accuracy** |
| Zhan Wu et al. [24] | IDRiD  DR’s Kaggle dataset | CF-DRNet | Multiclass  No DR, Mild, Moderate, Severe and PDR | 56.19% (IDRiD dataset)  83.10% (Kaggle Dataset) |
| [Rubina Sarki](https://pubmed.ncbi.nlm.nih.gov/?term=Sarki+R&cauthor_id=33088488) et al [25] | Messidor  Messidor-2  DRISHTI-GS  Retina dataset | Fine tuned vgg16 and Inception v3 model | Multiclass  (Normal, DR, DME, Glaucoma, Cataract)  Mild Multiclass  (Normal, Mild DR mild DME and Mild Glaucoma) | 88.3%(Multiclass)  85.95% (Mild Multiclass) |
| P. Saranya, S. Prabakaran [26] | Messidor  IDRiD | CNN | Multiclass  No DR, Mild, Moderate DR, Severe DR | 90.89%(Messidor)  90.29%(IDRID) |
| Harry Pratt et al,[27] | Kaggle Dataset | CNN | Multiclass  No DR, Mild, Moderate, Severe and PDR | 75% |
| [Zhentao Gao](https://ieeexplore.ieee.org/author/37086529226) et al [28] | Ophthalmology Department, Health Management Centre, and Endocrinology & Metabolism Department | Inception-v3  Inception@4 | Multiclass  Normal, Moderate, Heavy, and Severe | 88.72% |
| [Jagadish Nayak](https://link.springer.com/article/10.1007/s10916-007-9113-9#auth-Jagadish-Nayak-Aff1) et al [29] | Department of Ophthalmology, Kasturba Medical College, Manipal, | ANN | Multiclass  Normal, NPDR and PDR | 93% |
| HARSHIT KAUSHIK et al [30] | EyePacs | Stacked CNN | Binary  No DR and Having DR  Multiclass  No DR, Mild, Moderate, Severe and PDR | 97.92% (Binary Classification)  87.45% (Multiclass Classification) |
| **Ensembled Model (Our Proposed work)** | **IDRiD** | **Vgg16, vgg19 and MobileNet** | **Multiclass**  **No DR, Mild, Moderate, Severe and PDR** | **95.27%** |

# Conclusion

Our study developed and evaluated an ensemble deep learning model for DR (DR) classification using the IDRiD dataset. By combining VGG16, MobileNet, and VGG19 architectures, we achieved an impressive test accuracy of 95.27%. Precision, recall, Specificityand F1 score analyses demonstrated the model's effectiveness across DR severity grades.

The ensemble approach underscores the value of combining diverse models to enhance classification performance. Future work will focus on dataset expansion and real-world deployment to validate our model's utility in early DR detection and management.

This study contributes to advancing automated DR diagnosis, offering a reliable solution for improving patient care and reducing vision-related complications associated with diabetes.

