An Efficient Deep Learning Approach to Detect Diabetic Retinopathy : Analysis and Severity Prediction

by

MD. TAMZID HOSSAIN
19101121
UTSAV BHOWMIK
19101646
RIZA ASMAT MILA
20101590
MAHTAB SHAHRIAR CHOWDHURY
19101637
RONAK KARMAKAR
18101298

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
January 2024

Declaration

We declare that

1. The thesis we have submitted is our own original work, accomplished for the Brac University degree.
2. Except in cases where it is properly cited with complete and precise references, this thesis does not contain any previously published or authored content by a third party.
3. It doesn't include any content that has been approved or submitted for credit toward another degree or certification at any university or other establishment.
4. We have given credit to all significant sources of support.

Student's Full Name & Signature:

Md. Tamzid Hossain

MD. TAMZID HOSSAIN
19101121

Utsav

UTSAV BHOWMIK
19101646

Riza Asmat

RIZA ASMAT MILA
20101590

Mahtab

MAHTAB SHAHRIAR CHOWDHURY
19101637

Ronak Karmakar

RONAK KARMAKAR
18101298

Approval

The thesis/project titled ”An Efficient Deep Learning Approach to Detect Diabetic Retinopathy : Analysis and Severity Prediction.” submitted by

1. MD. TAMZID HOSSAIN (19101121)
2. UTSAV BHOWMIK (19101646)
3. RIZA ASMAT MILA (20101590)
4. MAHTAB SHAHRIAR CHOWDHURY (19101637)
5. RONAK KARMAKAR (18101298)

Examining Committee:

Supervisor:
(Member)

Dr. Md. Ashraful Alam
Assistant Professor
Department of Computer Science & Engineering
Brac University

Program Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam
Professor
Department of Computer Science & Engineering
Brac University

Head of Department:
(Chair)

DR. Sadia Hamid Kazi
Professor
Department of Computer Science & Engineering
Brac University

Abstract

Diabetic retinopathy is one complicated eye complication of diabetes and considered one of the major causes of preventable blindness worldwide. Diabetic retinopathy occurs when high glucose levels in the blood damage small blood vessels of the retina over time continuously, resulting in various problems with vision. In its early stages, DR typically shows no symptoms; thus, early detection is very important in order to avoid permanent loss of vision. Given the importance of early diagnosis, advanced machine learning systems, especially those applying deep learning, have been very important in eye care in recent times. This work presents a new deep learning model using ensemble learning combined with a hybrid architecture and proposes a deep learning model named DRDetector. The proposed DRDetector combines ResNet50 for feature extraction with Vision Transformer ViT layers to understand the global context. This methodology overcomes the challenge of diagnosis and prediction of diabetic retinopathy with enhanced accuracy while minimizing false positive and negative cases. DRDetector uses a Convolutional Neural Network (CNN) combined with a Vision Transformer architecture, with transfer learning for detection of DR stages. It classifies the retinal images into different classes including healthy, and different stages of DR: mild, moderate, NPDR, and PDR. The aim of this paper is to comprehensively assess the performance of DRDetector based on a large dataset of retinal images, so that its efficacy can be shown in clinics. This would lead to improved diagnosis with higher accuracy, reduction of diagnostic errors, and in effect, help the ophthalmologists39; quest for perfection. Moreover, an advanced grading system can assist healthcare practitioners in grading the severity of the disease for better management and treatment options for DR. This study has pointed out that optimized deep learning systems may support early detection, risk evaluation, and personalized treatment for diabetic retinopathy patients.

Keywords: Diabetic retinopathy; Symtomps; DRDetector; non-proliferative; Deep Learning.

Dedication

This paper is dedicated to our beloved parents for whom we exist.

Acknowledgement

We are deeply grateful to the Almighty for granting us the opportunity to pursue our bachelor39;s degree at BRAC University and for ensuring the successful completion of our thesis without significant challenges.

The journey towards earning our degree, and specifically completing this thesis, would not have been possible without the support of many individuals. We are immensely thankful to everyone who offered their guidance and encouragement along the way.

Our heartfelt appreciation goes to our Supervisor, Dr. Md. Ashraful Alam, PhD, whose invaluable support and assistance were essential throughout this project. His constant motivation and guidance were instrumental in completing this thesis. We feel fortunate to have worked under such an inspiring and approachable mentor.

We would also like to extend our thanks to the Department of Computer Science and Engineering at BRAC University and the CVIS research lab for providing us with essential resources and assistance. Lastly, we owe immense gratitude to our

family members—our parents, brothers, and sisters who are the foundation of our lives. Their unconditional love and support were vital in the completion of this research.

Table of Contents

List of Figures

Chapter 1

Introduction

Diabetic retinopathy is a sophisticated complication of diabetes affecting the fine vessels of the retina, characterized by manifestations such as leakage of blood and fluid. This will cause swelling of the retina and grave impairment of vision. These conditions can be pathologies for both type 1 and type 2 diabetic patients. The prevalence increases with aging. In the most case, if left unmanaged, DR may cause permanent loss of vision, since symptoms often appear when the disease is at an advanced stage. Global health governing bodies, such as the World Health Organization, are concerned with DR as one of the priorities in public health. For instance, it is estimated that DR accounts for about 2.6% of total world blindness, hence the need for effective mechanisms of screening, which is also accessible for early detection to prevent the progress of the disease [16].

Management of diabetic retinopathy becomes especially challenging in underdeveloped countries because of the inadequacy of facilities in the health sector. This is the case in India, which has close to 60 million victims of diabetes. The shortage of ophthalmologists-only two thousand for the entire country-translates into an inability of supply to keep up with demand for eye treatment services. One of the big challenges in DR management is a lack of awareness, given that many people with diabetes are unaware of the risks to their vision. In the initial stages, DR is generally asymptomatic, and diagnosis is possible only in the advanced stage of the disease. It has been estimated that 18% of patients with diabetes have DR, while diabetic individuals are 25 times more at risk of developing this eye condition than non-diabetic people [11]. The serious problem with this belated symptomatology is that it creates major challenges for the health care system to address the issue of DR.

Blurred vision, floaters, and partial vision loss will occur when much of the retinal damage has already been done. These conditions increase the chances of irreversible blindness if not promptly dealt with. This further creates an urgent need to raise efforts for the early detection of DR among people throughout the world, at least in resource settings. Management of DR thus greatly depends on regular screenings of the retina before the damage has spread widely [24]. Classic screening, however, requires specialized equipment and trained medical personnel, which are usually lacking in resource-poor settings. Scaling up the screening programs is also complicated, since the interpretation of the images requires expertise that is usually not available in areas of acute shortages of ophthalmologists.

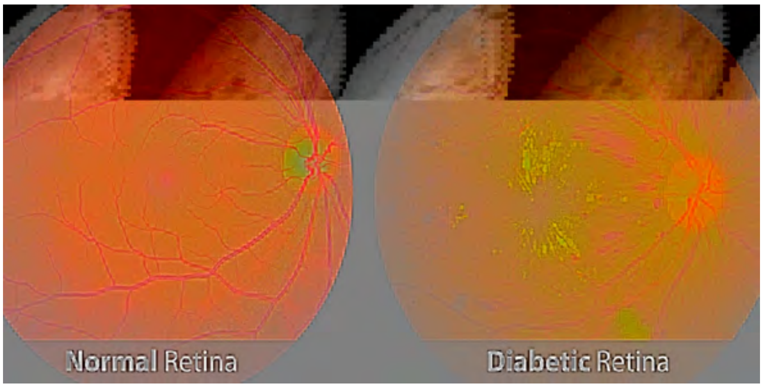

Figure 1.1: Diabetic Retina [8]

More recently, some of these health challenges have found promising solutions with the improvement in AI and deep learning. These algorithms, particularly deep learning algorithms, including CNNs and ViT models, have done a great job of analyzing medical images. They can learn complex patterns from large datasets of retinal images, thus enabling them to perform highly accurate automatic detection of diabetic retinopathy. With the use of a deep learning model for screening, this reduces the workload for health professionals to make certain diagnoses on time.
The research thesis focuses on employing deep learning in building a robust and efficient system for DR early detection and classification. The proposed model, called DRDetector, presents a hybrid model by using ResNet50 as the backbone for feature extraction combined with the ViT layers for modeling the global context. DRDetector aims to enhance the accuracy and reduce false positives and false negatives, increasing in turn the overall reliability of DR detection, by leveraging transfer learning and cutting-edge CNN and ViT architectures [21]. The model will be able to show various stages of DR detection, from mild to severe, including NPDR and PDR.

The ultimate goal of DRDetector is to provide a robust, scalable clinical solution that will aid ophthalmologists in timely diagnosing DR, especially in under-resourced settings.

1.1 Problem Statement

Diabetic retinopathy detection is based on the appearance of specific lesions in the retinal images, which include MAs, HMs, soft exudates, and hard exudates. MAs appear like small red dots; however, hemorrhages have larger and bigger irregular spots. Hard exudates are bright yellow in color due to leakage of plasma, while soft exudates are white spots caused by the swelling of nerve fibers. Each of these lesion types has distinct characteristics, with the major players in red lesions being MAs and HMs, whereas bright lesions represent both soft and hard exudates. DR

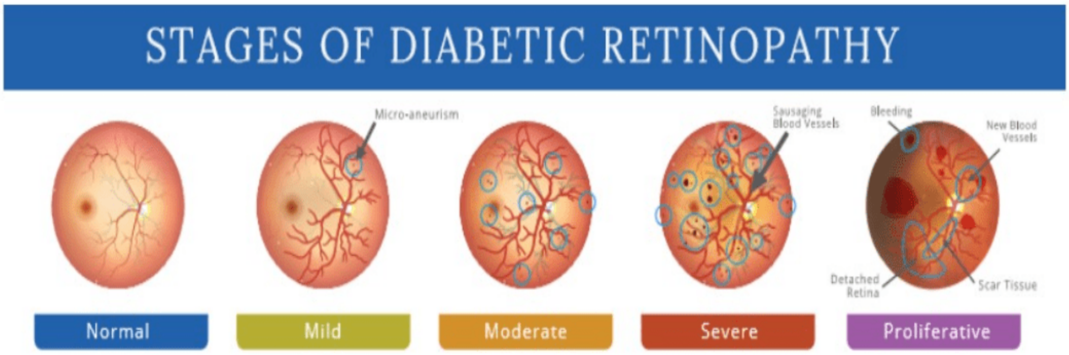

Figure 1.2: Stages of Diabetic Retinopathy [23]

severity is classified into five stages: no DR, mild, moderate, severe nonproliferative, and proliferative DR. The classification provides a basis for timely intervention and treatment.

However, there are some drawbacks with the traditional methods of identification. Many of them involve subjective examination by medical experts, which is both expensive and time- consuming. They tend to be subjective and prone to human error, leading to misdiagnosis and delays [22]. Traditional methods also have the problem of greatly limited availability due to a lack of necessary expertise and equipment, especially in resource-poor settings.

While detection with automation has become much easier, as well as time- and cost-effective, in contrast, much of this recent progress has been enabled by the recent resurgence of deep learning techniques. DL is a class of machine learning methods that are based on complex layers for nonlinear processing, unsupervised feature learning, and pattern recognition. Due to its unprecedented performance in medical image analysis including classification, segmentation, detection, retrieval, and registration, DL has become an active area of interest in DR detection. Techniques such as CNNs and ViTs are especially effective, since they can automatically learn complicated features from big datasets.

This forms the backbone of the CNNs for DL-based detection of DR, which basically comprises convolution layers, pooling layers, and fully connected layers. These networks apply different filters for feature extraction in retinal images, downsampling images with the help of different pooling techniques like average and max pooling. FC layers summarize the general properties of the image, often using SoftMax activation functions for classification.

This process is further accelerated with the pre-trained CNN architectures such as ResNet, VGGNet, Inception-v3, and AlexNet using transfer learning. Then again, one can tune layers individually or train the whole model from scratch.

