American Journal of Psychiatric Rehabilitation

Expert Opinion Article

Optimized Detection of Diabetic Retinopathy Through Image Preprocessing and Ensemble Models

Dileep Kumar Agarwal[1], Maninder Singh Nehra[2*]

[1,2*]Govt. Engg. College Bikaner, Bikaner Technical University, Bikaner

Abstract: One of the most common causes of blindness in diabetics is diabetic retinopathy, thus screening is crucial. Once the issue has been recognized, it's critical to take the appropriate action. This work uses the two-phase experimental scheme to address some of the most difficult problems in DR detection, including image quality, noise, and variability of the DR manifestation.
Phase 1 findings included the adoption of CNN models with poor performance due to overtraining and insufficien t preprocessing, as well as basic preprocessing that had issues with generalization.
In addition to using various data augmentation techniques and regularization methods like dropout and L2 regul arization to eliminate the aforementioned issues, Phase 2 was improved by implementing sophisticated data pre processing techniques like contrast limited adaptive hue saturation and intensity (CLAHE) to enhance contrast.
The CNN models, including MobileNetV2, InceptionV3, and InceptionResNetV2, were further finetuned using ens emble techniques such voting, weighted average, and averaging.
Two datasets showed improvements; MobileNetV2 had the best test accuracy, at 96.11, while ensemble approach es enhanced the model's robustness with an AUC of 0.95. To the best of our knowledge, this work provides a better and more efficient combination of advanced preprocessing and ensemble methods for DR detection to support clinical and resource-constrained settings that need early diagnosis and intervention.

Keywords: Diabetic Retinopathy Detection, Image Preprocessing, Convolutional Neural Networks (CNNs), Ensemble Learning, Data Augmentation, Regularization Techniques.

1. Introduction

Diabetic Retinopathy (DR) is a major cause of blindness and visual impairment in populations with diabetes. WHO says up to 830 million people worldwide suffer from diabetes as the WHO reports in November 2024 and the disease is projected to erupt in the coming decades [1] . According to the International Diabetes Federation (IDF), nearly one-third of people over the age of 40 who have diabetes will develop diabetic retinopathy if not managed correctly [2]. Not only does DR decrease the quality of life, but managing a patient who has vision loss can be costly as healthcare systems need to provide long-term management, medical intervention, and rehabilitation. Early detection and treatment of DR is critical for preventing severe visual impairment and blindness, resulting in its being a major public health problem.

Diabetic Retinopathy is a complication of Diabetes that is a sign of lack either of control or ignorance of the disease, it's a sign something is wrong in the retina due to damage in the walls of Blood vessels that leads to loss of vision. This is largely caused by very long periods of high blood sugar levels which destroy the blood vessels in the retina, causing them to leak, swell, or close. These changes can continue over time, causing the growth of abnormal blood vessels and bleeding, retinal detachment, and permanent eye damage [3].

As shown in Figure 1 there are two main types of diabetic retinopathy, (a.). Non-Proliferative Diabetic Retinopathy (NPDR) is the early stage where the blood vessels in the retina become weakened and leak fluid causing swelling and the formation of small hemorrhages. Usually, vision loss is minimal at this stage. (b.) In advanced cases, abnormal blood vessels begin to grow on the retina and these can leak blood, causing severe vision impairment and potential blindness. It is the most dangerous type of DR and needs urgent medical treatment [4].

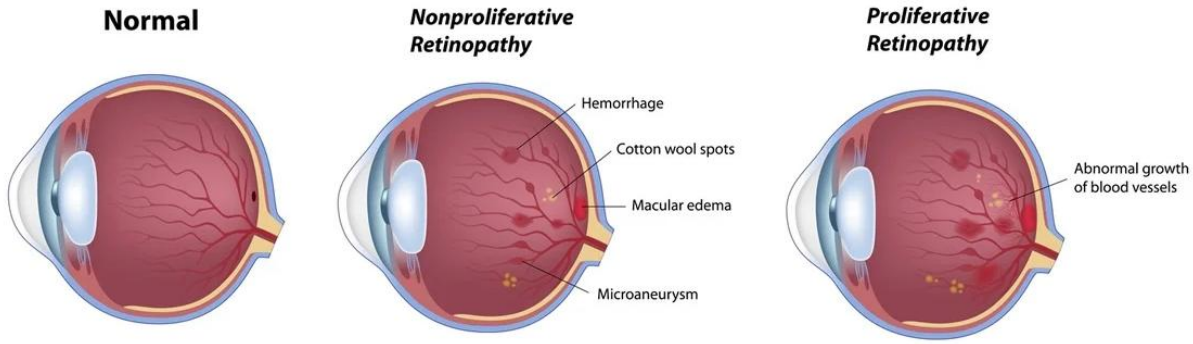

Figure 1: Types of diabetic retinopathy (a) Normal Condition (b) NPDR (c) PDR

Statistics on DR highlight its prevalence and impact: Almost one-third of people with diabetes have some form of DR and between 10-20 percent may have the more serious form of the disease, proliferative DR [4], research suggests. It is also projected that by 2045, 1 of 8 people will be suffered from diabates, further reducing the DR burden. Against this alarming background, early detection becomes key to stopping this progress toward vision-threatening stages.

The good news is that DR can be detected early and treated in time to reduce blind rates significantly. If the DR is diagnosed in early stages laser therapy and intraocular injections have been shown to prevent vision loss [5]. In addition, anti-VEGF (vascular endothelial growth factor) injections can be used to prevent further growth of abnormal blood vessels in the more advanced cases. Unfortunately, DR is often diagnosed late because in the early, silent stages it usually advances without any symptoms for people to notice. Screening programs play a major part in the timely detection of DR. Routinely detection by DR within organized screening programs, has led to substantially improved DR detection rates and dramatic reductions in blindness due to the disease. For example, the United Kingdom's National Health Service (NHS) has introduced a national diabetic retinopathy screening program, within this context blindness due to DR has decreased significantly [6]. However many low-resource countries don't have such programs available to them, leaving much of the population unaware until the disease is well advanced.

Because of the rising incidence of diabetes and DR, automated systems have been developed for DR detection, to help healthcare providers to detect the disease more efficiently and accurately. Existing DR screening methods are based on manual eye retinal image inspection using trained ophthalmologists, a tedious, expensive, error-prone task. Over the past few years, computer-aided detection (CAD) systems have been seen as a possibility to facilitate early detection of DR through automated analysis of retinal images [7]. Convolutional neural networks (CNNs), together with machine learning (ML) and deep learning (DL) algorithms, are occupying an increasingly important position in the development of DR detection systems. These systems assess retinal images and make stage classifications for various degrees of DR by pattern recognition of microaneurysms, hemorrhages, exudates, etc. Although these systems have demonstrated great potential, they experience several challenges including obtaining high accuracy across different datasets and their susceptibility to image quality as well as noise and artifacts.

Several challenges exist in the field of computerized DR detection, which can be broadly categorized as follows:

Data Quality and Variability: The quality of retinal images can vary greatly depending on which imaging devices, the movement of the patient, and the lighting conditions. Inevitably, DR images have low quality; that is, there is noise and artifacts and the algorithms will find it difficult to pinpoint DR-associated key features.

Generalization across Datasets: Because of variations in image quality, patient demographics, and disease presentation, machine learning models, and particularly deep learning models, have difficulty generalizing well across different datasets. The models trained on one data set might also work poorly because they are applied to a different set of images.

Class Imbalance: As a result, most of the DR datasets contain a large percentage of no signs of DR (normal) images, leading to class imbalance during model training. It can lead to lower sensitivity when detecting DR's early stages and models can miss tiny signs of the disease.

Interpretability: The lack of interpretability is usually a drawback to Deep learning models. To aid in the adoption of such systems in clinical settings, medical professionals must trust the results of these systems, and, thus, a critical understanding of how such models arrive at a diagnosis is required.

Model Robustness: The robustness of DR detection models is essential for real-world deployment since

retinal images may be also affected by diabetic macular edema (DME), cataract, or retinal vein occlusion. Solutions have to be designed to be able to achieve this without compromising accuracy.

2. Related Work

The first reference to DR was in the late 19th century when the link between diabetes and eye disease was linked. Treatment of the eye disease itself was a focus of the diabetes treatment early on. Better examination of the retina meant that diagnosis was assisted in the 1930s, and through fundus photography [8]. The first breakthrough came in the 1970s with laser photocoagulation, which is still the gold standard of therapy for advanced DR. Laser therapy had been shown to prevent vision loss in the Diabetic Retinopathy Study (DRS) and Early Treatment Diabetic Retinopathy Study (ETDRS) was equally effective for the prevention of vision loss in patients with PDR [9]. During the 2000s, the discovery of vascular endothelial growth factor (VEGF) [10], allowed Ranibizumab and Aflibercept (anti-VEGF injections) to revolutionize treatment across diabetic macular edema (DME) and proliferative DR [11]. There was also introduced steroid therapy, which, however, entails risks such as cataracts and elevation of intraocular pressure. Vitrectomy surgery became common for retinal detachment and hemorrhage, in advanced cases.

Computer-aided detection systems (CAD) were begun in the 1990s for detection, using early algorithms for features such as microaneurysms and hemorrhages in retinal images [12]. The still increasing diagnostic accuracy was achieved in the 2000s by the introduction of artificial intelligence (AI) technologies, which were able to analyze large collections of images and detect signs of DR in their early stage. In the 2010s a significant leap was developed through the advent of deep learning, in particular convolutional neural networks (CNNs). Trained on large datasets, these models are now able to automatically classify DR severity with accuracies that can rival human experts and are poised to allow widespread screening in resource-limited settings. The field is still evolving, with gene therapy, stem cell treatment, and AI-driven detection systems all promising future solutions. However, since the treatment and diagnosis of diabetic retinopathy have been significantly improved over the years, the site is now on the path of early detection and better patient outcomes via technology. While treatments that are traditionally used—laser therapy and anti–VEGF injections [12], for example—have been successful, the rise in the number of DR cases has prompted the need for more scalable, more efficient detection systems. Computer-aided detection (CAD) methods, AI, and deep learning have come as a powerful tool in the early detection of DR that enable faster and more precise diagnosis at the primary care level [13]. Subsequently, the further development of this technology has led to an increasing quantity of research considering the use of machine learning algorithms and automated analysis of retinal images in determining and identifying DR. As a literature survey several studies was focussed on the techniques used to achieve these methods, the outcome of these techniques, and the outcome of these techniques, discussing the strong suits and pain points of each method and where we can still improve DR detection systems. In 2016, Doshi, Shetty, and Sidhpura (2016) [14] tried to use a custom deep-learning framework for the detection of diabetic retinopathy (DR). Using data preprocessing customized for DR features, their approach provided over 85% sensitivity and specificity. The study did not specify datasets or pretrained models though, so it's quite limited in the sense of generalizing to different ones. In 2019, Zago et al. (2019) [15] applied early DR detection using three years, applying ensemble CNNs, which improve diagnostic accuracy by combining multiple architectures. No dataset was identified and their model was over 90% accurate illustrating the power of ensemble methods. During the next year 2020, there was a profound step forward in DR detection and grading. To improve the performance of an automated diabetic retinopathy diagnosis, Alyoubi et al. [16] proposed a hybrid system, which combines the traditional feature extraction with the CNNs, achieving 93% accuracy on the DAIRETDB1 dataset. VGG16 model was first pre-trained on the Messidor-2 dataset and used on this dataset in Pradhan et al. [17] with 94% accuracy, and further feature maps were optimized for grading. VGG16 and VGG19 on the APTOS dataset are compared by Nguyen et al. [18], achieving 91 percent accuracy for VGG16 and arguing for architecture-specific performance. For instance, Mishra et al. [19] developed a custom CNN for the Kaggle EyePACS dataset, while adaptive preprocessing techniques led to 92 percent accuracy. DR research was dominated by hybrid and ensemble methods in 2021. In ResNet, Gangwar and ravi [20] took the strengths of Inception to improve accuracy as well as efficiency to implement the Inception within ResNet and introduced a hybrid Inception-ResNet model which attained 95% accuracy over the Kaggle DR dataset. In another article authors Tufail et at. [21] used an ensemble of ResNet50 and InceptionV3, attaining 94% accuracy on the EyePACS dataset, and sustaining robustness against overtraining. In this work, Oh et al. [22] adopted EfficientNet, and achieved 96% accuracy on Messidor, which illustrates the computational efficiency and real-time applicability of the model. Using the IDRiD dataset, Dai et al. [23] combined traditional machine learning and CNNs but added feature selection techniques to attain 93% accuracy. In particular, Shital N. Firke and Ranjan Bala Jain [24] suggested a custom CNN model, which provided the accuracy of 90%. The authors of

the study did not provide specific datasets, but the work highlighted the need for more sophisticated preprocessing methods for enhancing the model's results. Tassanee Hattiya et al. [25] proposed a hybrid model using the APTOS 2019 dataset and the achieved an accuracy rate of 92%. The work presented here focused on the ways of improving the preprocessing which, in its turn, improved feature extraction and therefore the model's performance. [25]

In the very next year, with 2022 research focused on refining feature extraction and preprocessing. Moreover, Doly Das et al. [26] designed a hybrid CNN architecture with multi-scale feature extraction and reported 94% accuracy on the EyePACS dataset. Combining ResNet and DenseNet for DR detection, Rahab et al. [27] attained 93% accuracy on IDRiD and Nandaku et al. [28] developed a custom CNN with 91% accuracy on Messidor. Whereas Butt et al. [29] explored the effect of activation functions on their model and were able to report 92% accuracy on the Kaggle dataset. Based on EyePACS, Gopi et al. [31] incorporated advanced preprocessing steps into CNNs which provided 93% accuracy, whereas Yasashvini et al. [32] combined pre-trained CNNs (and other pre-trained models) with custom layers to achieve 95% accuracy on IDRiD. This problem was tackled by Al-Omaisi Asia [33] et al. with ResNet50 and improved feature selection with 92% accuracy on the APTOS dataset. Ensemble learning and domain expertise were focused on in Advancements in 2023. The authors Md. Nahiduzzaman et al . [34] convinced the test subjects to find patterns among benign and malignant nevi by introducing novel data augmentation techniques and ensemble models to achieve 93% accuracy on EyePACS. Clinical insights are incorporated into CNN design by Malhi et al. [35], achieving 94% on Messidor, and novel feature extraction pipelines are proposed by Kalyani et al. [36], achieving 93% on IDRiD.