References

1. [Kierstan Boyd](https://www.aao.org/biography/eac93724-3619-4834-8f38-50520c476300) , “DR: Causes, Symptoms, Treatment” American Academy of Ophthalmology, Nov 27 2023
2. Lai C.-M., Dunlop S. A., May L. A., et al. Generation of transgenic mice with mild and severe retinal neovascularisation. British Journal of Ophthalmology.  2005
3. Cecelia Koetting, OD, FAAO “The Four Stages of DR” Modern Optometry, June 2019
4. L. Z. Heng, O. Comyn1, T. Peto, C. Tadros, E. Ng, S. Sivaprasad, and P. G. Hykin “DR: pathogenesis, clinical grading, management and future developments”, Diabetic Medicine Volume 30, Issue 6, June 2013
5. K.N.D.van Eijk, J.W. Blom, J. Gussekloo, B.C.P. Polak, Y. Groeneveld “DR screening in patients with diabetes mellitus in primary care: Incentives and barriers to screening attendance” ,[Diabetes Research and Clinical Practice](https://www.sciencedirect.com/journal/diabetes-research-and-clinical-practice) , [Volume 96, Issue 1](https://www.sciencedirect.com/journal/diabetes-research-and-clinical-practice/vol/96/issue/1), April 2012,
6. Robert N. Frank, “DR” , [Progress in Retinal and Eye Research](https://www.sciencedirect.com/journal/progress-in-retinal-and-eye-research) [Volume 14, Issue 2](https://www.sciencedirect.com/journal/progress-in-retinal-and-eye-research/vol/14/issue/2), 1995,
7. R. Taylor and D. Batey, Handbook of Retinal Screening in Diabetes: Diagnosis and Management. Hoboken, NJ, USA: Wiley, 2012.
8. Wong TY, Sun J, Kawasaki R, et al. Guidelines on diabetic eye care: The International Council of Ophthalmology recommendations for screening, follow-up, referral, and treatment based on resource settings. Ophthalmology. 2018;125(10):1608-1622.
9. Mr. Jaykurnar Lachure Mr. A. V. Deorankar Mr. Sagar Lachure Miss. Swati Gupta Mr. Romit Jadhav, “DR using Morphological Operations and Machine Learning” International Advance Computing Conference 2015
10. Niall Patton, Tariq M. Aslam, Thomas MacGillivray, Ian J. Deary, Baljean Dhillon, RobertH. Eikenboom, Kanagasingam Yogesan, Ian J. Constable “Retinal image analysis: Concepts, applications and potential” [Progress in Retinal and Eye Research](https://www.sciencedirect.com/journal/progress-in-retinal-and-eye-research) [Volume 25, Issue 1](https://www.sciencedirect.com/journal/progress-in-retinal-and-eye-research/vol/25/issue/1), January 2006
11. W. L. Alyoubi, M. F. Abulkhair, and W. M. Shalash, ‘‘DR fundus image classification and lesions localization system using deep learning,’’ Sensors, vol. 21, no. 11, p. 3704, May 2021
12. [Rajendra Acharya U](https://link.springer.com/article/10.1007/s10916-008-9154-8#auth-Rajendra-Acharya_U-Aff1),  [Chua Kuang Chua](https://link.springer.com/article/10.1007/s10916-008-9154-8#auth-Chua_Kuang-Chua-Aff1), [E. Y. K. Ng](https://link.springer.com/article/10.1007/s10916-008-9154-8#auth-E__Y__K_-Ng-Aff4),  [Wenwei Yu](https://link.springer.com/article/10.1007/s10916-008-9154-8#auth-Wenwei-Yu-Aff2) &  [Caroline Chee](https://link.springer.com/article/10.1007/s10916-008-9154-8#auth-Caroline-Chee-Aff3) , “Application of Higher Order Spectra for the Identification of Diabetes Retinopathy Stages”  [Journal of Medical Systems](https://link.springer.com/journal/10916) 2008
13. Recep E. Hacisoftaoglu, Mahmut Karakaya, Ahmed B. Sallam “Deep learning frameworks for DR detection with smartphone-based retinal imaging systems” [Pattern Recognition Letters](https://www.sciencedirect.com/journal/pattern-recognition-letters) [Volume 135](https://www.sciencedirect.com/journal/pattern-recognition-letters/vol/135/suppl/C), July 2020,
14. Wejdan L. Alyoubi, Wafaa M. Shalash, Maysoon F. Abulkhair “DR detection through deep learning techniques: A review” [Informatics in Medicine Unlocked](https://www.sciencedirect.com/journal/informatics-in-medicine-unlocked) [Volume 20](https://www.sciencedirect.com/journal/informatics-in-medicine-unlocked/vol/20/suppl/C), 2020
15. Tao Li , Yingqi Gao , Kai Wang , Song Guo , Hanruo Liu , Hong Kang “Diagnostic assessment of deep learning algorithms for DR screening”, [Information Sciences](https://www.sciencedirect.com/journal/information-sciences), [Volume 501](https://www.sciencedirect.com/journal/information-sciences/vol/501/suppl/C), October 2019