ViT was developed to apply self-attention in holistic image processing based on the transformer architecture. ViTs divide an image into patches, flatten these patches, and treat them as a sequence of tokens. Given that, the ability of ViTs in modeling long-range dependencies could enhance overall understanding of the images.

Very key elements underpinning their design thus include patch embedding, position embedding, and self-attention mechanisms that provide a way for these models to achieve competitive or even state-of-the-art performance on large-scale image classification tasks.
The family includes pre-trained ViT models: Swin Transformer, DeiT, original ViT, and MedT; their use in a transfer learning framework can significantly improve the performance of the model at reduced training time. These models should be able to generalize their learned representation of global images well for specialized tasks in medical imaging, object detection, or segmentation.
Typically, dataset collection, preprocessing for image enhancement, and feeding into the DL model for feature extraction constitute the general workflow of most works on DL-based DR detection and classification.

1.2 Research Objective

In this study, we propose an efficient deep learning method for identifying and estimating the severity of retinal diseases, with a focus on DR. With the assistance of CNNs and ViT, specifically making a deep learning model with the help of CNN architectures and ViT architectures, we will apply it on the dataset that we formed, which has highest number of meaningful data available in terms of dataset volume,image quality and has enough data for every DR stages.Keeping all the prerequisites in mind we have formed a dataset namely " Diabetic RetinoScope", collecting the best images from the existing datasets like "Diabetic Retinopathy Detection 2015" and "Aptos 2019 Blindness Detection" which are available on Kaggle [8]. The objectives of our research are as follows:

1. Design an efficient deep learning model,,especially DRDetector, which can automatically detect DR in the retinal images.

2. Data collection shall be done, and a dataset should be created with emphasis on the production of a reliable training, validation, and testing dataset.

3. Utilize state-of-the-art image preprocessing techniques for enhancing the quality of retinal images, probably resulting in successful DR detection.

4. Investigate deep learning methods in grading different levels of severity of DR. The stages of the illness are separated out into several.

5. Calculate Sensitivity, Specificity, and Accuracy of deep learning-generated model in DR detection and severity prediction.

6. Compare the effectiveness and precision of the deep learning methodology to conventional DR diagnosis techniques.

7. Analyze the deep learning model's potential for scaling and its usefulness in actual clinical settings.

8. Some details are provided with the knowledge and suggestions of power deep learning methods in integrating them into standard DR screening and diagnosing procedures.

These are the research objectives to guide this study towards the development of an effective and robust deep learning strategy for the diagnosis of diabetic retinopathy, considering its applicability and potential influence on clinical practice.

Chapter 2

Literature Review

2.1 Background Study

2.1.1 Convolutional Neural Network (CNN)

In image processing, a CNN is particularly designed for grid-like data analysis and many other purposes like image classification, recognition, segmentation, and object detection. Convolutional Neural Networks can be one-dimensional, two-dimensional, or three-dimensional. The working of the CNNs took their inspiration from the visual cortex inside an animal, which acts in response to stimuli within certain areas of the visible field. It is designed such that CNNs learn features of interest from input data through various convolution filters or kernels applied to small regions of the output. These filters sum over element-wise multiplication across the entirety of the output, thus allowing precise feature detection within the data [25]. CNN architecture generally includes an Input Layer, a Hidden Layer, and an Output Layer.The major components of a Hidden Layer include the Convolution Layer, the Pooling Layer, the Activation Function, and the Fully Connected Layer. Probably the most important one here is the Convolution Layer, in which a set of filters is applied to small regions of input. Unlike SIFT, features in a CNN are not predefined; they are identified and learned during training. The Convolution Layer is succeeded by the Pooling Layer, which diminishes the size of the input by reducing parameters or weights and, thus accelerating training [9]. There are two types of pooling: Max-Pooling, where the highest value from a feature map is taken, and Average-Pooling, which calculates the mean of values in a pooling window. Non-linear activation functions, such as ReLU, work element-wise on feature maps by replacing the input values that are negative with zero. A Fully Connected Layer is fully connected to all neurons or nodes in one layer to every other neuron in the previous layer: it serves to map learned features into the final output.

2.1.2 Vision Transfomer (ViT)

The Vision Transformer represents any deep learning model that applies the Transformer to process visual information. For the first time developed to solve applications related to NLP, the Transformer architecture has been reused for execution of a wide range of different tasks including image classification, object detection, and image segmentation.In contrast to Convolutional Neural Networks (CNNs), which depend on convolutional operations for feature extraction from images, the Vision Transformer utilizes the self-attention mechanism to identify relationships among various segments of the image without employing convolutional layers [23]. The basic concept behind the Vision Transformer is to take an image as a one-dimensional sequence of small-sized image patches, similar to the words in a sentence, and then feed it into a Transformer model to process the patches and capture very local and global image features [22]. ViT models are designed to analyze two-dimensional grid-like data such as images, where the attention mechanism helps the model focus on important features from all the patches. A Vision Transformer will have three parts: an Input Layer, a Transformer Encoder and an output Layer. Each Transformer Encoder will also include Multi-Head Self-Attention mechanisms, Feed-Forward Neural Networks and Positional Encoding elements.

2.2 Related Works

For the last several years, deep learning has been one of the most studied topics, particularly in Convolutional Neural Networks and their use in medical image processing. Progress has been noted in computer science during recent years, especially within the departments of artificial intelligence and deep learning. All the above innovations have opened new routes for the timely detection of DR and, hence, have created ways for improving patient outcomes, reducing the risk of visual loss.
The paper [13] by Dai et al., which concerns the diagnosis of DR across its spectrum of disease, has proposed a framework known as DeepDR. DeepDR, supported by transfer learning, consists of three deep learning-specific sub-networks, namely image quality assessment, lesion-aware sub-networks, and a DR grading sub-network. The presented deep learning uses the integral framework of ResNet and Mask-RCNN architectures. Regarding general image quality, the proposed DeepDR system reached an overall 0.929 – 0.938. The average AUC from the system's lesion-aware subnetwork in detecting different types of retinal lesions was 0.94. Finally, the average AUC for diabetic retinopathy grading performed by the system was 0.955. It is this unique three-subnetwork architecture that helps DeepDR function in a way that allows the system to automatically grade the quality of a retinal image, detect lesions, and grade diabetic retinopathy. The approach holds much promise of providing better diagnosis of diabetic retinopathy.
In their study [2], Hann et al. present a conventional approach that involves the examination of the morphology and composition of digital photographs in order to derive diverse features from fundus images. This technique has advantages in terms of clarity and simplicity. The identification of exudates in close proximity to the macula in fundus pictures is a pivotal factor in the diagnosis of diabetic macular edema.This paper also focuses on the DR detection in remote areas which lacks in terms of internet speed and other facilities.So,they tried to maintain an acceptable accuracy by reducing the computational expense for that they have used low resource CNN architectures.
In their study, Pratt et al. [6] introduce a novel strategy utilizing deep learning techniques for the purpose of diagnosing and categorizing the severity of diabetic retinopathy (DR) through the analysis of digital fundus images. The authors of this study discuss the difficulties that arise from the manual diagnosis of diabetic retinopathy (DR). They argue that the best way of handling such a challenge is the

employment of a CNN model, which can automatically detect and classify different features of diabetic retinopathy, such as micro-aneurysms, exudates, and hemorrhages. The architecture of CNN used for the classification of diabetic retinopathy is reviewed in great detail in this paper. It explains the different components of CNN, comprising convolution layers, batch normalization, max-poling, dropout, and the activation functions involved in each step of classification. Various other aspects are also brought out by the authors in their discussion: the dataset used, hardware and software configuration, pre-processing steps that were involved-color normalization and scaling-and methodology for training followed, which integrated real-time data augmentation for improvement in performance of the network. The following sections provide the results of their CNN-based DR classification system that achieved a specificity of 95%, an accuracy of 75%, and a sensitivity of 30%. Numerical classification is assigned to different severity categories in the case of DR.
The work of Abràmoff et al. [3] compared the effectiveness of wavelet detectors with that of the k Nearest Neighbors method in the extraction of clinical characteristics from fundus images. The obtained AUC value with the presented extraction method was 0.86 with a Standard Error of 0.0084. It is also interesting to note that the dataset used in the research was obtained by employing ”non-mydriatic” digital cameras of the retina. The size of the fundus photographs ranges from 0.15 to 0.5 gigabytes.

The deep learning system for diagnosing diabetic retinopathy was assessed by Gulshan et al. in their study [5], utilizing a dataset obtained through the use of a smartphone. This study emphasizes the possibility of broadening the availability of diagnostic techniques for diabetic retinopathy to a more extensive and varied population.

In their publication [7], Mateen et al. propose a methodological enhancement that encompasses the integration of multiple methodologies, such as a Gaussian mixture model (GMM), visual geometry group network (VGGNet), singular value decomposition (SVD), principal component analysis (PCA), and softmax. Above-mentioned strategies have been applied to the level of region segmentation, extraction of high-dimensional features, selection of features, and classification of fundus images. It is believed, according to authors, that VGG-19 outperforms both AlexNet and SIFT by performing better in classification while faster in processing.

In another work related to that, Agurto et al. conducted research [4] where they presented an approach which was dependent on the multiscale amplitude-modulation frequency-modulation (AM-FM) techniques to classify fundus images between subjects with and without pathologies. These modulations were applied in various local regions in the fundus images that covered different lesion types. Then, the amplitude-frequency response was analyzed in order to get feature vectors for these locations. According to the authors, one can tell apart normal retinal structures from diseased lesions by the use of the AM-FM features and establishing statistical differences.

The authors of the Khalifa et al. work [10] introduced several computational models in their work. The evaluation of these chosen models was performed on the mentioned dataset APTOS 2019. The chosen models were AlexNet, ResNet18, SqueezeNet, GoogleNet, VGG16, and VGG19. The basis behind this selected model is that they contain lesser layers compared with the other known deep models such as DenseNet and InceptionResNet. Further reinforcement of the models was done, and certain methodologies were applied to restrict their complete absorption of information from the dataset. The examination included computing the testing accuracy and performance metrics to illustrate the robustness of the selected models. The AlexNet model reached the maximum testing accuracy of 97.9%. These performance metrics also further helped in quality improvement. It is important to note that AlexNet has been set up with a limited number of layers, which helps reduce computational complexity and training time

Supriya Mishra and Seema Hanchate used the same data with VGG-16 and DenseNet architectures in paper [12]. They have not applied VGG-16 pre-trained with QWK and ImageNet, but DenseNet121 has been applied with both QWK and ImageNet pretraining. The dataset used in this work consists of 3662 images for training and 1928 images for testing or validation. These are divided into five classes, namely, No DR, Mild DR, Moderate DR, Severe DR, and Prolific DR. Their result indicated that pretraining with ImageNet provided the most significant increase in the accuracy of DenseNet121 with an accuracy rate of 96.1%. Meanwhile, VGG16 resulted in a much lower accuracy of only 73%.

In the paper [1], Walter et al. have proposed a set of algorithms that could effectively extract both exudates and optic discs in the retina. The central idea of their paper is the point of extraction of the exudates by making use of large differences in grayscale level and detecting their outlines using morphological reconstruction techniques.

The authors AbdelMaksoud et al. in the paper [15] proposed a hybrid model, called EDenseNet, merging strengths from both the EyeNet and DenseNet models to precisely identify healthy and DR cases from diverse color fundus images sourced from four different benchmark datasets. In their contribution, an extended evaluation was made on the E-DenseNet model against some well-known architectures like ResNet50, Inception V3, and VGG-19 using the MESSIDOR and IDRiD datasets. The performance metrics considered are computation time (T), dice similarity coefficient (DSC), sensitivity (SEN), specificity (SPE), accuracy (ACC), and Youden's index. Results obtained prove that E-DenseNet performs well, especially when applied to the IDRiD dataset for DR grading, reaching accuracy of up to 93% and sensitivity of up to 96.7%. The work will contribute to efforts in place toward the automation of diabetic retinopathy diagnosis by showing the potential of hybrid models in improving detection and classification of the condition.