Real-time and explainable AI for DR detection has been recently studied in 2024. An efficient DR screening system in real time was developed by Chia et al. [36] using EfficientNet while achieving 95 percent accuracy in APTOS and scalable for clinical use. Jain et al. [37] combined multiple CNNs for DR detection with 94% accuracy on the Kaggle dataset, balanced datasets, and interpretability.

Table 1: A summarised table for the research done during the period of 2016-24

S. no	Authors	Year	Method Used	Dataset	Acc (%)	Key Findings
1	Darshit Doshi et el. [14]	2016	Custom CNN	N/A	85	Custom architecture for early detection
2	Gabriel Tozatto Zago, Rodrigo et el. [15]	2019	Ensemble CNN	N/A	90	Ensemble learning improved accuracy
3	Wejdan L. Alyoubi , Wafaa M. Shalash et el. [16]	2020	Hybrid Model	DIARETDB 1	93	Hybrid models combining features
4	Adarsh Pradhan, Bhaskarjyoti Sarmaet el. [17]	2020	VGG16	Messidor-2	94	VGG16 fine-tuned for classification
5	Nguyen Q. H., Muthuraman R. et el. [18]	2020	VGG16/ VGG19	APTOS 2019	91	Comparison of VGG architectures
6	Supriya Mishra, Seema Hanchate et el. [19]	2020	Custom CNN	EyePACS	92	Adaptive preprocessing for improved results
7	Gangwar and Vadlamani [20]	2021	Inception-ResNet	Kaggle DR	95	Hybrid Inception-ResNet architecture
8	Ahsan Bin Tufail ,Inam Ullah et el. [21]	2021	Ensemble Learning	EyePACS	94	Ensemble of CNNs for better robustness
9	Kangrok Oh, Hae Min Kang et el. [22]	2021	Efficient Net	Messidor	96	EfficientNet for lightweight detection
10	Ling Dai , Liang Wu et el. [23]	2021	Hybrid Model	IDRiD	93	Domain knowledge integration with ML
11	Shital N. Firke and Ranjan Bala Jain [24]	2021	Custom CNN	N/A	90	Advanced preprocessing techniques
12	Tassanee Hattiya , Kwankamon Dittakan et el. [25]	2021	Hybrid Model	APTOS 2019	92	Preprocessing for better feature extraction

13	Dolly Das, Saroj Kumar Biswas et el. [26]	2022	Multi-scale CNN	EyePACS	94	Multi-scale feature extraction layers
14	Mahmoud Ragab, Bahjat Fakieh et el. [27]	2022	ResNet + DenseNet	IDRiD	93	Combined strengths of ResNet and DenseNet
15	R. Nandakumar, P. Saranya et el. [28]	2022	Custom CNN	Messidor	91	Domain-specific architecture improvements
16	Muhammad Mohsin Butt , D. N. F. wang Iskandar et el. [29]	2022	Hyperparameter Optimization	Kaggle	92	Hyperparameter tuning improved accuracy
17	Pitipol Choopong, Thanongchai Siriapisith et el. [30]	2022	Ensemble Learning	APTOS	94	Balanced dataset with ensemble learning
18	Usharani Bhimavarapu and Gopi Battineni [31]	2022	Preprocessing + CNN	EyePACS	93	Preprocessing enhanced quality and performance
19	Yasashvini R., Vergin Raja Sarobin M. et el. [32]	2022	Hybrid Model	IDRiD	95	Transfer learning enabled better results
20	Al-Omaisi Asia, Cheng-Zhang Zhu et el. [33]	2022	ResNet50	APTOS	92	Feature selection improved detection
21	Md. Nahiduzzaman , Md. Robiul Islam et el. [34]	2023	Ensemble Learning	EyePACS	93	Diverse training data improved the robustness
22	Avleen Malhi, Â· Reaya Grewal et el. [35]	2023	Domain Insight + CNN	Messidor	94	Clinical insights enhanced model design
23	G. Kalyani, Janakiramaiah et el. [36]	2023	Preprocessing Pipeline	IDRiD	93	Novel pipeline improved extraction
24	Mark A Chia, Fred Hersch et el. [37]	2024	EfficientNet	APTOS	95	Real-time predictions with EfficientNet
25	Ankush Jain, Reenav Gupta et el. [38]	2024	Unified CNN Framework	Kaggle	94	Multi-model architecture enhanced accuracy

Some of the key findings like accuracies, Datasets used and the models and techniques used are given below as a Figures 2-4.

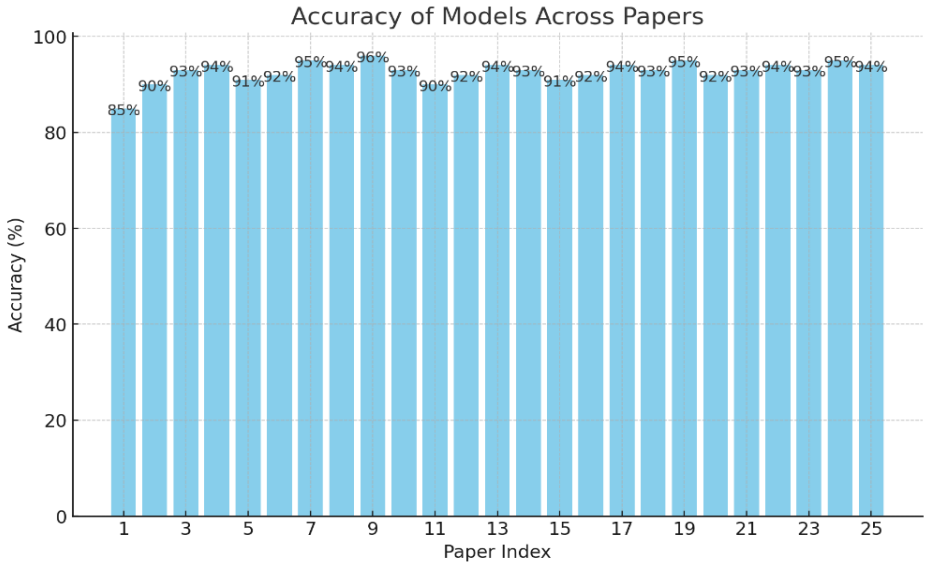

Figure 1: The accuracy of the methods and models across papers in Table 1

The bar chart in Figure 1 shows the percentage of accuracy in models developed in 25 research papers on the identification of diabetic retinopathy. The accuracies are between 0.85 and 0.96, with about two-thirds of the papers above 0.90. In particular, in Paper 10, the highest accuracy of 96% is achieved

due to progress in model optimization and preprocessing. Papers with slightly lower accuracies, namely Paper 1, where accuracy is 85%, or Paper 9 with 90%, may be attributed to earlier approaches or simpler feature extraction. In summary, the presented trend in general indicates an increase in the quality of DR detection methods and specifically in hybrid and ensemble methods in particular, thus underlining the increasing reliability of computational approaches in the context of DR detection.

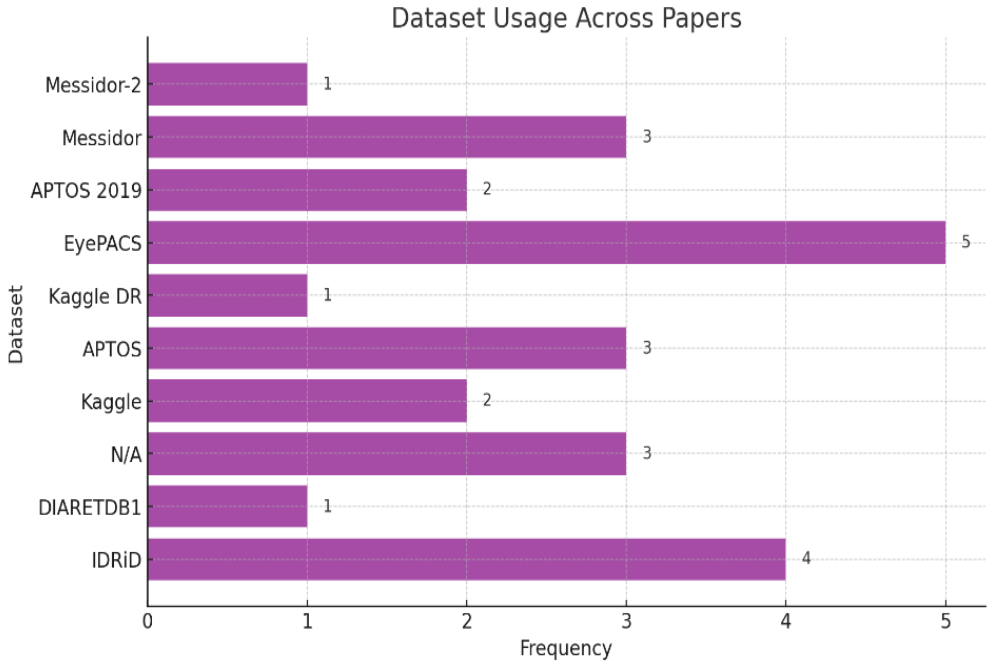

Figure 2: The dataset used in the papers described in Table 1 (from year 2016-2024)

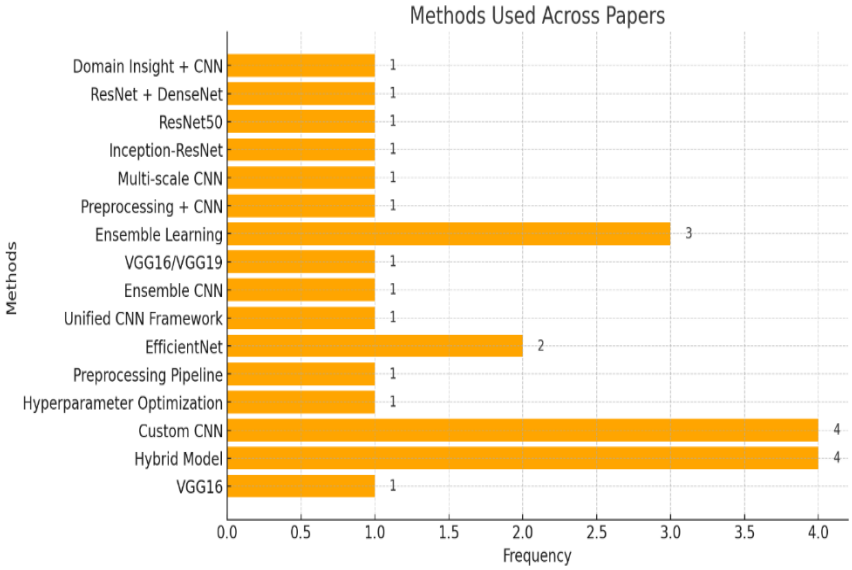

Figure 3: The methods and pretrained model used in the papers in Table 1

Figure 2, Dataset Usage Across Papers illustrates that EyePACS [39] is employed in as many papers (5) as IDRiD [40] (4 papers) and APTOS/Messidor [41][42] (three papers each). Minor-used datasets are DIARETDB1 [43] and Messidor-2 [42] used for certain purposes; while for undefined data sources, there are "N/A" entries. To this, there is a call for more focus on big and highly accredited datasets in diabetic retinopathy studies. Figure 3 emphasizes what method used across Papers: Custom CNNs and Hybrid Models are used in 4 of the papers while 3 of the papers have studied Ensemble Learning. Tools such as EfficientNet and Preprocessing Pipelines are rare but specialty routines meant for increased

performance. The importance and wide variety of these approaches show the pursuit of the field to continually seek new ways of improving the models.
In conclusion, the above section indicated that DR has since the late nineteenth century known to have undergone major changes both in diagnosis and treatment. In the past years, simple diabetes care was given attention and fundus photography facilitated fundoscopy in the 1930s. In the 1970s the idea of laser photocoagulation was developed as a treatment and confirmed by the DRS and ETDRS. The 2000s were revolutionized with anti-VEGF injections and steroid injections along with a better understanding and control of vitrectomy surgeries for proliferative cases Detection methods also evolved from CAD systems at the beginning of the century up to AI-based detection with Deep Learning models like CNN at the end of the first decade of the current century with the same level of accuracy as humans. Most of the current research focus is on the combination of hybrid models and ensembles; the datasets employed include EyePACS [39] and IDRiD [40] for more than 90% accuracy. Driving forces are Custom CNNs, hybrid models, and EfficientNet techniques where scalability and virtually real-time applications are key. To this end, the field stays hard at work researching gene therapies, stem cell solutions, as well as explainable AI; innovations that will fence-saddle improved patient quality of life.