16. [Roc Reguant](https://www.nature.com/articles/s41598-021-89225-0#auth-Roc-Reguant-Aff1-Aff2), [Søren Brunak](https://www.nature.com/articles/s41598-021-89225-0#auth-S_ren-Brunak-Aff1) & [Sajib Saha](https://www.nature.com/articles/s41598-021-89225-0#auth-Sajib-Saha-Aff2) Understanding inherent image features in CNN-based assessment of DR , [Scientific Reports](https://www.nature.com/srep/about) , 06 May 2021
17. T. Y. Goh, S. N. Basah, H. Yazid, M. J. A. Safar, and F. S. A. Saad, ‘‘Performance analysis of image thresholding: Otsu technique,’’ Measurement, vol. 114, pp. 298–307, Jan. 2018,
18. A. Buades, B. Coll, and J.-M. Morel, ‘‘A non-local algorithm for image denoising,’’ in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2, Jun. 2005, pp. 60–65.
19. G. Sathiya and P. Gayathri, ‘‘Automated detection of DR using GLCM,’’ Int. J. Appl. Eng. Res., vol. 9, no. 22, pp. 7019–7027, 2014.
20. F. Shao, Y. Yang, Q. Jiang, G. Jiang, and Y.-S. Ho, ‘‘Automated quality assessment of fundus images via analysis of illumination, naturalness, and structure,’’ IEEE Access, vol. 6, pp. 806–817, 2018.
21. S. Rasta, M. Partovi, H. Seyedarabi, and A. Javadzadeh, ‘‘A comparative study on preprocessing techniques in DR retinal images: Illumination correction and contrast enhancement,’’ J. Med. Signals Sensors, vol. 5, no. 1, p. 40, 2015.
22. L. Taylor and G. Nitschke, ‘‘Improving deep learning with generic data augmentation,’’ in Proc. IEEE Symp. Ser. Comput. Intell. (SSCI), Nov. 2018, pp. 1542–1547.
23. Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, Fabrice Meriaudeau “Indian DR Image Dataset (IDRiD): A Database for DR Screening Research” Data, 3 (2018), pp. 1-8
24. Zhan Wu, Gonglei Shi, YangChen, Fei Shi e, Xinjian Chen e, Gouenou Coatrieux f, Jian Yang g, Limin Luo a b c d, Shuo Li h, “Coarse-to-fine classification for DR grading using convolutional neural network” , [Artificial Intelligence in Medicine](https://www.sciencedirect.com/journal/artificial-intelligence-in-medicine), vol. 108, Aug. 2020,
25. R. Sarki, K. Ahmed, H. Wang, and Y. Zhang, ‘‘Automated detection of mild and multi-class diabetic eye diseases using deep learning,’’ Health Inf. Sci. Syst., vol. 8, no. 1, p. 32, Dec. 2020.
26. P. Saranya and S. Prabakaran, ‘‘Automatic detection of non-proliferative DR in retinal fundus images using convolution neural network,’’ J. Ambient Intell. Humanized Comput., no. 0123456789, 2020,
27. Harry Pratta, , Frans Coenenb, Deborah M Broadbentc , Simon P Hardinga ,Yalin Zhenga,” Convolutional Neural Networks for DR”, Procedia Computer Science 90 ( 2016 ) 200 – 205
28. [Zhentao Gao](https://ieeexplore.ieee.org/author/37086529226) ,[Jie Li](https://ieeexplore.ieee.org/author/37086585847), [Jixiang Guo](https://ieeexplore.ieee.org/author/37086527222), [Yuanyuan Chen](https://ieeexplore.ieee.org/author/37086555464), [Zhang Yi](https://ieeexplore.ieee.org/author/37267126100), [Jie Zhong](https://ieeexplore.ieee.org/author/37086570878),” Diagnosis of DR Using Deep Neural Networks”,IEEE Access , January 11, 2019
29. [Jagadish Nayak](https://link.springer.com/article/10.1007/s10916-007-9113-9#auth-Jagadish-Nayak-Aff1), [P Subbanna Bhat](https://link.springer.com/article/10.1007/s10916-007-9113-9#auth-P_Subbanna-Bhat-Aff2), [Rajendra Acharya U](https://link.springer.com/article/10.1007/s10916-007-9113-9#auth-Rajendra-Acharya_U-Aff3), [C. M. Lim](https://link.springer.com/article/10.1007/s10916-007-9113-9#auth-C__M_-Lim-Aff3) & [Manjunath Kagathi](https://link.springer.com/article/10.1007/s10916-007-9113-9#auth-Manjunath-Kagathi-Aff4) “Automated Identification of DR Stages Using Digital Fundus Images”  [Journal of Medical System](https://link.springer.com/journal/10916) vol 32 ,13 Nov 2007
30. HARSHIT KAUSHIK, DILBAG SINGH MANJIT KAUR HAMMAM ALSHAZLY, ATEF ZAGUIA 4, AND HABIB HAMAM , DR Diagnosis From Fundus Images Using Stacked Generalization of Deep Models , IEEE Access, August 9, 2021.