The authors Mushtaq et al., in their paper [14], proposed a new architecture of CNN that consists of many deep layers. They used a DenseNet-169 for early diagnosis of the same through their deep learning approach. For this work, two datasets are considered: 'Diabetic Retinopathy Detection 2015' and 'APTOS 2019 blindness detection' from Kaggle. The model propounds a pretty promising outcome, for which the training accuracy reaches up to 95% and the validation accuracy is up to 90%.Directly compared, the proposed model outperforms the regression model, yielding a validation accuracy of 78%. In addition, this study provides a comprehensive review of a wide range of approaches that span deep learning to the use of classic machine learning classifiers: the Support Vector Machine, the Decision Tree, and K-Nearest Neighbour. Among them, the proposed DenseNet-169-based model had a maximum accuracy rate of 90%. Also, it will provide a feasible path for DR in early diagnosis.

Study [18] Sajan et al. refers to the fact that Diabetic Retinopathy is the most common cause of blindness among adults with diabetes, and it needs to be identified at the earliest possible time.It examines the application of the models of machine learning with the purpose of enabling an automated classification system to classify DR into 5 types-80% sensitivity, 82% accuracy, 82% specificity, and 0.904 AUC. First, the strategy followed for the paper involves collection of diversified retinal fundus image with variable conditions, pre-processing to enhance image quality and extract useful features, training on CNN models, which in turn classifies images into various grades concerning the severity of diabetic retinopathy.

In the paper **dihin**, Dihin et. all produced ST by replacing MSA with SW-MSA and leaving rest of the layers intact. They used the Swin-T model combined with multi-wavelet transformation for feature extraction, yielding an accuracy of 97% in the case of the test datasets. They also achieved validation accuracy of 0.9891. The model was more sensitive to detecting No-DR cases, where the sensitivity was 0.97 and specificity was 0.9867. In the case of DR cases, they obtained a sensitivity of 0.9798 and specificity of 0.96.

Another paper [27] by Yang et al. A pre-trained ViT with MAE was used for better improvement in referable DR. MAE pre-trained it with over 100,000 publicly available fundus retinal images with dimensions greater than 224×224. Then, the pre-trained ViT was used for the classification of referable DR and compared the results to ImageNet. The obtained results were an accuracy of 93.42%, AUC value of 0.9853, sensitivity of 0.973, and specificity of 0.9539.

In the paper [1], Fernandes implemented the DeiT model-based diabetic retinopathy detection and also compared the results with the performance of the ResNet18 model. She has shown how learning rates of 1E-04 and 1E-05 can improve the F1 score significantly high as 40% compared to the higher learning rates. In this paper, it was shown that the DeiT model outperformed ResNet18 with a 13% higher F1 score for all the learning rates. This paper presented a relatively better choice of training parameters for DeiT and showed how, with the right mix of parameters, a transformer-based model can outperform several CNN-based models.

Chapter 3

Dataset

3.1 Description of the Dataset

In this study, we introduce a custom dataset called the Diabetic Retinoscope, which is designed for diabetic retinopathy classification tasks. The dataset consists of 10,000 fundus images divided into five classes, each class consisting of 2,000 images. Images were collected from two publicly available Kaggle datasets: APTOS 2019 Blindness Detection dataset (2019) consisting of 3,662 training images; and Diabetic Retinopathy Detection (2015) provided by EyePACS, which consists of 35,126 images. To create the Diabetic Retinoscope dataset, we utilized all images in APTOS 2019 and randomly chose 15,000 images from EyePACS. The reason for choosing most images from APTOS is that this dataset contains better-quality images, and its overall size is smaller than that of EyePACS hence making it computationally more efficient. On the other hand, although more images were collected from the EyePACS dataset, it was less utilized because of its high storage requirement and inherent noise, which included overexposure and shadowing artifacts. The reason behind this was to ensure that only quality images made it to the final dataset. We hence designed a deep learning algorithm using state-of-the-art models: ResNet50 and VGG16. This model was utilized for filtering out the images containing significant visual artifacts. This mainly involved the removal of dark images in which critical features could not be extracted well. We removed all the images for which brightness was below a threshold value of 40 because the important retinal details in such images were hard to view. After doing the quality check, we remained with 8,700 high-quality images. For making the dataset balanced and for the desired total of 10,000 images, perform class-wise data augmentation by cropping, flipping, and other transformations such that the augmented images retain critical features related to diabetic retinopathy classification. We then split the dataset into 7:2:1, where 7000 images allotted for training, 2000 images for validation and 1000 images were allocated for the test part. The dataset size is 3.5 GB, featuring accompanying CSV files specifying labels for each image.

Table 3.1: Severity Levels for Diabetic Retinopathy

ASSIGNED CLASS VALUE	SEVERITY
class_0	No DR
class_1	Mild NPDR
class_2	Moderate NPDR
class_3	Severe NPDR
class_4	Proliferative DR

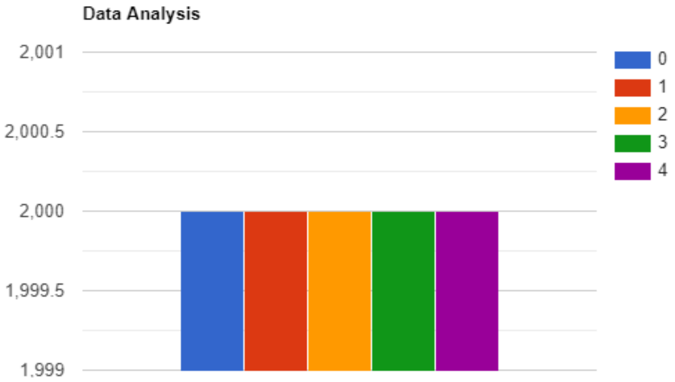

Figure 3.1: Number of images in each class

3.2 Data Sample

Below are images of different diabetic retinopathy eye's conditions. We have five different categories, ranging from healthy (normal) to varying severity levels of DR, including mild, moderate, non-proliferative diabetic retinopathy (NPDR) or severe and proliferative diabetic retinopathy (PDR). It's hard to differentiate the issues with human eyes usually that's why we have used layers so that it is enough good for the model to identify the stages precisely.

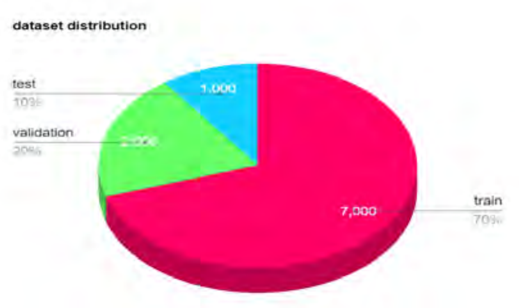

Figure 3.2: Dataset Distribution

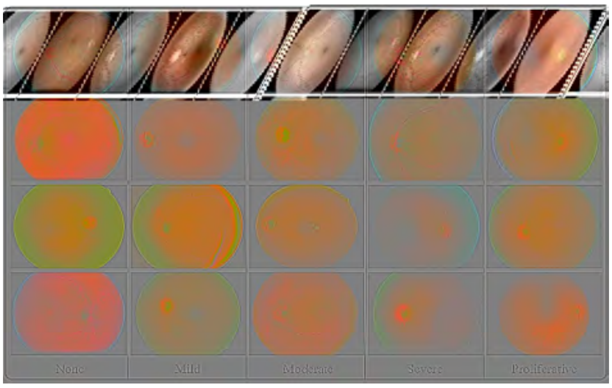

Figure 3.3: DR samples

3.3 Data Pre-Processing

Preprocessing DR fundus image analysis accentuates the mentioned features, especially the blood vessel network and lesions. CLAHE enhances the features of the green channel in fundus images as a preprocessing step, and is particularly useful for enhancing such features. The green channel provides the best contrast for blood vessel extraction when compared to red and blue channels. What is CLAHE? CLAHE is an advanced image enhancement technique that improves the contrast of an image through performing adaptive histogram equalization in small, neighbourhood regions of the image, referred to as tiles. CLAHE differs from the classic histogram equalization in that, during enhancement, it limits amplification at each tile-with the result of not over-saturating high-density pixel intensity areas. The Green Channel: Why choose CLAHE? Blood Vessel Detection : In fundus images, the green channel is usually most obvious in contrast. It increases this contrast of blood vessels by making them much more distinguishable from their background. It, therefore, improves the clarity in detecting it. Lesion Extraction: The manifestations of lesions such as microaneurysms, hemorrhages, and exudates often appear as slight changes in intensity that serve as principal markers for DR. CLAHE enhanced the features in the green channel relating to small, localized variations and improved the detection ability of the model with regard to these abnormalities. In this way, it enhances the input the model gets, and thus, it shall be better at detecting the subtleties that are specifically required for DR.