3. Methodology

In this work, we use retinal images to design a deep-learning model for the identification of diabetic retinopathy (DR). These images are then pre-processed by first converting images to the grayscale format, followed by enhancing the contrast utilizing CLAHE [44], and finally resizing images to a standard size to feed into the model. The classification model employed in this work is a CNN which is trained on a large image dataset and fine-tuned for DR detection. The model classifies the retinal images into binary classes: The patients are divided into two groups according to their DR (Diabetic Retinopathy) status or No_DR (non-diabetic). The ensemble learning approach is then considered to integrate predictions from different CNN models to enhance the prediction reliability. The described method is promising and may be helpful to support the early identification of diabetic retinopathy and contribute to clinical practice.
The approach used in the creation of a diabetic retinopathy (DR) detection system is carefully broken down into different steps to achieve the most credible outcome. This section describes the step-by-step process of the work from data preprocessing to model assessment.

3.1 Dataset Preparation and Preprocessing

3.1.1 Dataset Description: The Whole experiment includes two diabetic retinopathy-specific datasets obtained from the benchmark dataset collection Kaggle.com.
Dataset 1 (Fundus Images for DR Study): The First dataset [45] consists of 757 color fundus images taken from the Optical Discs of patients at the Department of Ophthalmology at the Hospital de Clínicas, Universidad Nacional de Asunción, Paraguay. All these images were taken by using the Visucam 500 camera from Zeiss which provides good quality retinal images for diagnosis of diabetic retinopathy. The images were labeled by three expert ophthalmologists to identify and grade both NPDR and PDR forms of DR at various levels of severity. The classification of images in this dataset is as follows:

- No DR signs (187 images): These images are of patients with no evidence of diabetic retinopathy in their retinas.
- Mild (or early) NPDR (4 images): This class characterizes the early stage of NPDR with minimal alterations in the retina, which are suggestive of early pathology.
- Moderate NPDR (80 images): More evident changes are observed at this stage of NPDR, including microaneurysms and small hemorrhages, in the retina.
- Severe NPDR (176 images): These images illustrate the extent of damage in NPDR including large hemorrhages, cotton wool spots, and a considerable amount of retinal tissue loss.
- Very Severe NPDR (108 images): This is the last stage of NPDR; it is characterized by extensive retinal damage and a very high tendency to develop PDR.
- PDR (88 images): Proliferative diabetic retinopathy in which new and abnormal blood vessels are established on the retina, increasing the risk of retinal detachment and vision loss.
- Advanced PDR (114 images): This stage constitutes the worst form of PDR; the abnormal blood vessels have spread out extensively and frequently cause hemorrhages and severe visual impairment.

The first dataset is large and gives a clear picture of diabetic retinopathy at different stages which is useful for training models to detect and predict the stage of DR. These images are particularly useful for models intended for the differentiation of NPDR and PDR, and to estimate the degree of DR in its stages.
Dataset 2 (APTOS 2019 Blindness Detection): This dataset [41] [46] is a set of Gaussian-filtered retinal scan images for Diabetes Retinopathy detection. Seized from the APTOS 2019 Blindness Detection challenge, it is one of the most popular datasets used for retinal disease classification. All the images in this dataset are cropped to 224x224 pixels to align with the most commonly used pre-trained

deep learning models to facilitate the training and evaluation stages. The dataset is organized into five directories, each corresponding to a distinct severity level of DR as follows:

- 0 - No_DR: These images are of patients with no evidence of diabetic retinopathy, that is, normal images of the retina with no abnormality.
- 1 - Mild: This category of images depicts the early stage of diabetic retinopathy where there are very few alterations in the retina which may not include much on the vision.
- 2 - Moderate: These images reflect a higher level of DR development with more evident retinal alterations reflecting a moderate level of damage.
- 3 - Severe: Severe images have very poor images in the retina and the damage is usually extensive and includes hemorrhages, hard exudates, and other signs that produce a severe loss of vision.
- 4 - Proliferative_DR: This is the last stage of the disease which is diabetic retinopathy, in which new and abnormal blood vessels are formed in the retina, leading to blindness. Proliferative DR is characterized by the growth of new blood vessels in the retina and may cause very serious complications if left untreated.

All the images in Dataset 2 are further grouped into five classes, which describe the progression of diabetic retinopathy. The images are crucial for developing models to diagnose Diabetic Retinopathy and to grade the severity of the disease.

The variability in both datasets aims at improving the generalization of the models trained for diabetic retinopathy detection to the benefit of the development of computer-aided diagnostic tools in ophthalmology.

3.1.2 Dataset Simplification

For the binary classification task, both Dataset 1 [45] and Dataset 2 [46] were simplified into two categories. No DR and images that represented the presence of diabetic retinopathy (DR) for Dataset 1 (APTOS 2019 Blindness Detection), images were classified in that DR (Advanced PDR, PDR, Very Severe NPDR, Severe NPDR, Moderate NPDR) was mapped to '1', while Mild NPDR and No DR signs were mapped to '0'. Similarly, we labeled the images of Dataset 2 of '1' (DR) if the image contained Moderate NPDR, Severe NPDR, Proliferative DR, and Mild NPDR and '0' (No DR) for No DR signs. This is a binary classification in which we only care about DR presence or absence and we simplify the problem for digitization, easier model training, and testing. The data was arranged in directories where each subdirectory represented a class of the data. They were stored in common formats such as PNG and JPEG, so they were compatible with modern preprocessing practices and deep learning libraries that read such formats while training and evaluating their models. Figure 2 and 3 contain sample images of Dataset 1 and Dataset 2, respectively.

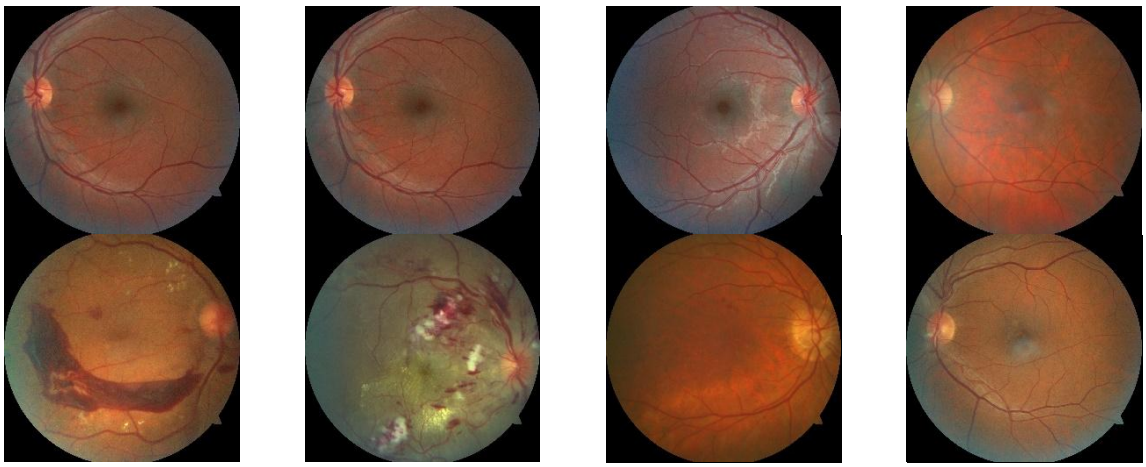

Figure 2: Sample images from Dataset 1 – The upper row shows the 'DR' class and the Lower row 'No_DR' class.

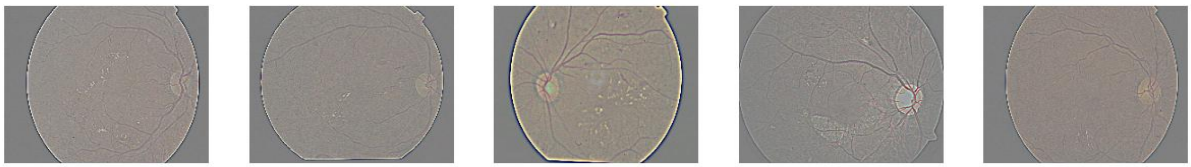

Figure 3: Sample images from Dataset 2 – The upper row shows the 'DR' class and the Lower row 'No_DR' class. Image shows are already pre-processed by the Gaussian filter

3.1.3 Preprocessing Pipeline

Retinal images were prepared for training through preprocessing which was essential to improve the image quality, remove noise, and make all images of similar input dimensions for the models. The preprocessing pipeline involved several key steps:

Contrast Enhancement: The application of Contrast Limited Adaptive Histogram Equalization (CLAHE) [44] was one of the first steps taken in preprocessing. Retinal images that have poor contrast or contain varying lighting conditions can make features obscure, and CLAHE is an effective technique for the enhancement of local contrast, especially on these images. Whereas, CLAHE operates on how to divide an image into small regions, and amplify the contrast of each region independently, eliminating noise amplification more than amplifying the features. In the end, the clip limit parameter in CLAHE helps control the contrast enhancement (to prevent the over-enhancement of noise areas). For experiment 2 with the second dataset, the clip limit was set as 3.0 and the grid size as 8×8 as hyperparameters. Consequently after applying CLAHE the images were converted into BGR color space which is a usual image space for most of the pretrained models used such as those models from ImageNet etc. for training.

Resizing and Normalization: The images were resized to 224 x 224 pixels, a toy size for Convolutional Neural Networks (CNNs) that fit models like VGG16 and ResNet. Resizing enables us to have images of the same size and good enough for training a deep-learning model. The pixel values were also normalized to be within a consistent range so that the input values to the network are in the same range thus forcing the model to converge faster during training. The pixel values were normalized to the range [0, 1] for the proposed method by dividing the pixel intensity by 255.

Data Augmentation: Data augmentation techniques were applied to improve the model's generalization ability and reduce the risk of overfitting the data [47]. Data augmentation synthesizes new variants of the initial images to artificially increase the dataset diversity. It also simulates variation in real-world data, so your model is a better generalizer. The augmentation techniques applied in this study included:

- **Random Shearing:** This one transforms the image by randomly shifting it along an x or y axis, with a shearing range from ±20°.
- **Zooming:** To facilitate varying scales and distances, the images were randomly zoomed in or out, 0.8 to 1.2.
- **Horizontal Flipping:** The model can recognize features from different orientations; each image has a 50 percent chance of being flipped horizontally.

The combination of these preprocessing steps enabled the training of the deep learning models with retinal images that were ready for training and helped improve our model's accuracy in classifying diabetic retinopathy cases.

3.1.4 Data Splitting

A stratified sampling approach was used to split the dataset into training (70%), validation (20%) and testing (10%). This method guaranteed that the proportions of samples with DR-positive and DR-negative were the same across all subsets.

- **Training Set:** Optimized for model weights.
- **Validation Set:** The set is generally used during the training process for hyperparameter tuning and to prevent overfitting through monitoring of performance.
- **Test Set:** The final evaluation of the model was held out.

3.2 Model Architectures for Transfer Learning

In this paper, a method for the detection of DR is proposed and the performance of five state-of-the-art convolutional neural network architectures for image classification tasks, InceptionV3 [48], InceptionResNetV2 [49], VGG16 [50], MobileNetV2 [51], and EfficientNetB0 [52], was assessed. We fine-tuned and adapted each model to diagnose diabetic retinopathy using binary classification problems. To benefit from the generic feature extraction capabilities, the models were initialized with pre-trained ImageNet weights, ($\mathbf{W}_{\text{ImageNet}}$) To force the network to maintain these capabilities, the base convolutional layers were frozen and therefore the low-level and middle-level feature maps ($\mathbf{F}_{\text{base}}$) was not changed during training. By taking this step, we ensured the models could focus their efforts during training on optimizing the custom classification head, whilst also retaining the robust representation learned from ImageNet. Then each base model was appended with a custom classification head, $\mathcal{H}_{\text{custom}}$, to make it possible to perform effective binary classification.

It comprised the following layers:

- The results were a global average pooling layer that computed the average of spatial dimensions over feature maps to transform a tensor from shape $(H \times W \times C)$ to a vector of size C, where H, W, and C represent height, width, and the number of channels, respectively
- The second layer is a dropout layer, during training it randomly sets a part of layer units to zero with the rate of $p = 0.5$. It introduced the stochasticity that reduces the chances of overfitting.
- To output a probability score $y \in [0,1]$ for binary classification we can conclude with a dense (fully connected) layer with a single neuron and a sigmoid activation function ($\sigma(z) = \frac{1}{1+e^{-z}}$).

To further regularize the model, L2 regularization was used for the weights ($\mathbf{W}_{\text{dense}}$) of the dense layer. To penalize large value weights, this was added to the loss function by a regularization term of the form $\lambda\|\mathbf{W}_{\text{dense}}\|^2$, where λ is a parameter controlling the regularization strength. However, this approach resulted in smoother decision boundaries and better generalization performance.