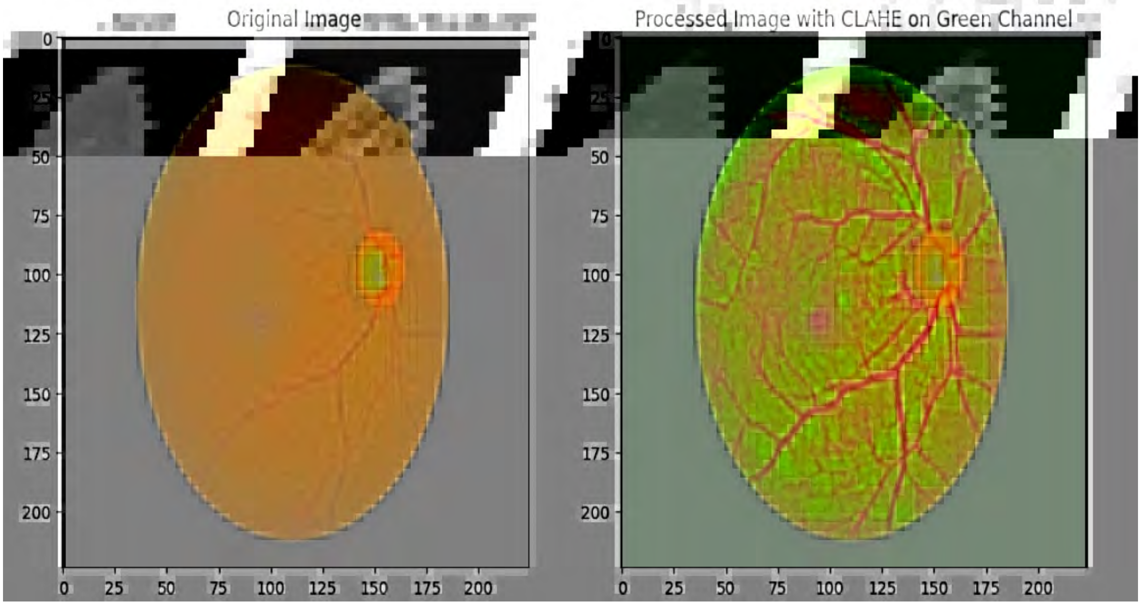

Figure 3.4: Application of CLAHE-01

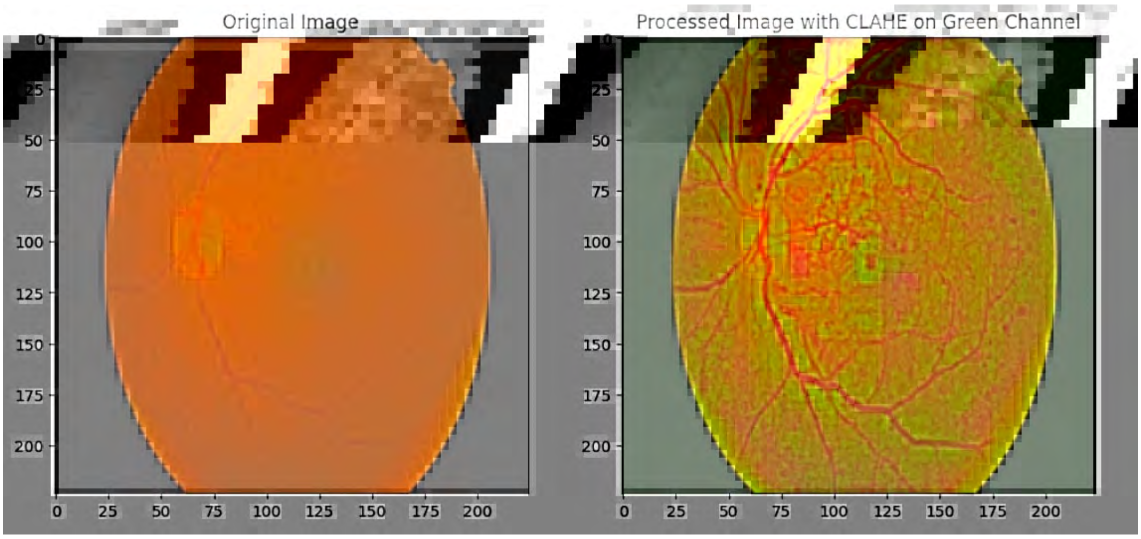

Figure 3.5: Application of CLAHE-02

Chapter 4

Methodology

In order to determine the most efficient approach, it is imperative that we ensure a clear understanding of the proposed problem. Our primary objective is to acquire a comprehensive dataset of diabetic retinopathy, which consists of high quality tomographic retinal images. To optimize our model's performance and achieve the utmost accuracy, meticulous fine-tuning and dataset balancing are essential. Firstly ,after building our dataset, we have applied necessary pre-processing techniques on the data for example resizing all the images to the same size,removing the noisy images,adjust the contrast of the images etc. After that, we have judged the diversity of our data through all the DR stages that we are intended to classify.First of all, the initial data was highly imbalanced. We balanced our data using several augmentation techniques, which reduced the overfitting of the model. Therefore, we will segment the data into three: 70% training, 20% validation, and 10% testing. The key intention of this thesis is to build a deep learning model that can be much better than existing models, yielding more precise and meaningful results based on their observations. The classification of images for diabetic retinopathy will be done in the following groups: mild, moderate, severe non-proliferative diabetic retinopathy (NPDR), and the most advanced form of the disease, proliferative diabetic retinopathy (PDR). Our model will be built on CNN architecture and ViT architecture, since our task involves the processing of images. Dai et al. [13] further performed a multilayer architecture using ResNet; results from each layer were remarkable in terms of AUC scores. Additionally, according to the work by Mateen et al. [7], the VGG-16 model is more effective than AlexNet in terms of classification accuracy and computational efficiency. Based on the performance of these studies, we have done our own model, DRDetector, expecting it to perform better. These will finally be compared with other available empirical results using CNN and ViT architectures. The proposed system performs the task of automatic classification of retinal images, which were normal (healthy) to different severity levels of diabetic retinopathy that included mild and moderate, NPDR, and PDR. We will carry out an in-depth assessment of the DRDetector with a large data set of retinal images, which indeed provides valuable information regarding performance and reliability in a clinical real-world setting. Our method not only improves the diagnostic accuracy but also minimizes false positives and false negatives, becoming an efficient tool for ophthalmologists.

The part below consists of the system architecture of our proposed model as shown in figure 3.1 .

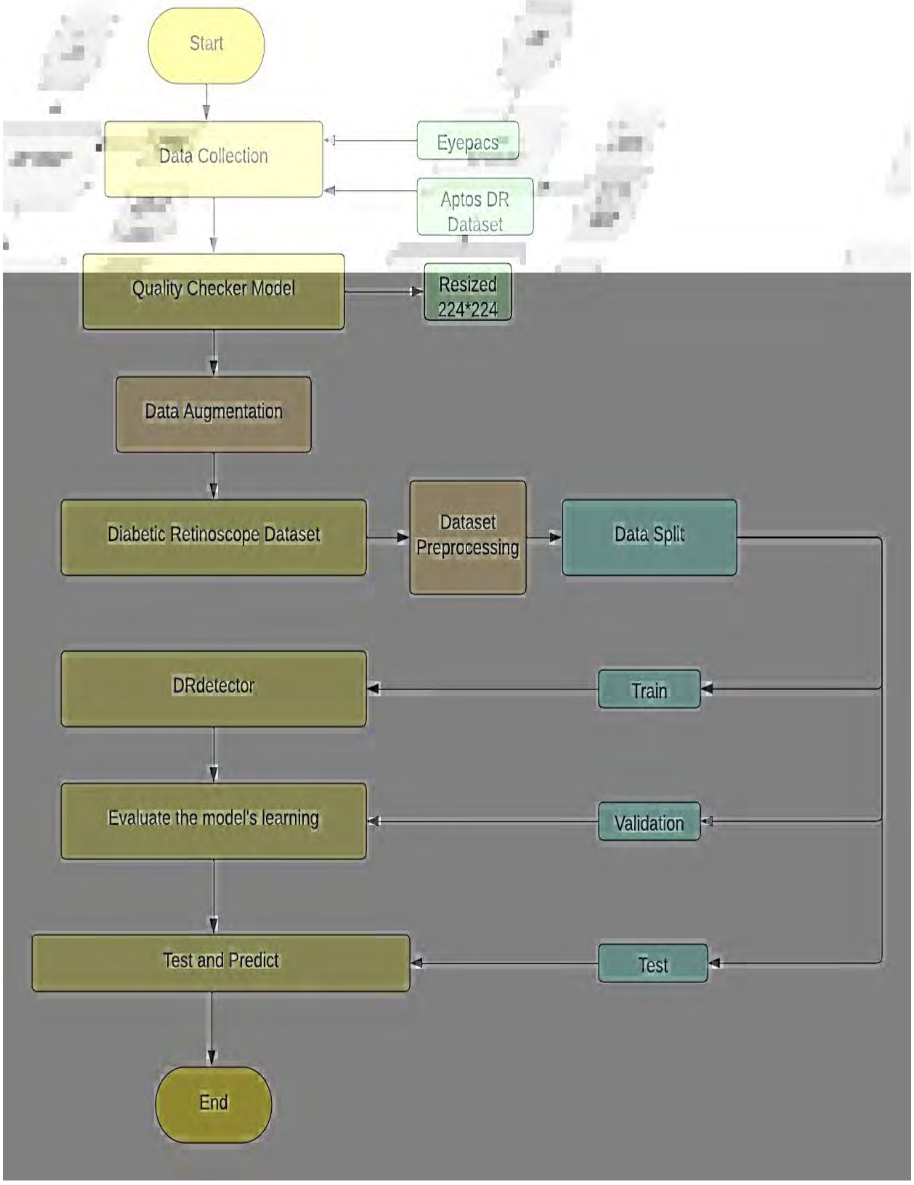

Figure 4.1: Workflow Diagram

4.1 DR Detector Model

4.1.1 Input Layer:

The model accepts input images with a shape of (224, 224, 3), representing 224x224 pixels with three color channels (RGB).

4.1.2 Backbone (ResNet50):

i) Backbone: Based on the architecture of ResNet50-ResNet50. ResNet50 is a deep convolutional neural network proposed to be built with residual connections so that training will be easier and help mitigate problems involving vanishing gradients.

ii) include_top=False in ResNet50 removes the default classification head in the network and keeps only the convolution layers for feature extraction.

iii) The output from the last layer of ResNet50 is retained; this provides a rich set of learned features.

4.1.3 Flattening:

After the ResNet50 backbone, feature maps pass through a Flatten layer to convert the multidimensional feature maps into a one-dimensional vector. This step prepares the features for the Vision Transformer section.

4.1.4 Vision Transformer (ViT) Architecture:

i) Embedding layer: This consists of a dense layer with 128 units in order to reduce the dimensionalities of the flattened feature maps. This way, it's going to help in the transformation of high-dimensional output into a decent size which could be afforded.

ii) The output from the dense layer is reshaped as a sequence of tokens of shape (1, 128). This indicates that the image is treated as a single token, sequence length = 1, and embedding size is 128.

iii) Transformer Encoder Blocks: This is the composition of two transformer encoder blocks. Each of them contains a MultiHead Self-Attention layer with 4 attention heads with key dimension 128. This attention allows the model to learn the relationships from other parts of an image, or in other words, sequences. It helps to capture both local and global dependencies. Residual Connections after the self-attention layer and after the feed-forward layers help in preserving information so that the network learns more efficiently. Layer Normalization after each residual connection is used to stabilize the learning process. Further, this output of the attention mechanism is fed into a Feed Forward Network, which is implemented as a dense layer with 128 units and ReLU activation.

4.1.5 Flattening (Post Transformer):

After the transformer blocks, a flattening layer has been used to arrange the output for the final classification layer.

4.1.6 Classification Head:

Further feature refinement is done by a dense layer of 128 units with ReLU activation. The final output layer is a Dense layer with num_classes as 5, softmax activation function returning a probability distribution across the five categories of diabetic retinopathy.

4.1.7 Compilation:

i) Optimizer: One can use Adam Optimizer because the learning rate will tune itself with the gradients, which could indicate better convergence speed.

ii) Loss Function: It will be the categorical cross-entropy. This is the proper choice for a multi-class problem where classes of output are completely mutually exclusive.

iii) Metric: Accuracy is the metric on which the model's performance will be measured at both training and evaluation.

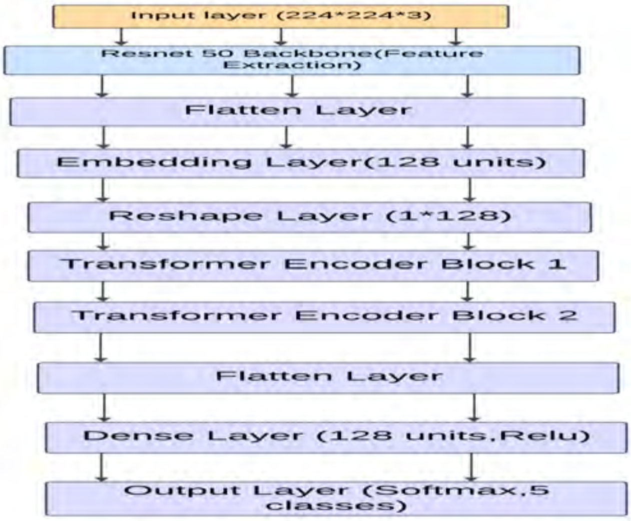
Input layer (224*224*3)
Resnet 50 Backbone(Feature Extraction)
Flatten Layer
Embedding Layer(128 units)
Reshape Layer (1*128)
Transformer Encoder Block 1
Transformer Encoder Block 2
Flatten Layer
Dense Layer (128 units,Relu)
Output Layer (Softmax,5 classes)

Figure 4.2: Architecture of DR Detector

4.2 DR Detector Model Built

The custom model uses the advantages of CNN and transformer-based architectures for classifying the DR fundus images into five severity levels. It starts with a refined version of the ResNet50 backbone, which is from those classes of deep CNN architectures that are known for their powerful feature extraction capabilities. The model is the ResNet50 model, without the top classification layer, and concentrates the model on learning complex visual patterns in the input images, such as networks of blood vessels and lesions of the retina. This output from ResNet50 is flattened to turn those multi-dimensional feature maps into a one-dimensional vector ready for further processing by the Vision Transformer. Then, the embedding layer of the Vision Transformer architecture reduces the flattened features to a 128-dimensional vector that is reshaped into a sequence to be fed into transformer encoder blocks. Each block has a multihead self-attention mechanism with feed-forward networks and residual connections since the model will establish the relations among different regions in the image. This attention mechanism will allow the model to focus on important parts of the image, such as small lesions or abnormal blood vessels, which are key manifestations of DR. These features are further flattened after being processed through the transformer blocks and passed onto a dense layer to reach the softmax classification head, which gives the probability distribution across five DR classes. It was then compiled using the Adam optimizer and categorical cross-entropy loss for the model to classify the input images with high accuracy. This combination of CNNs and transformers provides the most definite methodology of analyzing fundus images and their capability of detecting diabetic retinopathy stages precisely.

4.3 DR Detector Model Application

In this regard, the diabetic retinopathy is detected by locating major retinal features that include lesions, blood vessels, and other abnormalities in fundus images; this model serves with convolutional layers on one hand-on ResNet50-and transformer blocks-Vision Transformer-on the other for the precise feature detection.

4.3.1 Blood Vessel Detection:

The convolutional layers are very powerful in ResNet50 in detecting patterns such as blood vessels by recognizing edges, contours, and textures in the fundus images. It is well known that CNNs extract low-level features like edges in their early layers, which play an important role in the discrimination of blood vessels from the surrounding retinal tissue. Blood vessels represent critical indicators in diabetic retinopathy, as alterations of structure-vessel dilation, microaneurysms-point to the progression of the disease. ResNet50 extracts detailed representations of the network of blood vessels so the model can detect certain anomalies, including vessel leakage or occlusion.

4.3.2 Lesion Detection:

The small lesions that develop, such as microaneurysms, hemorrhages, and exudates, are the important signs reflective of the severity of diabetic retinopathy. Most of these often appear as inconspicuous findings of small dark or bright spots or deposits on the retina. In ResNet50, deeper layers learn complex feature hierarchies that focus on high-level features of the lesion with respect to texture and shape. It is in the fine-grained details of such structures that the model performs the classification to demarcate normal and abnormal regions of the retina.

4.3.3 Vision Transformer (ViT) for Feature Relationships:

It gives a boost to the capability of capturing globally occurring relationships between different features of an image once Vision Transformer is introduced. While ResNet50 captures very local relationships between features, the transformer blocks introduce a layer of understanding spatial dependencies that relate specific features of interest, for example, lesions and blood vessels. For instance, transformers can aid in capturing the model to understand how microaneurysms are related with the general health of the retina by modeling long-range dependencies. This is especially important when detecting more complex features of DR such as cotton wool spots or neovascularization-features that involve several regions of the retina. Its multi-head attention mechanism in the transformer block provides the model with the ability to dwell deeper into relevant parts of the image, especially in areas of cooccurrence between lesions and vessel abnormalities, amplifying the signals that provide evidence of disease.

4.3.4 Classification into Diabetic Retinopathy Stages:

Once this model has extracted these crucial features related to vessel abnormalities and lesions, among others, the information goes through the transformer blocks to refine this feature map, focusing on important patterns across the image. The dense layers and softmax classification head output from this model finally predict a probability distribution across five classes representing different levels of severity in diabetic retinopathy. These might range from no apparent DR-class 0 to proliferative DR-class 4 depending on the lesion type, extent, and vessel damage.

4.3.5 Why the Model is Effective:

It leverages the strengths of a CNN in the detection of local features with those of transformers in the capture of global relationships, hence an overall understanding of the retinal image. These two approaches ensure that the system will be able to detect subtle changes in blood vessels and lesions-those features considered critical in diagnosing and grading diabetic retinopathy.

Chapter 5

Pre-Trained Model

5.1 ResNet-50

ResNet 50 is very powerful in image classification, proposed by Microsoft Research. This is built upon the ”Residual Network” architecture. It is some kind of deep convolutional neural network designed in working with image classification and having a depth of 50 layers. This architecture works well in recognizing images through residual connections that help in mapping input to the desired output. This would enable the model to learn the connections among them better and thus probably give better performance than in earlier stages.Architecture-wise, ResNet 50 starts with convolutional layers in order to derive features from the input image. Generally speaking, there are two types of blocks where the processing and transformation of inputs take place: the identity block and the convolutional block. Lastly, the final classification in ResNet-50 is done by the fully connected layers. In other words, this was the overall description of the architecture of ResNet 50.The significant advantage of this model is that it resolves the problem of a vanishing gradient in a highly innovative way. Hence, it can handle the gradient even when the number of layers reaches up to a thousand or so, out of reach for many models. ResNet 50 is preferred due to the simplification of the training process as a result of residual mapping, which can bear heavy complexities for mapping and make the process of deep learning simpler [17]. Most of the core components remain similar in extensions of ResNet across different configurations. Adding residual connections after a number of blocks that includes convolutional and max pooling layers combined, we are going to focus on constructing ResNet 50 using images that are each 224x224 pixels. These optimized variants include tuned hyperparameters for our needs of learning rate and dropout rates, among others, for further improvements in optimization and avoidance of overfitting.Then, for all the variants of ResNet we experimented with, we set the search space for learning rates to be from 0.0001 to 0.1, and the dropout rate ranged from 0.0 to 0.9. We iterated 10 times, randomly selecting the learning and dropout rates, running each configuration for 3 epochs to get the best settings based on testing accuracy. The key difference between various ResNet models is the size of the input images and the number of convolutional layers, which will be elaborated on in the implementation section.

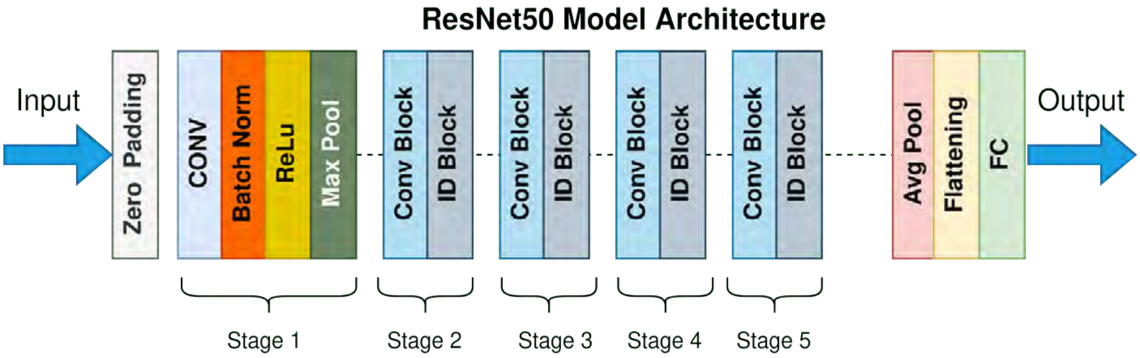

Figure 5.1: Architecture of Resnet-50 [9]

5.2 VGG 16

VGG16 is one of the most widely used and user-friendly models for image classification. It consists of sixteen layers, each containing learnable parameters, also referred to as weights. This model can accurately classify images into one of 1000 predefined categories. The overall architecture includes 21 layers in total, of which 13 are convolutional layers, 5 are Max Pooling layers, and 3 are Dense layers.. From these, only 16 are weight layers, that is, layers with learnable parameters. One of the salient features of VGG16 is its simplicity in design instead of relying on a large number of hyperparameters; it relies on stacks of convolutional filters of size 3x3 with stride 1, while padding is always the same. Max Pooling layers use 2x2 filters with a stride of 2. It is in this consistent pattern of convolution and pooling layers that the architecture for VGG16 is defined.

While VGG19, being more complex and having 19 layers, yields better accuracy with reduced loss, even VGG16 is computationally expensive- demanding much from the processor. However, correctly implemented, the VGG16 gives an accuracy of more than 90%. VGGNet, in all its variants, takes in images whose input dimensions are 224x224. For instance, its convolutional layers have very small receptive fields, for example, 3x3 kernels with the stride of 1 and padding of 1; thus, capturing very directed features. It also uses 1x1 con-volutions to linearly transform the input. VGG uses ReLU activation functions just like AlexNet, which provide positive outputs for any positive input but zero out for any non-positive input. VGG does not include LRN since its memory usage is too high.

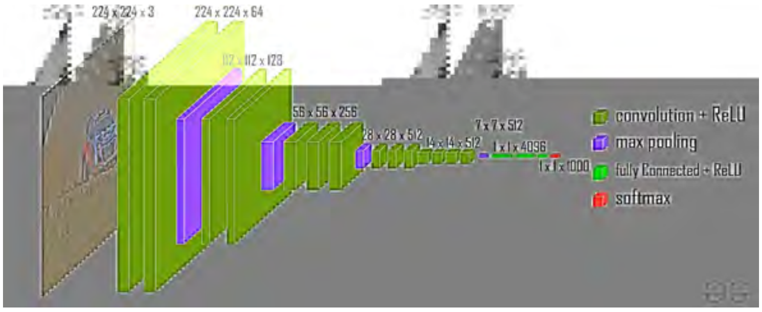

Figure 5.2: Architecture of VGG 16 [9]

Max Pooling follows each convolutional block for reducing the spatial dimensions of feature maps without losing key features. Fully connected layers follow after the final convolution and pooling to help in making predictions. The last fully connected layer is connected to the output with neurons representing target classes, using the SoftMax activation function for classification. Training a VGG16 from scratch can thus become very resource-intensive, since successive convolutional layers lead to computationally intensive procedures. Transfer learning by means of pre-trained VGG16 can allow a more efficient approach where changes are necessary only at the fully connected layers in order to adapt to certain classification tasks.

5.3 InceptionV3

InceptionV3 is a very deep (up to 42 layers) convolutional neural network architecture that was introduced in the year of 2015 by Google. It has been widely popular in different areas, including medical image analysis because of its great performance for solving the problem of Image Classification. This model is very useful in identifying the different stages of diabetic retinopathy, thanks to its capability to extract highly sophisticated features from images of retinaueling Fast data processing power InceptionV3 is actually the third version of Inception series and includes lots of improvement compared to its predecessors. Inception was highly influenced by the success of previous architectures like AlexNet and VGG in 2012 ImageNet competition, hence wanting to achieve both great accuracy but also effective use of computational resources.Indeed the key characteristic that drastically separates Inception from its predecessors is a more extensive utilization for a diversified scale including multi-scales operation enabled with inception modules. It does this by using 1x1,3x3 and 5 x 5 convolutions at parallel in Conv layer then the combined output is used to train the next layers. The architecture also innovates by using factorized convolutions and asymmetric convolutions, which lower computational cost without losing accuracy. As such, rather than utilizing large convolution filters (which—let's say—a 5x5 will have too many parameters), InceptionV3 factorizes them and then parametrize the resulting smaller operations. For instance two 3x3 convolutions instead of a single one with bigger size. It does not only increase in efficiency, but also feature extraction.

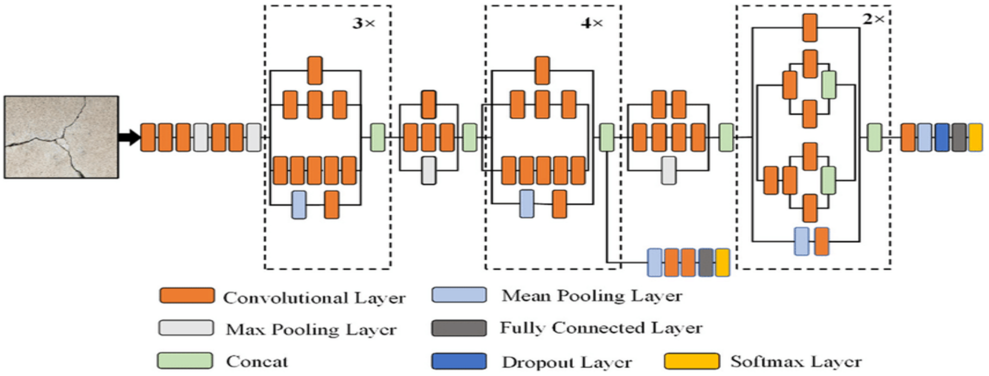

Figure 5.3: Architecture of InceptionV3 [9]

Furthermore, InceptionV3 uses batch normalization extensively to stabilize and accelerate training by normalizing intermediate layers, leading to faster convergence and better generalization. Another notable feature is global average pooling, which replaces fully connected layers with a pooling mechanism that significantly reduces the number of parameters while retaining essential information. This contributes to the model's efficiency, making it ideal for large-scale image analysis, including medical imaging tasks.