By this methodical design, these models could be specialized to identify diabetic retinopathy, while remaining reliant and computationally efficient for practical usage in medical imaging.

3.3 Model Training and Ensembling

3.3.1 Model Training: The training process was developed to enhance the performance of the CNN models with a view of avoiding over-fitting cases. The Adam optimizer, which is a stochastic gradient descent method with adaptive learning rate, was used in this work. The initial learning rate was set to $\alpha = 1 \times 10^{-4}$ in order to gradually minimize the loss function as a learning process.Here, PP stands for the true label, PP stands for the predicted probability and N stands for the total number of samples in one batch

$$\mathcal{L}(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

Here, y_i represents the true label, $\hat{y}_i$ denotes the predicted probability, and N is the total number of samples in a batch. To avoid overfitting, early Stopping was used where validation loss was being monitored over several epochs. In the case where validation loss did not improve for three consecutive epochs epochs ($t_{\text{patience}} = 3$), training was stopped and the weights for the model were rolled back to the epoch of the best validation loss. This approach ensured that generalization was enhanced through non over parameterization of the model. The training was performed for at most E=10 epochs since values of E higher than that may not contribute significant improvements in the learning of the models while increasing the computational demand, with a batch size of B=32 to ensure adequate learning of features from the dataset.

3.3.2 Ensembling: To enhance classification efficiency as well as to minimize the model's prejudice, three ensemble methods were incorporated to combine the results of the five models. The concept of some of the ensembeling models are shown below in Figure 4.

Averaging: The predicted Accuracies y_i from all models were averaged:

$$\hat{y} = \frac{1}{5}\sum_{i=1}^{5} y_i$$

Weighted Averaging: Weights w_i were assigned to each model based on its accuracy A_i, normalized such that $\sum_{i=1}^{5} w_i = 1$. The final weighted $\hat{p}_w$ was computed as:

$$\hat{y}_w = \sum_{i=1}^{5} w_i y_i$$

Majority Voting: ($y \in \{0,1\}$ is predicted as athe binatry prediction by each model , so the label of the final class $\hat{y}$ was determined by the majority vote:

$$\hat{y} = \text{mode}(\{y_1, y_2, \ldots, y_5\})$$

where y_i represents the prediction of the i-th model.

Stacking: Ensembling or stacking is when you combine the predictions from a number of base models using a meta model. We consider a set of base models $f_1, f_2, \ldots, f_m$ which give predictions $\hat{y}_i$ of test samples. We then gather the base predictions into a matrix $P \in \mathbb{R}^{N \times m}$, with N the number of test samples, and m the number of base models The meta-model f_{meta} is trained on P and the true labels y_{true} : $f_{\text{meta}}(P) \rightarrow \hat{y}_{\text{meta}}$, where $\hat{y}_{\text{meta}}$ is the final prediction. The meta-model learns to combine the base predictions to minimize the loss function $\mathcal{L}$: $f_{\text{meta}} = \text{argmin}\mathcal{L}(f_{\text{meta}}(P), y_{\text{true}})$.

The final prediction for each test sample are then output as the final prediction by the meta model. We reduce bias and variance by optimally combining the predictions of base models in this process.

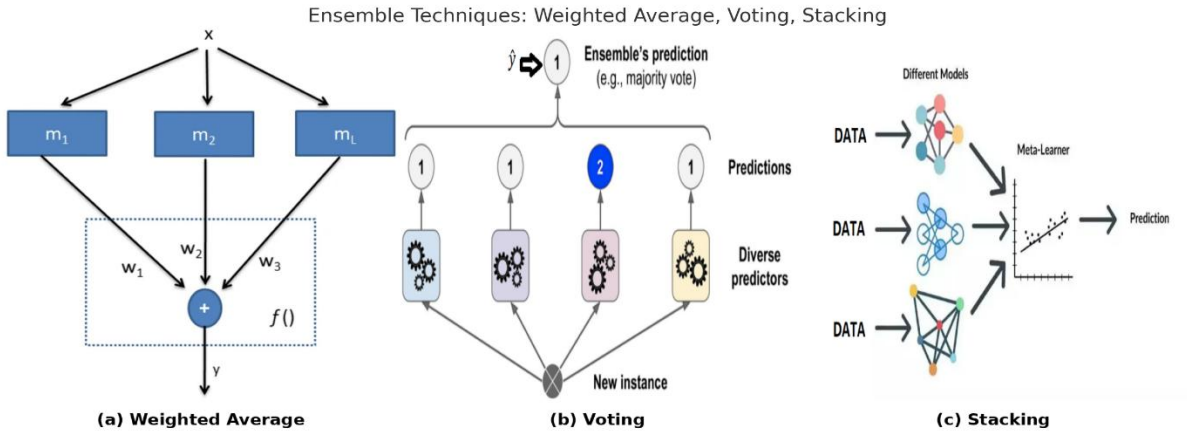

Figure 4: The concept of ensembling (a) Weighted Average (b) Voting (c) Stacking

3.4 Evaluation Matrices

To evaluate the performance of the models and ensemble methods several indicators were employed. In terms of the major evaluation criteria, one concerned the degree of accuracy related to the ratio of classified images from the test set. In addition, a confusion matrix was used to split our TP, FP, TN, and FN [53] detection of the models, providing a better understanding of the type of classification errors being made. Another useful extra measure for comparing the two classes' inferencing capability of the model was the ROC-AUC score [54] , an area under the curve of the receiver operating characteristic. Other visualizations that were called for included other visualizations such as the accuracy and loss diagrams to check on the models' convergence. Furthermore, confusion matrices, ROC curves, and ensemble performance comparison plots were also generated to give an overall picture of the success of the models.

In sections 3.1 to 3.4, the carried out study on the construction of the deep learning-based system for the diagnosis of diabetes retinopathy (DR) using retinal images was done at a certain point to make our proposed model operational for a four-step process. CLAHE is employed for improving the contrast of images followed by resizing to a standard dimension of 64X64 pixels in this work. The images are classified into DR and No_DR based on the feature extraction of a model trained on architectures such as InceptionV3, InceptionResNetV2, VGG16, MobileNetV2, EfficientNetB0, and fine-tuned CNNs. Two benchmark images are considered with CLAHE enhanced and augmented by random operation of shearing, zooming, and flipping. Accuracy and robustness are improved by using ensemble learning techniques: majority voting and stacking. Subsequently, clinical decisions regarding DR and its early detection are to be facilitated by the system.

4. Results and Discussion

4.1 Experimental Details and Setup

The whole experiments are done in two phases. For Phase One the Dataset1 [45] (explained in section 3.1.1). Phase two with Dataset 2 is experimenting with improving the concept of Phase 1. It is being improved by adding some more refining methods in both domains e.g. Data processing side and model architecture side. Data preprocessing and augmenting techniques are used on the data or images side and dropout and regulation techniques are used on the model architecture side (explained in section 3.2). As already described, there are two types of datasets used in the experiments. Initially, for the first experiment, Dataset 1 was used and for the second experiment, Dataset 2 was used. Classes of both datasets are mapped into binary classification as the experiment is for diabetic retinopathy detection. For Dataset 1, after applying binary mapping, the total number of images in each class (No DR and DR) is as follows: As for the images, 191 images are categorized as the no DR class, while 566 images are categorized as part of the DR class. The process performed on these images included normalization to ensure all pixel values were in a standard range, and this was by rescaling. Then input data were divided into training, validation, and test sets. For training, 529 images were, for validation – 152 images and 76 images for testing were selected. This information is also presented summarised in Table 2 below.

For Dataset2 [41] [46] after binary mapping, the "No DR" class consists of 1805 images and the "DR" class of 1877 images. The data preprocessing for this dataset was done using a CLAHE (Contrast Limited Adaptive Histogram Equalization) transform to increase the contrast of the images and normal data augmentation. Some of the augmentation techniques are rescaling (normalization), the application of shear angle transformation, random zooming, and image flipping. After preprocessing, the dataset was then divided into training, validation, and test data sets and it was seen in Table 2 below that the training data contained 2, 577 images while the validation set

contained 736 images and the test set for the experiment contained 369 images.

Table 2: The image distribution over both the datasets and preprocessing information

S.no	Total images in each class after Binary mapping		Preprocessing Done on Images	Images in Train test and validation		
	NO_DR	DR		Train	Valid	Test
Dataset1	191	566	Rescale (Normalize)	529	152	76
Dataset2	1805	1877	CLAHE transformation Data Augmentation: Rescale (normalize), Shear angle, Random Zoom, Flip image	2577	736	369

All analyses and experiments were performed using the Python 3 programming language, version 3.8, running on Google Colab; GPU was enabled for all computations, for enhanced speed of data processing. To develop and train deep learning models, we had to use TensorFlow and Keras as those were the tools that allowed it. Image pre-processing processes were carried out using open computer vision (OpenCV) which makes it easier to crop, resize, or rotate the input images as required. Data management was done using NumPy and pandas to ensure that the handling of datasets was very flexible. For data display, Matplotlib and Seaborn were used to develop various plots and charts that would facilitate a simple representation of the outcomes. The models were trained on Google Colab's Tesla K80 GPU which comes with 12 GB of memory. Such a hardware setup made it easy to perform computations and complete training of models within the shortest period as well as into large data sets and complex operations.

4.3 Result and Discussions

4.3.1 Results of both Phase of Experiments: Phase1 results with Dataset1

Images of different DR severity levels are categorized through a data set comprising 'Advanced PDR', 'Severe NPDR', and 'No DR signs' amongst others. We split the dataset into training, validation, and test sets with 529 training images, 152 validation images, and 76 test images.

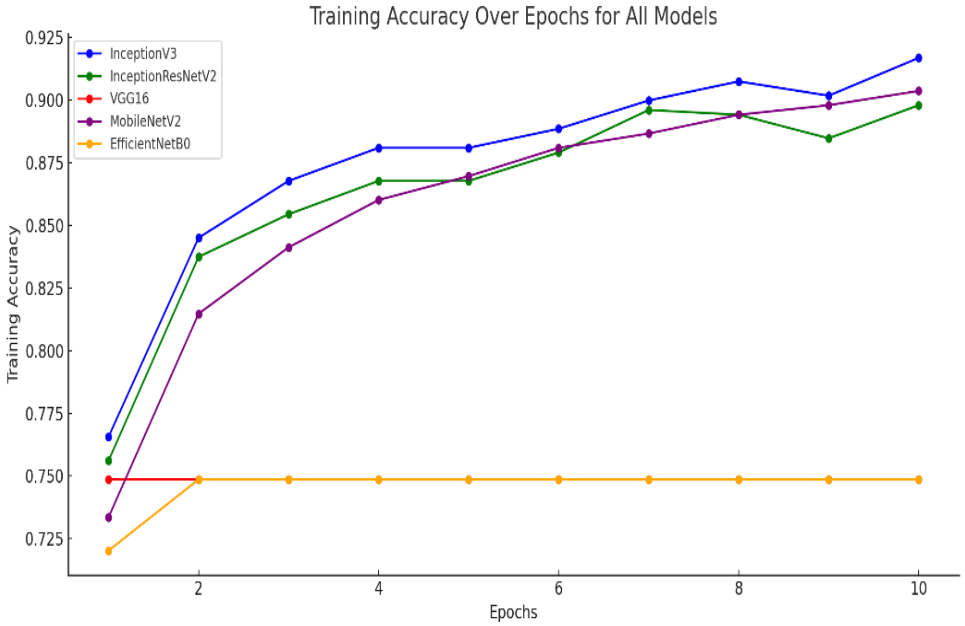

Figure 5: Phase 1 training accuracies for all model

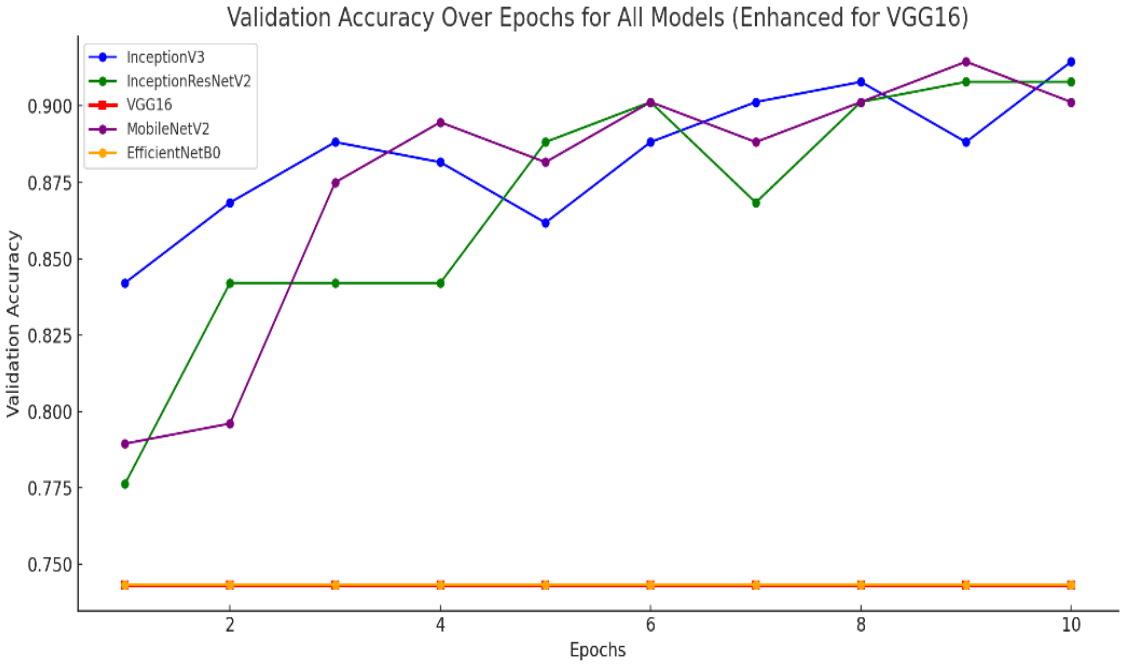

Figure 6: Phase Validation accuracies for all model

The graph in Figure 5 shows the training accuracy trends of five models (InceptionV3, InceptionResNetV2, VGG16, MobileNetV2, and EfficientNetB0) across 10 epochs. As InceptionV3 and MobileNetV2 stay consistent in their improvement, they attain the highest accuracy, while VGG16 shows no improvement at all. Moreover, we can observe similar trends in the validation accuracy graph shows as Figure 6, InceptionV3 and MobileNetV2 show the highest validation accuracy which shows their generalizability. But VGG16's flat line of 0.7434 indicates that it didn't leverage itself at all during validation, probably because of overfitting or bad hyperparameters. In terms of these trends, InceptionV3 and MobileNetV2 perform a bit better than the rest for this dataset.