5.4 Xception

The Xception architecture is a sophisticated convolutional neural network developed by researchers at Google, utilizing depth-wise separable convolutions. It bases its architecture on the reinterpretation of Inception modules in convolutional networks to serve as a bridge between the depth-wise separable convolutions and the usual convolutions. A depthwise separable convolution can be seen as an Inception module with the maximum number of towers. It comprises a depthwise convolution followed by a pointwise convolution. This led to the creation of Xception, which builds on Inception by replacing its modules with depthwise separable convolutions for improved computational efficiency. Instead of traditional convolutional layers, the architecture employs depthwise separable convolutions, where the pointwise convolution—a 1×1 convolution—merges the output channels to enhance interactions, while the depthwise convolution decreases the overall computations. To facilitate better gradient flow and convergence during training, Xception incorporates residual connections around several layers. Due to these connections, the model will be able to learn more powerful representations since it reduces the risk of vanishing gradients for deeper layers. Besides, the residual connection of this model helps to remove the problems about disappearing gradients in deeper layers. Diabetic retinopathy is a kind of detection process, finding particular lesions and abnormalities such as microaneurysms or hemorrhages from the global and local features in the retinal fundus images for better accuracy [9]. The Xception model extracts both tiny details and global patterns from the images using depth wise separable convolutions, reducing the computational load for swifter training and inference. This algorithm will classify DR phases based on feature correlations and will learn how to identify DR-related features such as textures and forms based on annotated fundus images. Xception is an ideal architecture for medical imaging owing to its excellent precision and effectiveness in DR stage detection. While Xception models are still quite expensive to train, they are a significant improvement on Inception. A part of the solution to adapt such algorithms for your specific purpose is transfer learning. Instead of using parallel convolutions of different sizes, as in Inception modules. This was replaced, in Xception, with depth wise separable convolutions where the convolution is split into two parts: Depthwise convolution: Only one convolution filter is applied on each channel. Pointwise convolution: It uses 1x1 convolutions in order to combine the output of the depthwise convolution. It achieves very good performance with drastically reduced total number of parameters and computational cost. For tasks like classification into different stages of diabetic retinopathy, this separation of the convolutions allows Xception to capture much more intricate spatial details and better hierarchical information. The architecture of the Xception model is made up of 36 convolutional layers that are structured in a modular fashion. Each module in Xception ends with a residual connection similar to ResNet. These residual connections allow the network to be much deeper, as they avoid the problem of vanishing gradients.

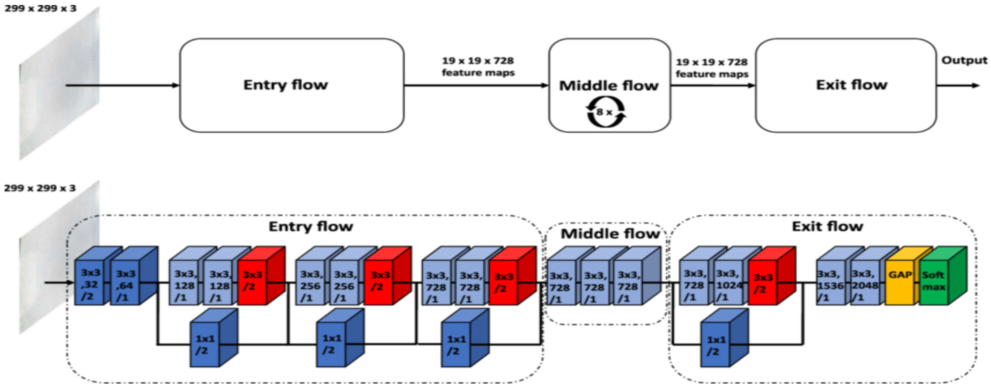

Figure 5.4: Architecture of Xception [9]

This approach significantly reduces the number of parameters and computational cost, while still maintaining high performance. By separating these convolutions, Xception allows the network to capture more complex spatial features and learn better hierarchical representations, which are crucial for tasks like identifying different stages of diabetic retinopathy. Xception's architecture consists of 36 convolutional layers structured into modules, with each module ending in a residual connection, similar to ResNet. These residual connections help combat the vanishing gradient problem and allow the network to go deeper without losing performance.

5.5 ViT

Google researchers introduced the Vision Transformer (ViT) as a novel approach for image recognition, based on the Transformer architecture commonly used in natural language processing. ViT eliminates convolutional layers in aid of a self-attention mechanism that treats image patches as sequences of tokens, similar to words in a sentence, this is known as patch embedding. Since transformers do not have built-in awareness of where each patch sits in the overall image, the ViT first adds positional encoding, which keeps track of each patch's original location. This helps a model understand the spatial relationships within an image-for instance, where certain features such as vessels or lesions appear in a retinal scan.In these layers, each patch can interact with all others using a self-attention mechanism, which allows the model to "see" the entire image context at once. The self-attention mechanism will help ViT determine such complex patterns by learning how different regions of an image are related to one another, which is highly useful when analyzing retina images for diabetic retinopathy. Another great thing in ViT is the special token it has, [CLS], which serves like a summary of all patch information so that the model may summarize the whole image for the classification task, say, classifying the level of diabetic retinopathy. The architecture is efficient in handling the image to the extent that both small changes and larger patterns which include the structure of blood vessels can be learnt. It works better in pre-training on large datasets and then fine-tuning with it to grab details of the image specifically [23]. Diabetic retinopathy detection from the images of the retina requires both local and global features detection. It would cause minor changes in the retina, including microaneurysms, hemorrhages, and blood vessels distortions, which are not constrained to only one area. The self-attention mechanism of the ViT model detects long-range dependencies; hence, it will analyze an entire image and decide how different regions interact with one another. Meanwhile, considering diabetic retinopathy with complex patterns, ViT is very effective due to its ability to process detailed local features and global image structure simultaneously.

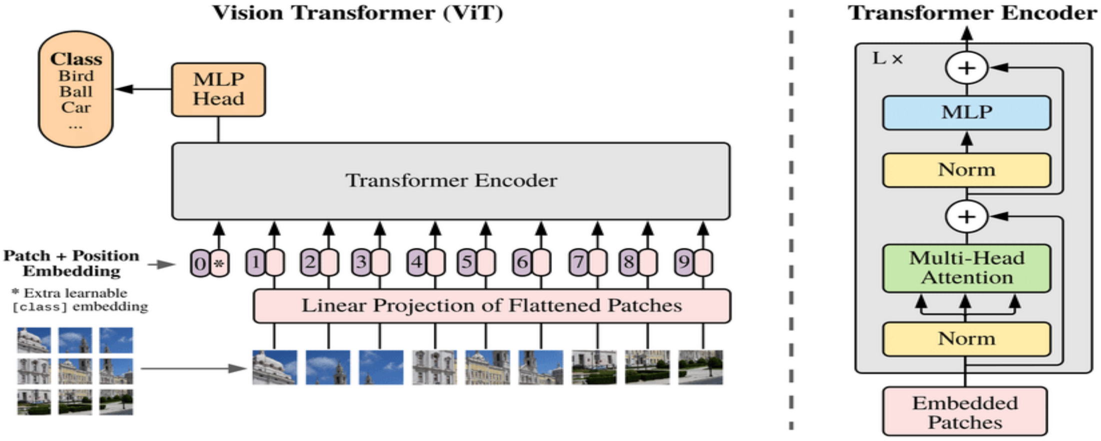

Figure 5.5: Architecture of standard ViT [23]

This model is particularly effective in analyzing retinal images because it can understand both small changes (like lesions) and larger patterns (such as blood vessel structure). When pre-trained on huge datasets and fine-tuned to capture the image's unique characteristics, this architecture performs better. Both local and global features must be found in retinal imaging in order to diagnose diabetic retinopathy. It results in minor, non-localized alterations in the retina, including microaneurysms, hemorrhages, and blood vessel abnormalities. By identifying long-range relationships, ViT's self-attention mechanism is able to examine the entire image and ascertain how various parts interact with one another. All things considered, ViT is especially useful for recognizing the intricate patterns connected to diabetic retinopathy because of its capacity to comprehend both global image structure and fine-grained local information at the same time.

5.6 DeiT

Data-efficient Image Transformer is a vision transformer model designed to perform image classification efficiently, especially when the dataset provided is small. DeiT takes the standard Vision Transformer (ViT) structure and optimizes it for scenarios where large labeled datasets may not be available, which is a common challenge in image classification. The process starts with a patch embedding layer and moving the patches into a high-dimensional space. Then, using positional encoding, it maintains spatial relationships between patches, which ensures that patch location is maintained when we are identifying patterns across the images.The patches with distillation tokens are passed through all the transformation layers. Self-attention is utilized in every layer such that the patched images can interact with each other. This enables DeiT to grasp complex spatial relationships, which may be very informative for those images in which salient features may not reveal themselves distinctly in one specific region alone, as in the case of fundus images for diabetic retinopathy detection. The reason this distillation token is so important in the process is that it is trained from the CNN teacher model. It will, in fact, help DeiT identify the most salient features present in an image and even improve its performance when dealing with smaller sets of data. This setup is especially useful in medical imaging, where labeled data can be scarce and quite costly. For diabetic retinopathy detection, DeiT's architecture can effectively identify finegrained patterns and larger structures in fundus images. Diabetic retinopathy involves a range of abnormalities, from small microaneurysms to larger hemorrhages, that require both detailed local and global context analysis. This is achieved by transformer layers in DeiT by keeping long-range dependencies and hence enabling it to recognize features scattered in an image. The distillation token aids this process of understanding such key DR-specific patterns by DeiT [20]. Further, it will contribute to learning both the exact details, like the shape and color variations of lesions, and the general spatial arrangement of vessels or abnormal spots on the retina. Effectiveness: The second strength with DeiT is. The design of DeiT represents an efficient adaptation of the Vision Transformer for data-constrained environments, hence very suitable for diabetic retinopathy detection. By combining transformer learning of global context with the CNN-guided supervision through a distillation token, DeiT can handle complex retinal images with high accuracy. This makes it appropriate for medical applications, given the need for accuracy in classification and efficiency in data usage, hence allowing for efficient and scalable detection of DR stages in fundus images.

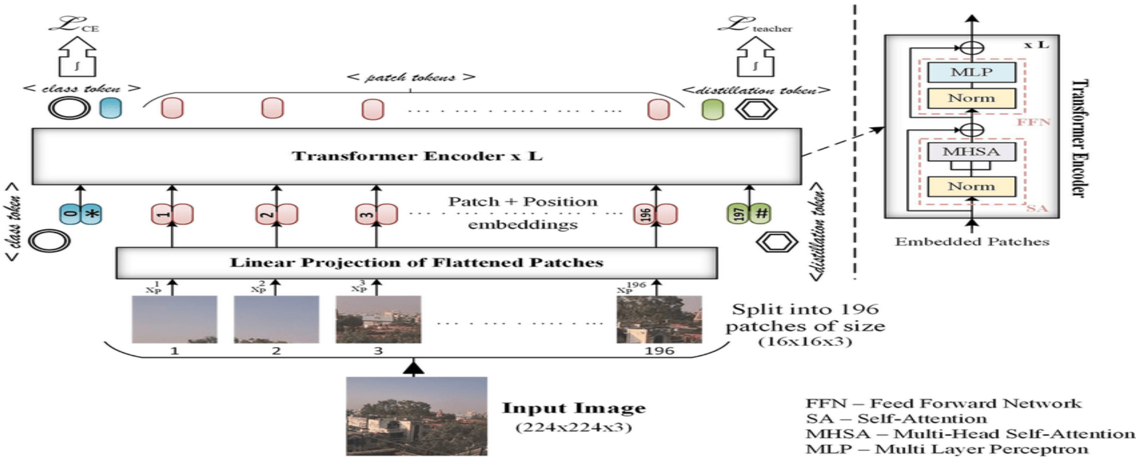

Figure 5.6: Architecture of DeiT [20]

For diabetic retinopathy detection, DeiT's architecture can effectively identify fine-grained patterns and larger structures in fundus images. Diabetic retinopathy involves a range of abnormalities, from small microaneurysms to larger hemorrhages, that require both detailed local and global context analysis. DeiT's transformer layers handle this complexity by preserving long-range dependencies, which allows it to recognize features scattered across an image. The distillation token assists DeiT in understanding these critical DR-specific patterns. It helps in learning both exact details, such as the form and color variations of lesions, and the overall spatial arrangement of vessels or aberrant spots in the retina.Additionally, DeiT provides an advantage in terms of effectiveness. DeiT's design is an efficient adaptation of the Vision Transformer for data-constrained environments, making it an excellent choice for diabetic retinopathy detection. By combining the transformer's global context learning with CNN-guided supervision through the distillation token, DeiT can handle complex retinal images with high accuracy. This makes it suitable for medical applications where precise classification and efficient data use are essential, allowing for effective and scalable detection of DR stages in fundus images.