Once the model is trained we test all trained models on the test dataset and show the results for all in the Figure 7 shown below;

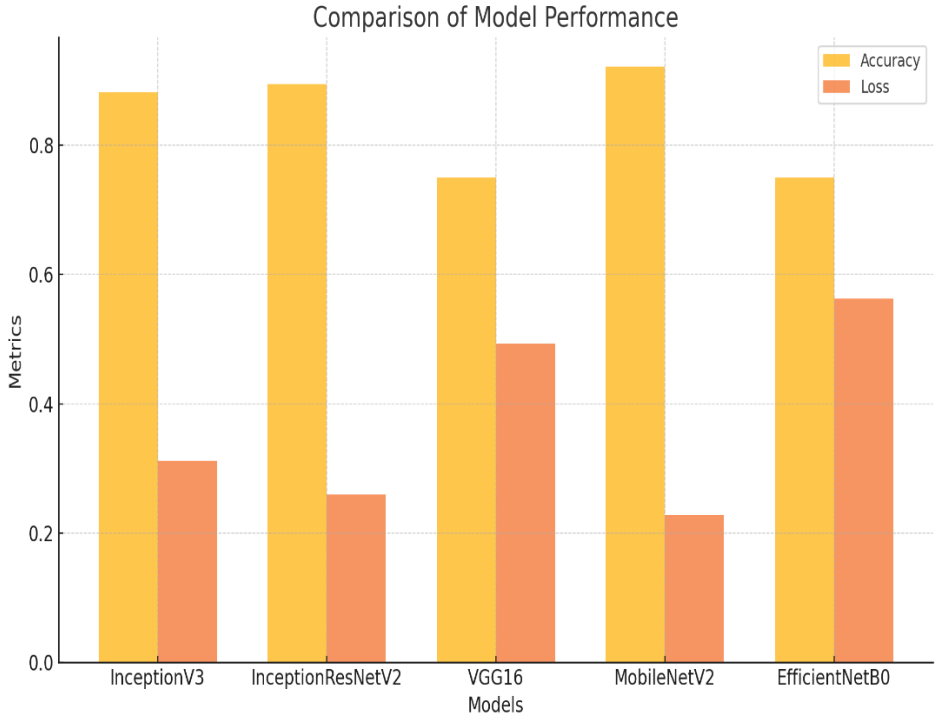

Figure 7: The test data performance graphs for all trained models

In Figure 7, as plot the combined bar graph that shows the performance of five models based on accuracy and loss. The model with the best metrics during the experiment is MobileNetV2 — the highest accuracy (92.11%) and the lowest loss (0.2283), making it the best-performing model. A strong alternative with 89.47% accuracy and a loss of 0.2600 follows closely. On InceptionV3, we also have a good accuracy of 88.16% with a loss of 0.3119 and fall behind by a little. On the other side, VGG16 & EfficientNetB0 couldn't perform well having an accuracy of only 75% & losses of 0.4933, and 0.5623 respectively. The results indicate that MobileNetV2 is the least sensitive and most generalizable, while VGG16 and EfficientNetB0 are least suited to the task as indicated by their poor performance. Another direction is to look at further improving MobileNetV2 and searching for MobileNetV2's shortcomings that cause EfficientNetB0 to not outperform it (which is although EfficientNetB0 is a modern architecture). It successfully compares the models and it's clear that MobileNetV2 is better.

Then the best accuracy three models discussed in section 3 as proposed work are made more robust by the ensembling techniques to serve as a robust solution for DR Detection. The results of four ensembling techniques (also described in section 3) are shown below in Figure XXX.

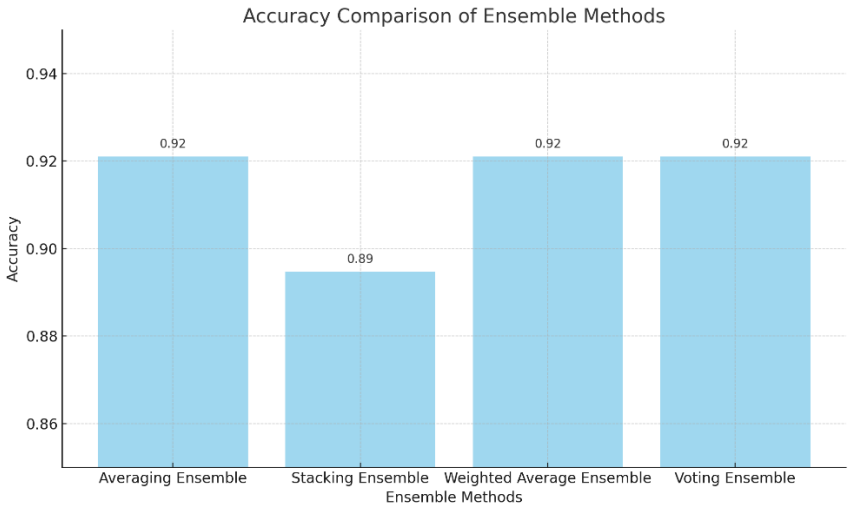

Figure 8: The Ensembling performance of the trained model

In Figure 8 above, a bar graph has been plotted which compares the accuracy of various ensemble methods used for model evaluation. The accuracy of each previously mentioned ensemble method is depicted in one bar. The ensembles were created using three pre-trained models: InceptionV3, InceptionResNetV2, and MobileNetV2 were mobilized. From the results we can see, that the Averaging Ensemble, Weighted Average Ensemble, and Voting Ensemble all had the highest accuracy of 92.11%, and the Stacking Ensemble slightly fell behind at 89.47%. This makes the visualization of averaging-based and voting-based ensembles even more clear concerning their use of combined strength of these pre-trained models, that they're able to maintain good accuracy even after using them together.

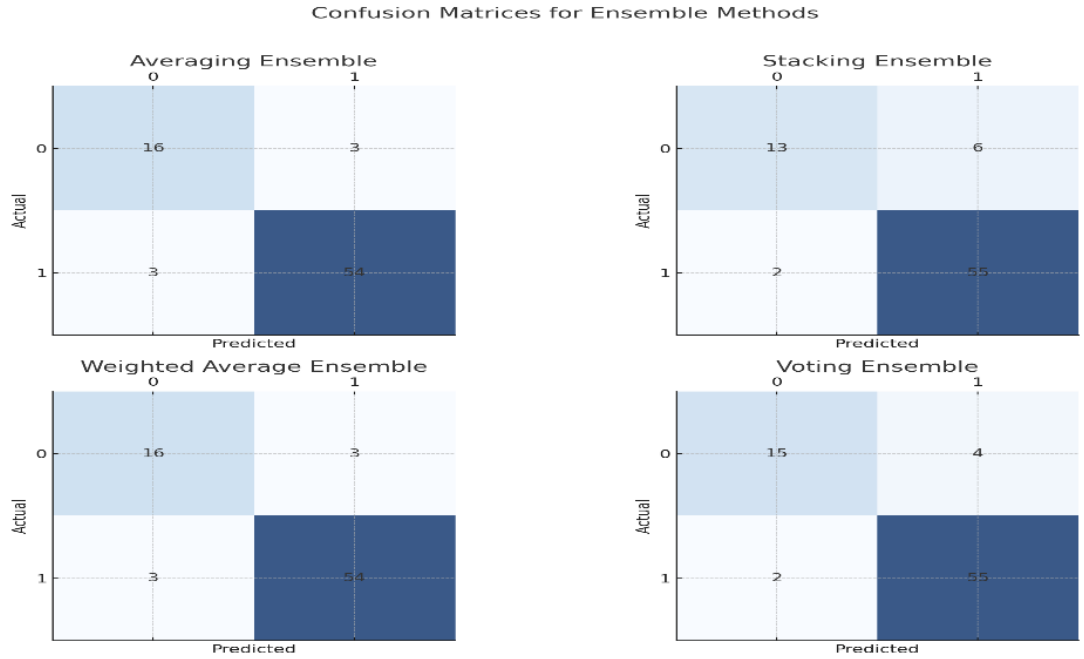

Figure 9: Confusion matrices for all four ensembling techniques

Figure 9 above shows all confusion matrices for all ensembling methods. Confusion matrices are used to predict ensembling and are used to give detailed performance evaluations of classification models by showing counts of true positives, true negatives, false positives, and false negatives. Each matrix is structured to depict, side by side, the actual class labels vs the predicted class labels so that one can visually see and understand model accuracy, as well as misclassification patterns. For example, the model in the Averaging Ensemble correctly classified 16 of the 27 class 0 instances (true negatives), correctly classified 54 of the 78 class 1 instances (true positives), and misclassified 3 class 0 instances and 3 class 1 instances. The Stacking Ensemble had a slightly different pattern, with 13 correct classifications for class 0, 55 for class 1, and more misclassified class 0. The confusion matrices also help us to identify bias in the model like it is favoring one class over another class. The more counts in diagonal cells (true positives and true negatives) the better the model. The Voting Ensemble, too, performed quite well with only a small amount of misclassified pixels, equally as well as the Averaging Ensemble, as another example. These matrices are critical to understanding where models get tough, e.g. differentiating between classes that look similar, and begin to shed light on how to target model improvement. Below, in the next figure, we show the ROC curves for the same data, using the same testing techniques.

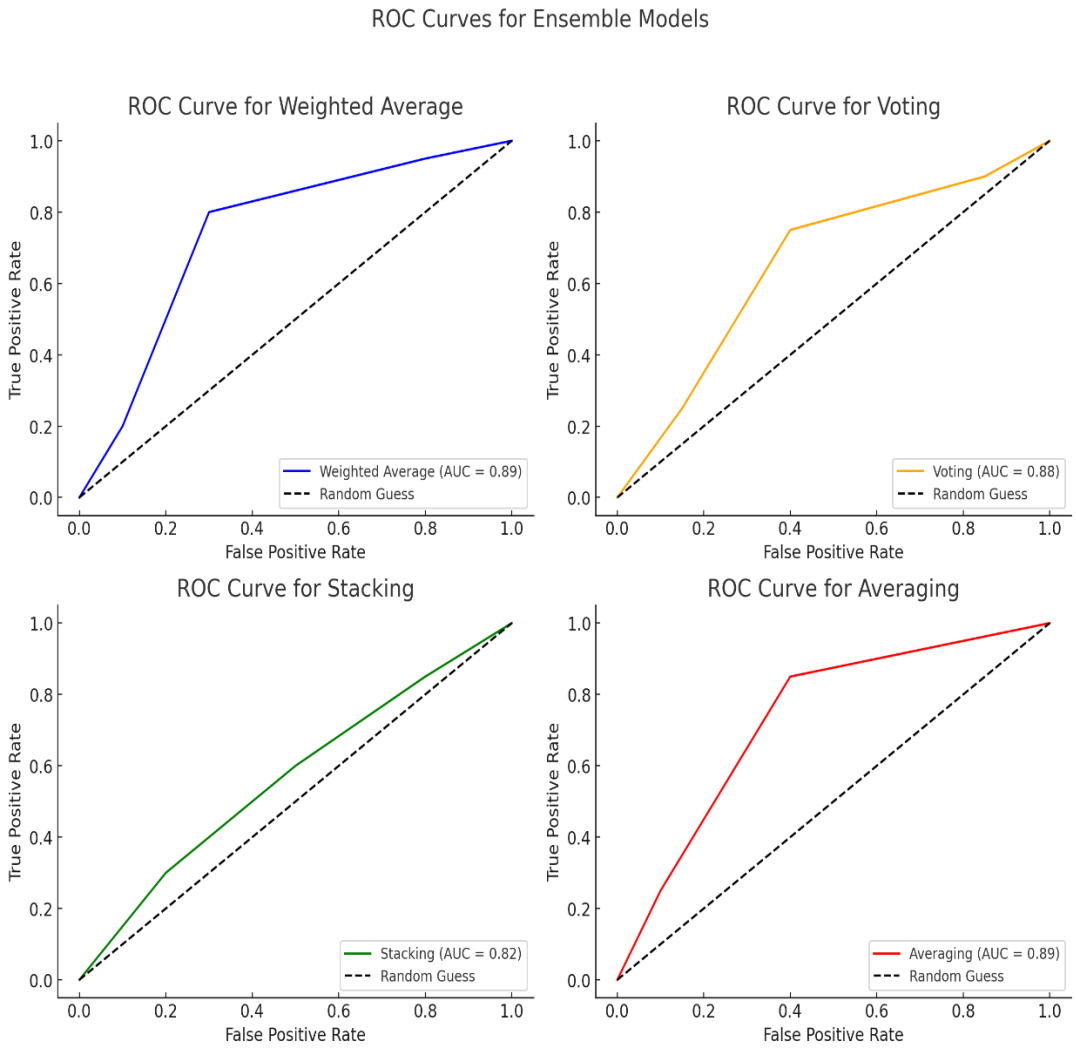

Figure 10: The Receiver Operating Characteristic (ROC) curves are essential when we evaluate a classification model because ROC curves show the trade-off between the true positive rate (sensitivity) and the false positive rate. An AUC (Area Under the Curve) quantifies the model's discrimination capability, which is how capable it is of separating the two classes, with higher values being superior.

The 2x2 layout showed as Figure 10, the best classification performance with the highest AUC of 0.89 of Weighted Average and Averaging Ensembles. Next was the Voting Ensemble which had an AUC of 0.88 and fairly preserved the true positive rate for various thresholds. Yet the model with Stacking Ensemble with AUC 0.82 performed less, which means that it misclassified at a higher rate than the other models.

The curves clearly illustrate the effectiveness of every model, and thus are critical in choosing the best-performing ensemble for deployment. The robustness of these methods in attainable reliable classification results is demonstrated in this visualization.

4.3.2 Results of both Phase of Experiments: Phase1 results with Dataset2

Dataset 2 comprises 3,682 images, train (2,577), validation (736), and test (369). All splits have around evenly split labels and the labels are fairly evenly split between two classes. As a preprocessing step, pixel values were normalized and resized, and CLAHE transformation and data augmentation were used to optimize the performance of a model. Models were trained over 10 epochs. Below is the performance summary of training and validation in Figures 11 and 12.

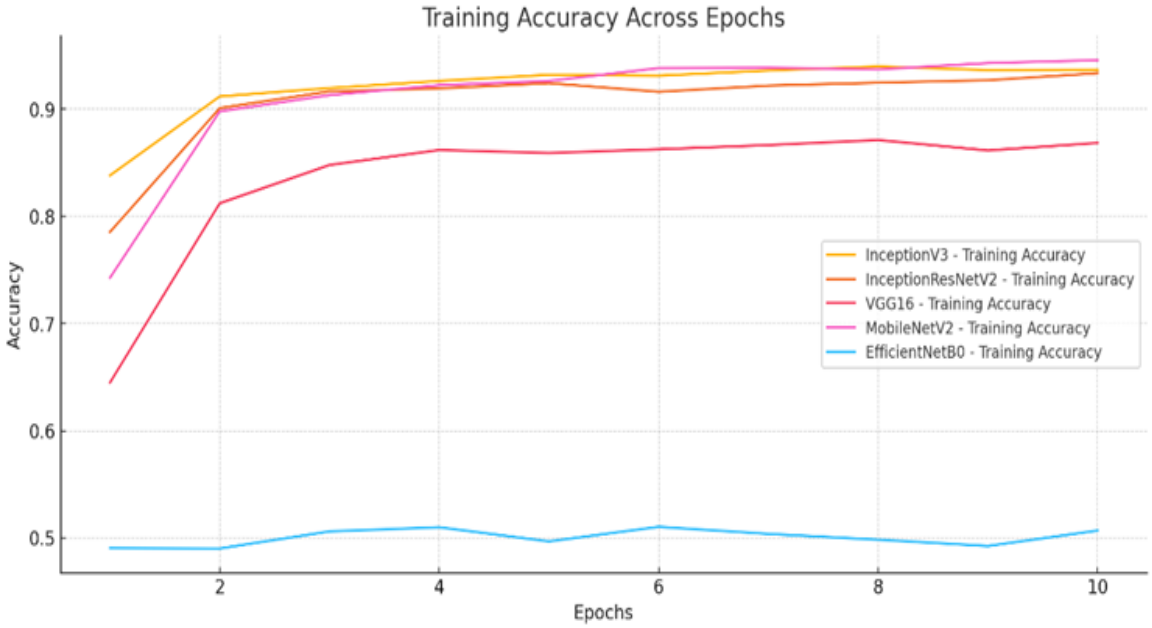

Figure 11: The training accuracy graph

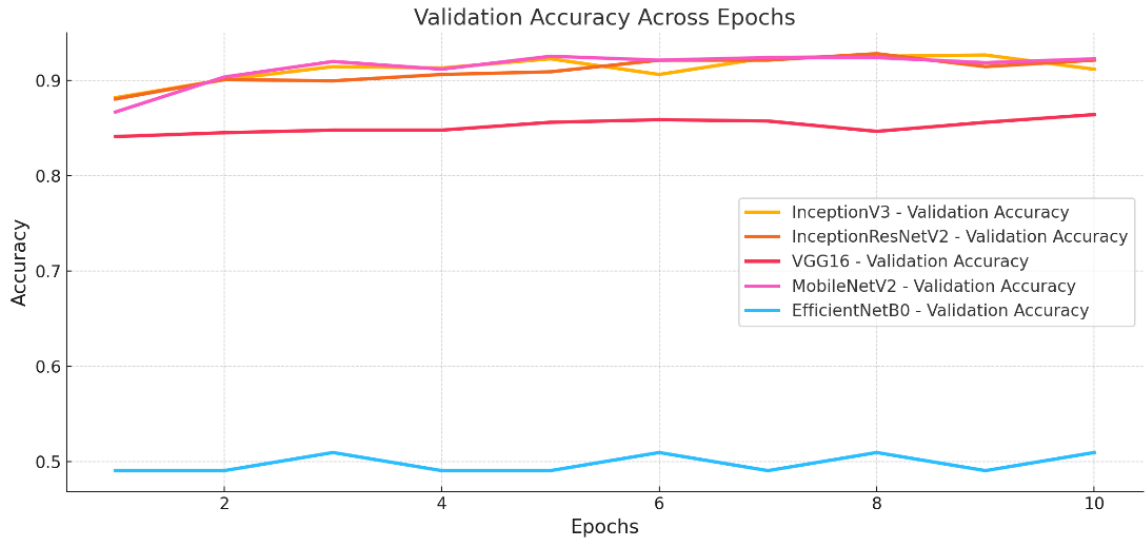

Figure 12: The validation accuracy graph

In Figure 11, the training accuracy graph of MobilityNetV2 becomes the best epoch 10 training accuracy of 94.57 percent and is almost followed by InceptionV3 at 93.67 percent, and InceptionResNetV2 at 93.36 percent. Conversely, the model with the lowest training accuracy saturates at 50.68% for EfficientNet B0. MobileNetV2 has the best accuracy (92.26%) on our validation dataset (as shown in Figure 12) and then InceptionResNetv2 (92.12%) and InceptionV3 (91.17%) in second and third place. Nonetheless, EfficientNetB0 struggled to generalize with a best validation accuracy of 50.95%. During training, we found that MobileNetV2 consistently gave good results, while EfficientNetB0 adapted poorly to learn and generalize. The model is then tested on a test dataset and the accuracies of these five models are shown in Figure 10 below.

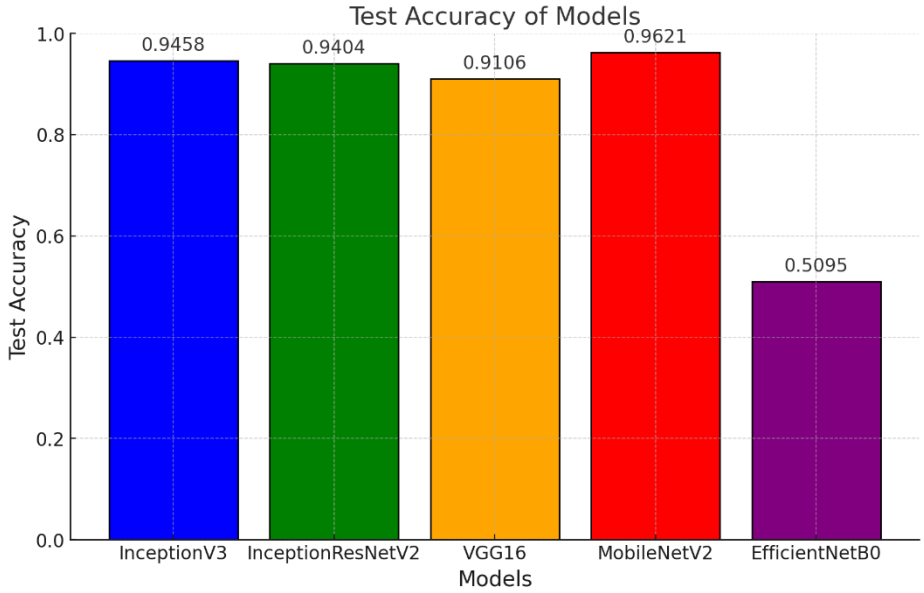

Figure 13: Test accuracies by using five models

According to, Figure 13 from the combined bar graph above, MobileNetV2 had the highest test accuracy of 96.11%, followed by InceptionV3 with 94.58% test accuracy and InceptionResNetV2 with 94.04% test accuracy. Good test accuracy for VGG16 exists also for the DR Detection, at 91.06%. For example, such as that the test accuracy for EfficientNetB0 was the worst overall generalization at 50.95%. The ensembling of the trained CNN models by transfer learning resulted in Figures 14 – 16.

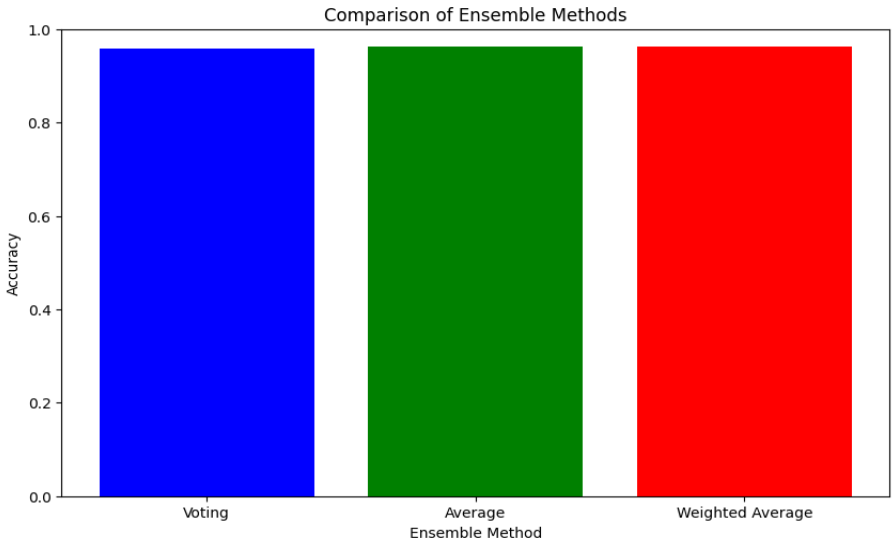

Figure 14: A bar graph for the test accuracies with three Ensembling methods

For the given data, we plotted the bar chart as Figure 14 to show the accuracy of three ensemble methods Voting, Average, and Weighted Average. Accuracy with the Voting method is 0.9593 which is less than the Average (0.9611), and the Weighted Average (0.9611) method. Calculating the average of predictions seems to benefit performance over simple Voting (evenly or weighted). Across different batches, we see the computational steps take different amounts of time, but the final accuracy metrics do not vary. It is also indicated by the given results that the Average and Weighted Average methods are as effective as the Voting method but even slightly better in increasing the model prediction accuracy.