5.7 Swin

Swin Transformer extends all previous models of visual processing and relies on an effective mechanism for image processing, namely shifted windowing. Previous models of CNN are always based on fixed receptive fields. Unlike these, in Swin Transformer, images are treated by first dividing them into non-overlapping windows and then performing self-attention within those local windows:. The central novelty of Swin resides in its hierarchically representational architecture, the mechanism of shifted windows, allowing both local and global feature extraction without sacrificing computational efficiency as in [19]. This provides the model with the ability to represent fine-grained details and to capture broader contextual information over multiple scales. The self-attention proposed in the Swin Transformer confines itself within several small-sized windowing and reduces the computational complexity drastically compared to global attention mechanisms. In this way, the model still captures long-range dependencies across the whole image by shifting the position of the window in different layers. This design makes the Swin Transformer excel in tasks that require both localized and large-scale pattern understandings; hence, it is well-suited for applications like DR detection. Diabetic retinopathy detection refers to the identification of subtle lesions and abnormalities such as microaneurysms and hemorrhages scattered within the retina. Hence, the Swin Transformer can model not only the fine details within a small window but also the greater structure of the retina for more precise feature detection. The model learns the DR-related patterns at several levels in the retina, such as changes in the structure of blood vessels, through a hierarchical structure [26]. It also involves layer normalization and residual connections that improve the flow of gradients, hence allowing for better convergence during training. This prevents a number of issues such as vanishing gradients, therefore enabling the model to learn even on deep architectures. Besides, its efficiency opens ways to large-scale medical image analysis, enabling faster training and high-accuracy inference. Applications such as DR detection give Swin Transformer a good edge since Swin deals with highresolution images that encompass feature extraction in both the local and global sense. This provides performance increments through training on big datasets and subsequent fine-tuning regarding specific medical tasks such as the analysis of retinal images. Hence, making the Swin Transformer a strong tool in medical image areas for improved accuracy in detection and enhancing computational efficiency with respect to the traditional convolutional models.

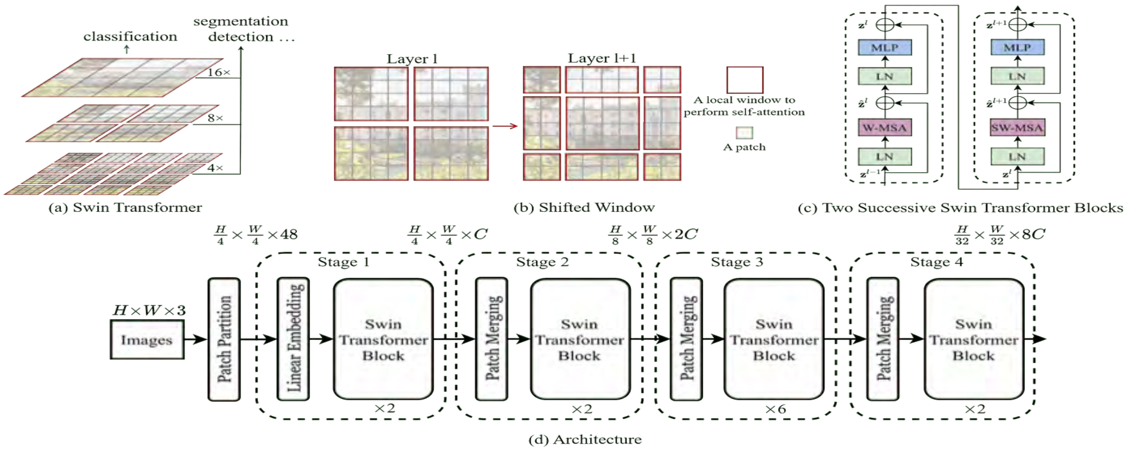

Figure 5.7: Architecture of Swin [19]

The model learns the DR-related patterns at several levels in the retina, such as changes in the structure of blood vessels, through a hierarchical structure. It also involves layer normalization and residual connections that improve the flow of gradients, hence allowing for better convergence during training. This prevents a number of issues such as vanishing gradients, therefore enabling the model to learn even on deep architectures. Besides, its efficiency opens ways to large-scale medical image analysis, enabling faster training and high-accuracy inference. Applications such as DR detection give Swin Transformer a good edge since Swin deals with high-resolution images that encompass feature extraction in both the local and global sense. This provides performance increments through training on big datasets and subsequent fine-tuning regarding specific medical tasks such as the analysis of retinal images. Hence, making the Swin Transformer a strong tool in medical image areas for improved accuracy in detection and enhancing computational efficiency with respect to the traditional convolutional models.

Chapter 6

Model Implementation

6.1 Performance of Pre-Trained Models

6.1.1 ResNet-50

Prior to placing the input images into the model, we scaled them to 224 x 224. The output layer was adjusted to fit five classes in order to determine the degree of diabetic retinopathy severity. Both Xception and Inception-v3 provide completely connected layers. It was then adjusted using our unique dataset. We obtained the validation accuracy of 90.83% and a training accuracy of 96.30% using the Resnet-50 model. The training and validation figures are shown in the picture below. The figure suggests that training accuracy increases with time.

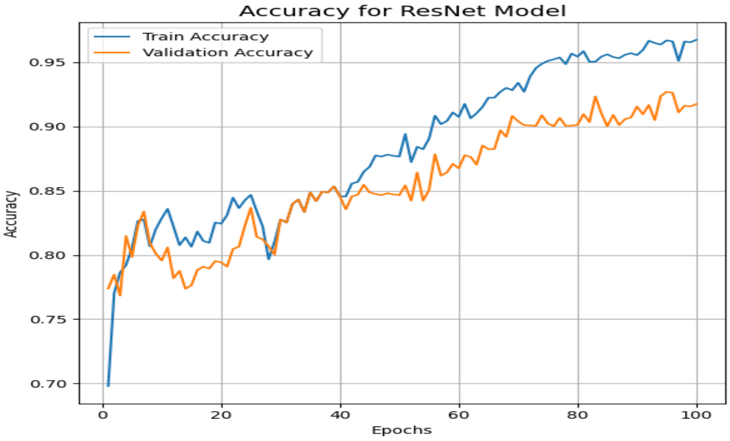

Figure 6.1: Graph of Resnet-50

6.1.2 VGG 16

We achieved the validation accuracy of 88.33% and the training accuracy of 94.72% using the VGG 16 model on our dataset. The training and validation graphs are shown in the figure below. The figure suggests that training accuracy increases with time.

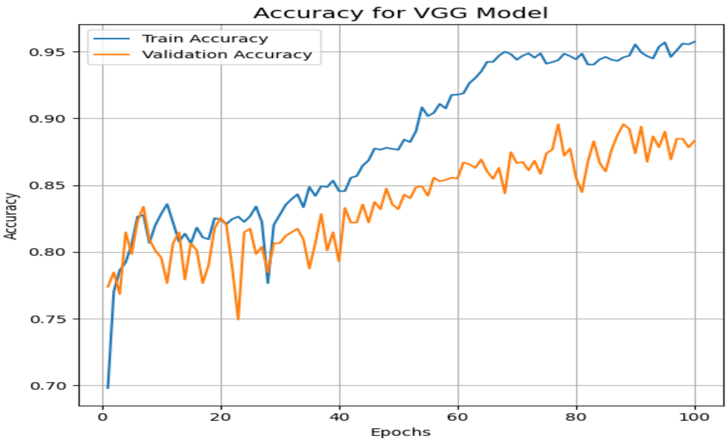

Figure 6.2: Graph of VGG 16

6.1.3 InceptionV3

Before feeding the input images into the model, we reduced them to 224 by 224, just like the Xception model. The output layer was modified to fit five classes in order to determine the degree of diabetic retinopathy severity. Both Xception and Inception-v3 provide fully connected layers. Then, just like we did for Resnet-50, it was adjusted on our own dataset. In the end, we achieved 95% train accuracy, 92% validation accuracy, 98% train auc, and 95% val auc. The figure suggests that training accuracy improve with time.

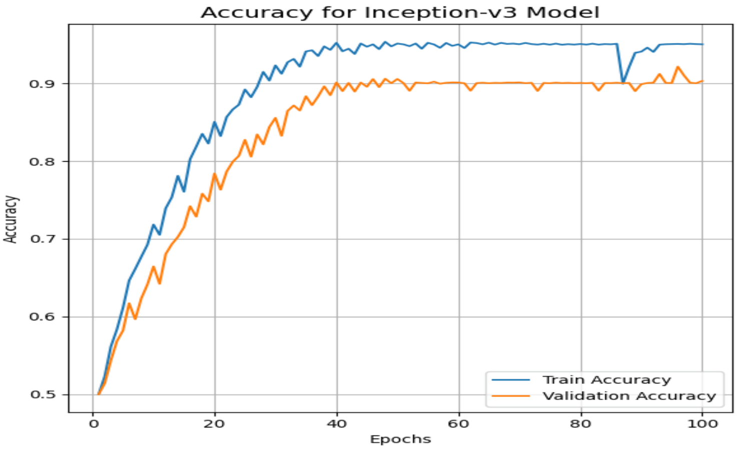

Figure 6.3: Graph of InceptionV3

6.1.4 Xception

We've used pretrained Xception model and then fine-tuned it on our custom dataset. For fine- tuning firstly we freezed the early layers and then later we unfreeze the last few layers gradually. As for learning rate, we start with a lower learning rate which is 0.0001 but we used a learning rate scheduler and optimizer to adjust the learning rate while training. We resized the image to 224×224 before putting them into the model. Besides, we did data augmentation for enhancing generalization. Finally, we got 96% train accuracy and 92.33% validation accuracy. As for the auc value we got 98% for train data and 96% for validation data. We can observe from the figure that the training accuracy improves over time.

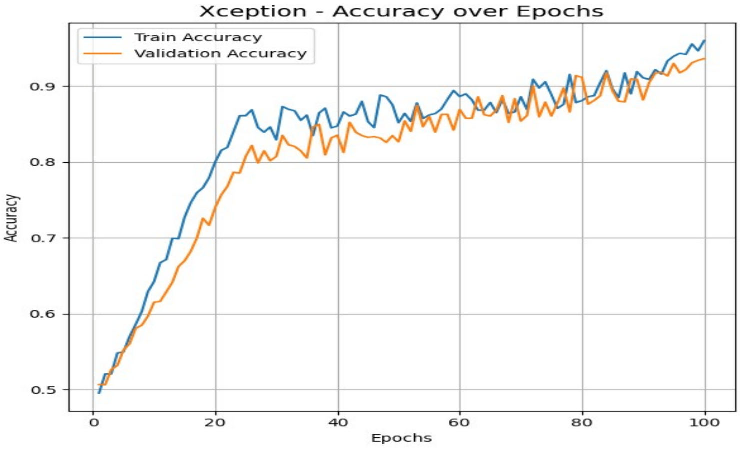

Figure 6.4: Graph of Xception

6.1.5 ViT

We fine-tuned the pre-trained model by starting with a learning rate of 3e-4 and then gradually reduced to 1e-4 for fine tuning. Also, we used the AdamW optimizer and Cosine Annealing Scheduler here. By freezing earlier and then unfreezing later gradually helped us performing better. Finally, we achieved 94.38% accuracy on the train data and 89.22% on the validation data. As for auc value we got 97% on train data and 92% on validation data. for auc value we got 97% on train data and 92% on validation data. for auc value we got 97% on train data and 92% on validation

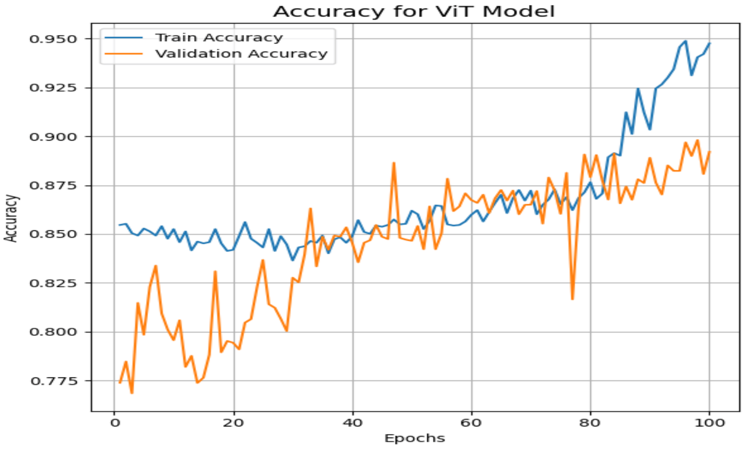

Figure 6.5: Graph of standard ViT

data.