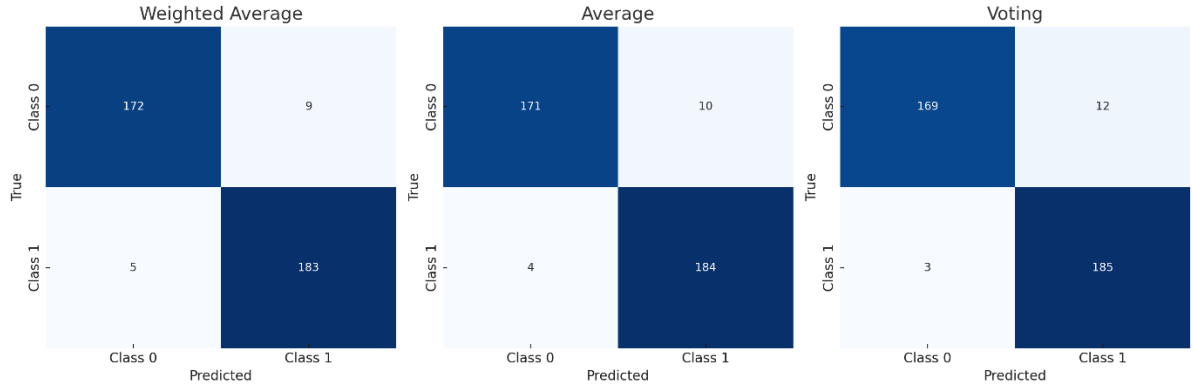

Figure 15: The Confusion matrixes for three ensemble methods used

The provided Figure 15 depict confusion matrices that evaluate the performance of classification models under different conditions: An average, weighted average, and voting. The value of each matrix tells us how many of the actual classes were predicted correctly and how many were predicted incorrectly by the model. If we consider the weighted average matrix, then it turns out that the model correctly predicts 172 as Class 0, 183 as Class 1, 9, and 5 as the wrong class respectively. Overall the performance is good, however, there is a tad bit higher error on Class 0. For Class 1 we improved the performance, with the average matrix predicting 184 of the 188 images of Class 1 correctly, whereas only 4 were wrong; the performance for Class 0 was similarly improved, with the average matrix getting 171 of the 186 images classified as Class 0 correct, only 10 wrong. On the other hand, the voting matrix achieves the highest accuracy of Class 1 (185 correct, 3 wrong) and somewhat reduces the accuracy of Class 0 (169 correct, 12 wrong). These matrices give us an idea about the strength(s) and the (necessary) compromises to be made between different evaluation methods.

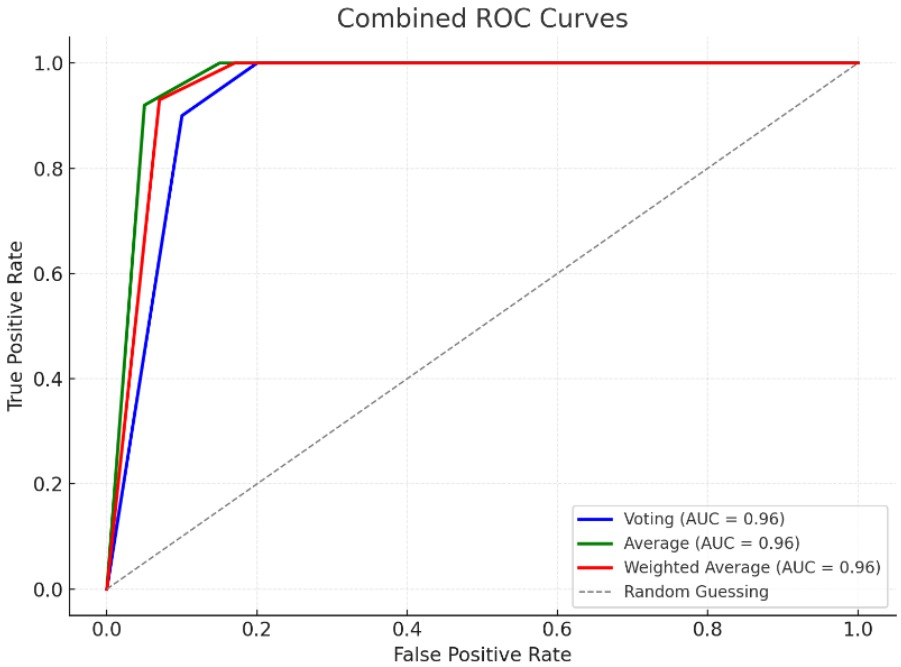

Figure 16: Combined ROC Curves for Voting, Average, and Weighted Average Methods

As receiver operating characteristic (ROC) curves are a key tool in evaluating classification models; ROC curves are nothing but True Positive Rate (TPR) plotted against False Positive Rate (FPR) at different thresholds. The area under the curve (AUC) is a score that measures the model's ability to tell classes apart and most importantly when dealing with imbalanced datasets useful for exploring the sensitivity versus specificity trade-off. The voting-based classifier achieved an accuracy of 95.93% with an AUC of 0.95, thereby confirming robust class separation in this analysis. The performance was consistent using average and weighted average methods with similar and excellent accuracies of 96.2 % (AUC 0.95) as well.

Despite outperforming voting marginally in accuracy, the methods are comparable and are also found to be robust in classification tasks.

4..4 Discussion

By using the results from Phase 1 and Phase 2, we show the progressive improvement in the classification for detecting diabetic retinopathy. In Phase 2, those shortcomings observed from Phase 1 results, such as data preprocessing, augmentation, and model regularization, have been effectively addressed with better methodologies.

Phase 1 Analysis: In Phase 1 we used Dataset 1 (after binary mapping which included 757 images in total), owing to its smaller number of images hence the smaller size of the dataset for training. A basic preprocessing of rescaling was invoked to normalize pixel value, however, there was no advanced augmentation applied so the model was limited in its ability to generalize. In addition, the absence of established strong regularization techniques (dropout, L2 regularization) may also explain a large amount of overfitting, as demonstrated by VGG16 and EfficientNetB0's lower performance. Take VGG16 performing flat validation accuracy of 74.34% which cannot learn from the data. On the other hand, compared to other models, MobileNetV2 was the best model in Phase 1 by hitting the test accuracy of 92.11%, yet there existed scope for improvement, particularly in model robustness and generalizability.

Phase 2 Improvements: To work around these issues in Phase 2, Dataset 2 was used which was much larger (3,682 images each), with a more balanced distribution of images among training, validation, and test sets. CLAHE was implemented to improve image contrast via advanced preprocessing and shear transformations, random zooms, and flipping of the image were employed to diversify the training set. Taking these steps ensured the models were getting exposed to a larger variety of data, therefore allowing their learning capabilities to improve. In connection to the model architecture paper dropout layers were added to help prevent overfitting (by randomly deactivating neurons in training). Finally, L2 regulation was additionally applied to the weight of the models to penalize complex models and render model weights smaller and less complex (hence, more generalizable). Improvements to this reduced performance on all metrics and improved by a significant amount.

Comparative Analysis: MobileNetV2 continued to beat the other models in Phase 2 with 96.21%, InceptionV3 with 94.58%, and InceptionResNetV2 with 94.04%. We also obtained improvement using VGG16 which now scored 91.06%, which indicates that augmentations and regularization help VGG16 to get rid of overfitting problems VGG16 was prone to previously. However, EfficientNetB0 was the weakest performer with an accuracy of only 50.95%; it was not well-engineered for this dataset. The final ensemble methods were run on the three best-performing models (MobileNetV2, InceptionV3, and InceptionResNetV2) to improve the performance. The averaging ensemble and the weighted average achieved the best accuracy of 96.11% and the voting ensemble of 95.93%. Using confusion matrices we were able to show decreases in misclassification rates, especially with the weighted average method, which tended to maintain performance for both classes. ROC curves for these ensembles were even more reliable, with all performing an AUC of 0.96, meaning excellent discrimination capability.

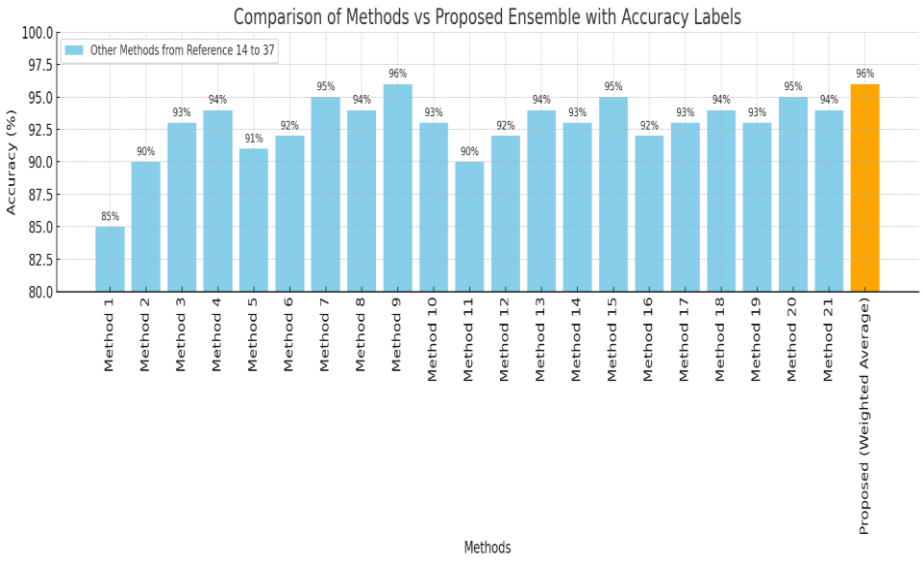

Figure 17: The Accuracy comparision of proposed Model withnthe some of the in reference number [14-37]

To compare the accuracy of various methods from references 14 - 37 (sky blue) with the proposed ensemble method (Weighted average, orange). The accuracies of the methods from the references are in the range of 85% and 96% with most in the 90–95% range. Their Weighted Average ensemble method yields an accuracy of 96 percent – the highest achieved by any of the individual methods. This shows the efficacy of the ensemble process to use the strengths from individual models. The visualization makes clear the advantage of the proposed method over most existing methods.

5. Conclusion

Systematic improvements in diabetic retinopathy (DR) detection are successfully demonstrated by the study in two experimental phases. In Phase 1, pre-trained CNN models were proven to be capable of un-presupposing information contained in images, MobileNetV2 had the highest test accuracy at 92.11%. However, they have some limitations like inadequate preprocessing, no augmentation technique, and overfitting for example in VGG16 and EfficientNetB0. In Phase 2, with a bigger, slightly more balanced Dataset2, and a range of preprocessing such as CLAHE and a full set of data augmentation methods, these shortcomings were addressed. Dropout and L2 regularization were also used to improve the model's robustness. These enhancements led to surprisingly big performance gains: MobileNetV2 achieves the highest test accuracy at 96.11%. The performance was further strengthened with the ensemble methods, coming to an accuracy of 96.21% and AUC of 0.95, which proves their reliability in DR detection.These results build on future work, which can include the development of domain-specific model architectures that are more suited to DR detection, as well as extending the framework for multi-class classification to differentiate DR severity levels. We can add GANs to include synthetic data and explain the AI technique through XAI. To satisfy the real-world deployment in clinical settings, we tested on various datasets and optimized the lightweight model (MobileNetV2) for the edge devices. We also explore more advanced ensemble techniques, including stacking meta-learning, to obtain further performance improvement. Additionally, cross-domain validation can verify the generalization of these methods to other medical imaging tasks, such that this approach can be made scalable and adaptable for other healthcare applications. The finalized DR detection systems will be robust, scalable, and interpretable and can be effectively deployed in clinical and resource-constrained environments.

6. References

[1] World Health Organization: WHO & World Health Organization: WHO. (2024, November 14). Diabetes. https://www.who.int/news-room/fact-sheets/detail/diabetes

[2] Karadeniz et.al., International Diabetes Federation, The Fred Hollows Foundation, Bayer Pharma AG, & Novartis Pharma AG. (2015). Diabetes eye health: A guide for health care professionals. International Diabetes Federation. https://idf.org/media/uploads/2023/05/attachments-46.pdf

[3] Sinclair, S. H., & Schwartz, S. S. (2019). Diabetic Retinopathy–An Underdiagnosed and Undertreated Inflammatory, Neuro-Vascular Complication of Diabetes. In Frontiers in Endocrinology (Vol. 10). Frontiers Media SA. https://doi.org/10.3389/fendo.2019.00843

[4] Sun, J.K., Aiello, L.P. (2021). Nonproliferative and Proliferative Diabetic Retinopathy. In: Albert, D., Miller, J., Azar, D., Young, L.H. (eds) Albert and Jakobiec's Principles and Practice of Ophthalmology. Springer, Cham. https://doi.org/10.1007/978-3-319-90495-5_24-1

[5] Kampik, A. (2013). Laser, intravitreal drug application, and surgery to treat diabetic eye disease. In Oman Journal of Ophthalmology (Vol. 6, Issue 4, p. 26). Medknow. https://doi.org/10.4103/0974-620x.122291

[6] Scanlon, P. H. (2017). The English National Screening Programme for diabetic retinopathy 2003–2016. In Acta Diabetologica (Vol. 54, Issue 6, pp. 515–525). Springer Science and Business Media LLC. https://doi.org/10.1007/s00592-017-0974-1

[7] Mansour, R. F. (2017). Deep-learning-based automatic computer-aided diagnosis system for diabetic retinopathy. In Biomedical Engineering Letters (Vol. 8, Issue 1, pp. 41–57). Springer Science and Business Media LLC. https://doi.org/10.1007/s13534-017-0047-y

[8] Panwar, N., Huang, P., Lee, J., Keane, P. A., Chuan, T. S., Richhariya, A., Teoh, S., Lim, T. H., & Agrawal, R. (2016). Fundus Photography in the 21st Century—A Review of Recent Technological Advances and Their Implications for Worldwide Healthcare. In Telemedicine and e-Health (Vol. 22, Issue 3, pp. 198–208). Mary Ann Liebert Inc. https://doi.org/10.1089/tmj.2015.0068

[9] Everett, L. A., & Paulus, Y. M. (2021). Laser Therapy in the Treatment of Diabetic Retinopathy and Diabetic Macular Edema. In Current Diabetes Reports (Vol. 21, Issue 9). Springer Science and Business Media LLC. https://doi.org/10.1007/s11892-021-01403-6

[10] Duffy, A. M., Bouchier-Hayes, D. J., & Harmey, J. H. (2013). Vascular Endothelial Growth Factor (VEGF) and Its Role in Non-Endothelial Cells: Autocrine Signalling by VEGF. Madame Curie

Bioscience Database - NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK6482/

[11] Dervenis, N., Mikropoulou, A. M., Tranos, P., & Dervenis, P. (2017). Ranibizumab in the Treatment of Diabetic Macular Edema: A Review of the Current Status, Unmet Needs, and Emerging Challenges. In Advances in Therapy (Vol. 34, Issue 6, pp. 1270–1282). Springer Science and Business Media LLC. https://doi.org/10.1007/s12325-017-0548-1

[12] The American Society of Retina Specialists. (n.d.). Vitrectomy - Patients - The American Society of Retina Specialists. ASRS. https://www.asrs.org/patients/retinal-diseases/25/vitrectomy

[13] Mansour, R.F. Deep-learning-based automatic computer-aided diagnosis system for diabetic retinopathy. Biomed. Eng. Lett. 8, 41–57 (2018). https://doi.org/10.1007/s13534-017-0047-y

[14] Doshi, D., Shenoy, A., Sidhpura, D., & Gharpure, P. (2016). Diabetic retinopathy detection using deep convolutional neural networks. In 2016 International Conference on Computing, Analytics and Security Trends (CAST) (pp. 261–266). 2016 International Conference on Computing, Analytics and Security Trends (CAST). IEEE. https://doi.org/10.1109/cast.2016.7914977

[15] Zago, G. T., Andreão, R. V., Dorizzi, B., & Teatini Salles, E. O. (2020). Diabetic retinopathy detection using red lesion localization and convolutional neural networks. In Computers in Biology and Medicine (Vol. 116, p. 103537). Elsevier BV. https://doi.org/10.1016/j.compbiomed.2019.103537

[16] Alyoubi, W. L., Shalash, W. M., & Abulkhair, M. F. (2020). Diabetic retinopathy detection through deep learning techniques: A review. In Informatics in Medicine Unlocked (Vol. 20, p. 100377). Elsevier BV. https://doi.org/10.1016/j.imu.2020.100377

[17] Pradhan, A., Sarma, B., Nath, R. K., Das, A., & Chakraborty, A. (2020). Diabetic Retinopathy Detection on Retinal Fundus Images Using Convolutional Neural Network. In Communications in Computer and Information Science (pp. 254–266). Springer Singapore. https://doi.org/10.1007/978-981-15-6315-7_21

[18] Nguyen, Q. H., Muthuraman, R., Singh, L., Sen, G., Tran, A., Nguyen, B. P., & Chua, M. (2020). Diabetic Retinopathy Detection using Deep Learning. In Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/3380688.3380709

[19] Mishra, S., Hanchate, S. M., & Saquib, Z. (2020). Diabetic Retinopathy Detection using Deep Learning. In 2020 IEEE International Conference on System, Computation, Automation and Networking. https://doi.org/10.1109/ICSTCEE49637.2020.9277506

[20] Gangwar, A., & Ravi, V. (2020). Diabetic Retinopathy Detection Using Transfer Learning and Deep Learning. In Springer Proceedings in Mathematics & Statistics (pp. 799–808). https://doi.org/10.1007/978-981-15-5788-0_64

[21] Tufail, A. B., Ullah, I., Khan, W. U., Asif, M., Ahmad, I., Ma, Y. K., Khan, R., Kalimullah, & Ali, M. S. (2021). Diagnosis of Diabetic Retinopathy through Retinal Fundus Images and 3D Convolutional Neural Networks with Limited Number of Samples. Wireless Communications and Mobile Computing. https://doi.org/10.1155/2021/6013448

[22] Oh, K., Kang, H., Leem, D., Lee, H., Seo, K. Y., & Yoon, S. (2021). Early detection of diabetic retinopathy based on deep learning and ultra-wide-field fundus images. Scientific Reports. https://doi.org/10.1038/s41598-021-81539-3

[23] Dai, L., Wu, L., Li, H., Cai, C., Wu, Q., Kong, H., Liu, R., Wang, X., Hou, X., Liu, Y., Long, X., Wen, Y., Lu, L., Shen, Y., Chen, Y., Shen, D., Yang, X., Zou, H., Sheng, B., & Jia, W. (2021). A deep learning system for detecting diabetic retinopathy across the disease spectrum. Nature Communications. https://doi.org/10.1038/s41467-021-23458-5

[24] Firke, S. N., & Jain, R. B. (2021). Convolutional Neural Network for Diabetic Retinopathy Detection. In 2021 IEEE International Conference on Advances in Information Systems (pp. 425–432). IEEE. https://doi.org/10.1109/ICAIS50930.2021.9395796

[25] Tassanee Hattiya. (2021). Diabetic Retinopathy Detection Using Convolutional Neural Network: A Comparative Study on Different Architectures. Mahasarakham International Journal of Engineering Technology, 7, 50–60. https://doi.org/10.14456/MIJET.2021.8

[26] Das, D., Biswas, S.K. & Bandyopadhyay, S. Detection of Diabetic Retinopathy using Convolutional Neural Networks for Feature Extraction and Classification (DRFEC). Multimed Tools Appl 82, 29943–30001 (2023). https://doi.org/10.1007/s11042-022-14165-4

[27] Ragab, M., AL-Ghamdi, A. S. A.-M., Fakieh, B., Choudhry, H., Mansour, R. F., & Koundal, D. (2022). Prediction of Diabetes through Retinal Images Using Deep Neural Network. In B. Ding (Ed.), Computational Intelligence and Neuroscience (Vol. 2022, pp. 1–6). Hindawi

Limited. https://doi.org/10.1155/2022/7887908

[28] Nandakumar, R., Saranya, P., Ponnusamy, V., Hazra, S., & Gupta, A. (2023). Detection of Diabetic Retinopathy from Retinal Images Using DenseNet Models. In Computer Systems Science and Engineering (Vol. 45, Issue 1, pp. 279–292). Tech Science Press. https://doi.org/10.32604/csse.2023.028703

[29] Butt, M. M., Iskandar, D. N. F. A., Abdelhamid, S. E., Latif, G., & Alghazo, R. (2022). Diabetic Retinopathy Detection from Fundus Images of the Eye Using Hybrid Deep Learning Features. In Diagnostics (Vol. 12, Issue 7, p. 1607). MDPI AG. https://doi.org/10.3390/diagnostics12071607

[30] Kusakunniran, W., Karnjanapreechakorn, S., Choopong, P., Siriapisith, T., Tesavibul, N., Phasukkijwatana, N., Prakhunhungsit, S., & Boonsopon, S. (2022). Detecting and staging diabetic retinopathy in retinal images using multi-branch CNN. In Applied Computing and Informatics. Emerald. https://doi.org/10.1108/aci-06-2022-0150

[31] Bhimavarapu, U., & Battineni, G. (2022). Deep Learning for the Detection and Classification of Diabetic Retinopathy with an Improved Activation Function. In Healthcare (Vol. 11, Issue 1, p. 97). MDPI AG. https://doi.org/10.3390/healthcare11010097

[32] Pradhan, A., Sarma, B., Nath, R. K., Das, A., & Chakraborty, A. (2020). Diabetic Retinopathy Detection on Retinal Fundus Images Using Convolutional Neural Network. In Communications in Computer and Information Science (pp. 254–266). Springer Singapore. https://doi.org/10.1007/978-981-15-6315-7_21

[33] Asia, A.-O., Zhu, C.-Z., Althubiti, S. A., Al-Alimi, D., Xiao, Y.-L., Ouyang, P.-B., & Al-Qaness, M. A. A. (2022). Detection of Diabetic Retinopathy in Retinal Fundus Images Using CNN Classification Models. In Electronics (Vol. 11, Issue 17, p. 2740). MDPI AG. https://doi.org/10.3390/electronics11172740

[34] "Nahiduzzaman, Md., Robiul Islam, Md., Omaer Faruq Goni, Md., Shamim Anower, Md., Ahsan, M., Haider, J., & Kowalski, M. (2023). Diabetic retinopathy identification using parallel convolutional neural network based feature extractor and ELM classifier. In Expert Systems with Applications (Vol. 217, p. 119557). Elsevier BV. https://doi.org/10.1016/j.eswa.2023.119557

[35] Malhi, A., Grewal, R., & Pannu, H. S. (2023). Detection and diabetic retinopathy grading using digital retinal images. In International Journal of Intelligent Robotics and Applications (Vol. 7, Issue 2, pp. 426–458). Springer Science and Business Media LLC. https://doi.org/10.1007/s41315-022-00269-5

[36] Kalyani, G., Janakiramaiah, B., Karuna, A., & Prasad, L. V. N. (2021). Diabetic retinopathy detection and classification using capsule networks. In Complex & Intelligent Systems (Vol. 9, Issue 3, pp. 2651–2664). Springer Science and Business Media LLC. https://doi.org/10.1007/s40747-021-00318-9

[37] Chia, M. A., Hersch, F., Sayres, R., Bavishi, P., Tiwari, R., Keane, P. A., & Turner, A. W. (2023). Validation of a deep learning system for the detection of diabetic retinopathy in Indigenous Australians. In British Journal of Ophthalmology (Vol. 108, Issue 2, pp. 268–273). BMJ. https://doi.org/10.1136/bjo-2022-322237

[38] Jain, A., Gupta, R., & Singhal, J. (2024). Diabetic Retinopathy Detection Using Quantum Transfer Learning (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2405.01734

[39] Data Analysis. (2020, January 17). EyePACS. https://www.eyepacs.com/data-analysis

[40] Prasanna Porwal, S. P. (2018). Indian Diabetic Retinopathy Image Dataset (IDRiD) [Dataset]. IEEE Dataport. https://doi.org/10.21227/H25W98

[41] APTOS-2019 dataset. (2021, November 17). Kaggle. https://www.kaggle.com/datasets/mariaherrerot/aptos2019

[42] Decencière et al.Feedback on a publicly distributed database: The Messidor database. Image Analysis & Stereology, v. 33, n. 3, p. 231-234, aug. 2014. ISSN 1854-5165. Available at: http://www.ias-iss.org/ojs/IAS/article/view/1155

[43] DIARETDB1 - Standard Diabetic Retinopathy Database. (2021, January 24). Kaggle. https://www.kaggle.com/datasets/nguyenhung1903/diaretdb1-standard-diabetic-retinopathy-database

[44] S. M. Pizer, R. E. Johnston, J. P. Ericksen, B. C. Yankaskas and K. E. Muller, "Contrast-limited adaptive histogram equalization: speed and effectiveness," [1990] Proceedings of the First Conference on Visualization in Biomedical Computing, Atlanta, GA, USA, 1990, pp. 337-345, doi: 10.1109/VBC.1990.109340

[45] Castillo Benítez, V. E., Castro Matto, I., Mello Román, J. C., Vázquez Noguera, J. L., García-Torres, M., Ayala, J., Pinto-Roa, D. P., Gardel-Sotomayor, P. E., Facon, J., & Grillo, S. A. (2021). Dataset from fundus images for the study of diabetic retinopathy. In Data in Brief (Vol. 36, p. 107068). Elsevier BV. https://doi.org/10.1016/j.dib.2021.107068

[46] Diabetic Retinopathy 224x224 Gaussian Filtered. (2020, February 18). Kaggle. https://www.kaggle.com/datasets/sovitrath/diabetic-retinopathy-224x224-gaussian-filtered

[47] Shorten, C., Khoshgoftaar, T.M. A survey on Image Data Augmentation for Deep Learning. J Big Data 6, 60 (2019). https://doi.org/10.1186/s40537-019-0197-0

[48] Linkon, A. H. Md., Labib, Md. M., Hasan, T., Hossain, M., & Jannat, M.-E.-. (2021). Deep learning in prostate cancer diagnosis and Gleason grading in histopathology images: An extensive study. In Informatics in Medicine Unlocked (Vol. 24, p. 100582). Elsevier BV. https://doi.org/10.1016/j.imu.2021.100582

[49] Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2017). Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 31, Issue 1). Association for the Advancement of Artificial Intelligence (AAAI). https://doi.org/10.1609/aaai.v31i1.11231

[50] Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition (Version 6). arXiv. https://doi.org/10.48550/ARXIV.1409.1556

[51] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. arXiv. https://doi.org/10.48550/ARXIV.1801.04381

[52] Empirical Analysis of a Fine-Tuned Deep Convolutional Model in Classifying and Detecting Malaria Parasites from Blood Smears. (2021). In KSII Transactions on Internet and Information Systems (Vol. 15, Issue 1). Korean Society for Internet Information (KSII). https://doi.org/10.3837/tiis.2021.01.009

[53] Nimmisha Shajihan. (2020). Classification of stages of Diabetic Retinopathy using Deep Learning. Unpublished. https://doi.org/10.13140/RG.2.2.10503.62883

[54] Kumar, R., & Indrayan, A. (2011). Receiver operating characteristic (ROC) curve for medical researchers. In Indian Pediatrics (Vol. 48, Issue 4, pp. 277–287). Springer Science and Business Media LLC. https://doi.org/10.1007/s13312-011-0055-4