6.1.6 DeiT

We resized the image to 224×224 pixels. Then we used rotation, flipping, color jitter and CLAHE filter and after that we imported pre-trained DeiT model. We used the patch size as 32×32 and used the AdamW optimizer so that we can prevent overfitting by weight decay. Initially we started training with 0.0001 learning rate and then used Cosine Annealing Learning Rate Scheduler to reduce the learing rate followed by a cosine curve so that we get stability in terms of accuracy. Finally, we got 93% train accuracy, 85% validation accuracy, 96% train auc and 92% validation auc. We can observe from the figure that the training accuracy improves over time.

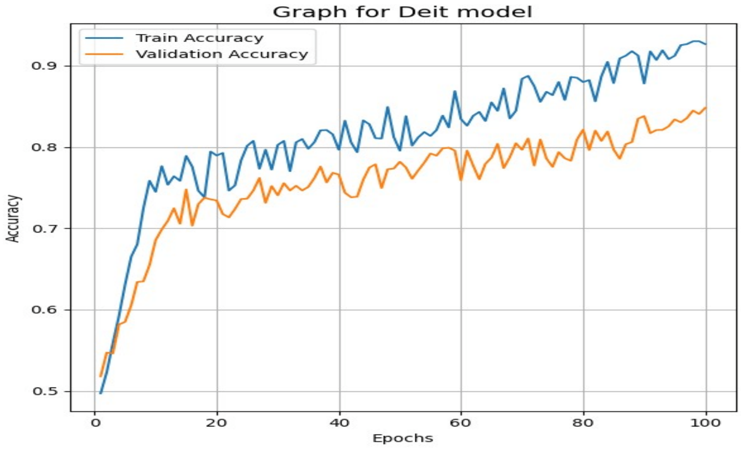

Figure 6.6: Graph of DeiT

6.1.7 Swin

After resizing, rotating,flipping and applying CLAHE filter on the image, we fine-tuned the model on our dataset by freezing, unfreezing layers. Here, AdamW was used and we used a small weight decay of 0.01 to regularize the model and prevent overfitting. Eventually we got 95.81% train accuracy, 89.66% validation accuracy, 98% train auc and 94% validation auc.

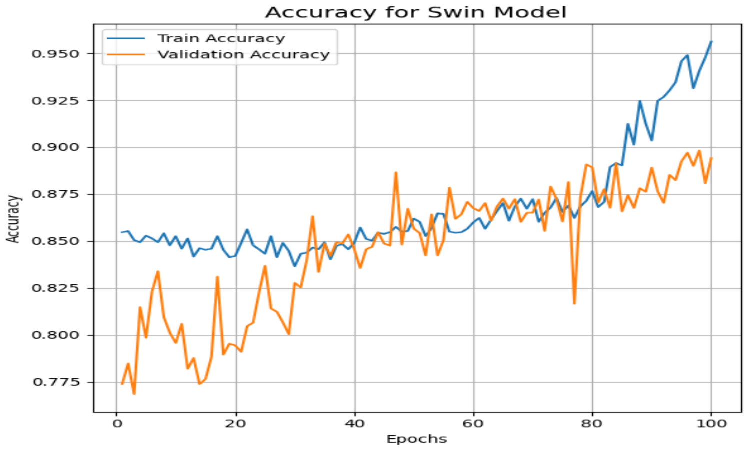

Figure 6.7: Graph of Swin

6.2 Performance of DR Detector

After implementing DiabeticRetinoScope dataset in our DRdetector model ,it shows an accuracy of 97.38% and a validation accuracy of 91.23% . For every class, the model's test accuracy was 89.40%. Comparing our pre-trained Resnet-50 model to other pre-trained models, we managed to get superior outcomes for the transformer layers, which include the Multi-Head Self-Attention layer, Residual Connections, Layer Normalization, and A Feed Forward Network. With adequate dataset preprocessing, we were able to get this outcome. As we continue to examine the shortcomings, we hope to increase the accuracy of this dataset on the DRDetector model. As we can see in the figure, both the training and validation accuracy increase over time.

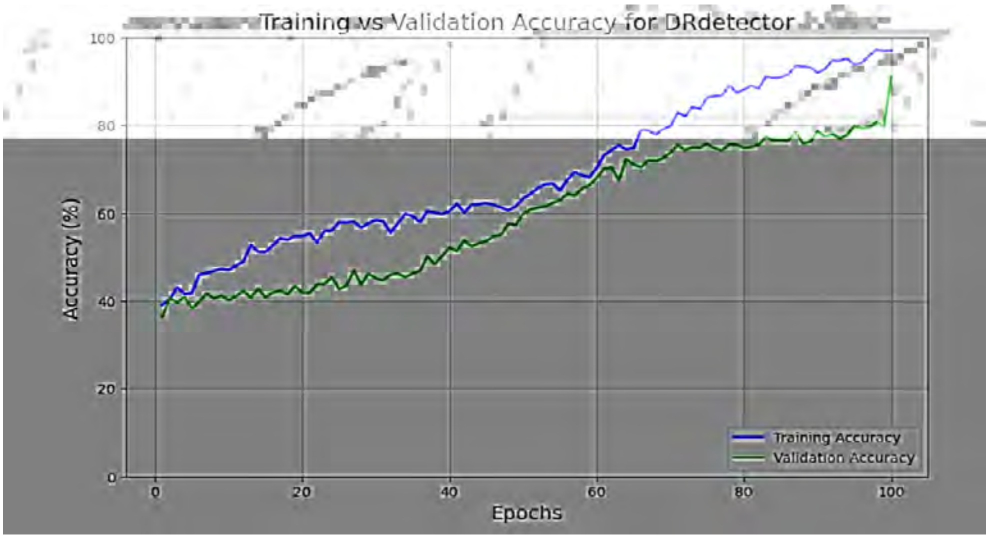

Figure 6.8: Accuracy Graph of DRDetector

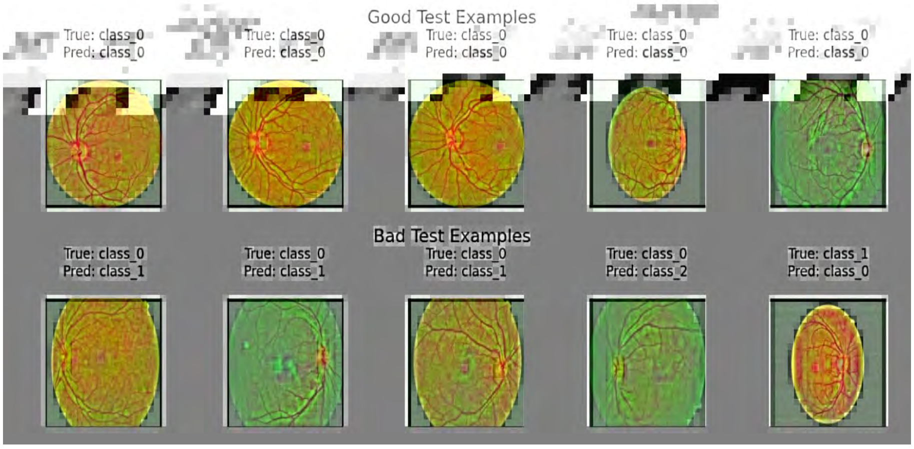

Figure 6.9: Good Test and Bad Test for DRDetector

6.2.1 Class 0:

For class 0 our DR Detector model got a precision accuracy of 87.8%, recall of 91% and F-1 score of 89.37%.

6.2.2 Class 1:

Our model performed well for class 1 also. It got a precision accuracy of 90.2%, recall of 93% and F-1 score of 91.6%.

6.2.3 Class 2:

For Class 2 this model gained a precision accuracy of 87.8%, recall of 86.5% and F-1 score of 87.18%.

6.2.4 Class 3:

DR Detector gained a well enough performance for class 3, precision accuracy of 89.5%, recall of 89.5% and F-1 score of 89.5%.

6.2.5 Class 4:

For class 4 our DR Detector model got a precision accuracy of 91.5%, recall of 87% and F-1 score of 89.1%.

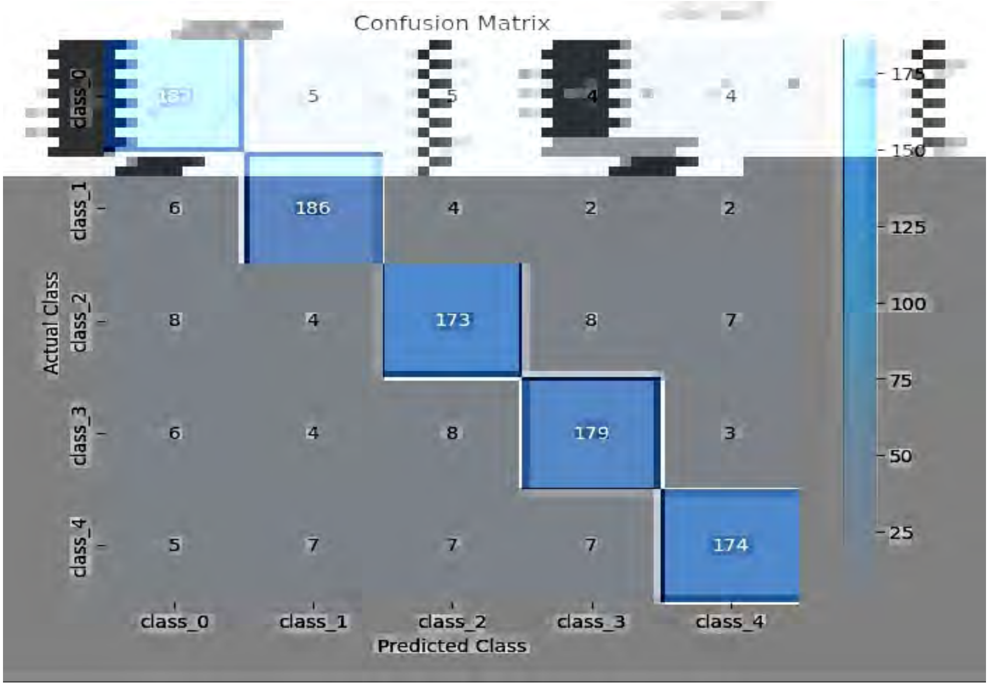

Figure 6.10: Confusion Matrix of DR Detector model class wise

Chapter 7

Performance Analysis

We underlined in the thesis that our main focus would be on two deep learning models, namely ResNet-50 and layers of vision transformers. The choice of ResNet and ViT is because of the detailed related literature where these models have shown better performance in predicting retinopathy compared to the availability of other models. Our first choice was ResNet because it introduced the use of skip connections in its architecture. ResNet50, one variation in the ResNet family of models, has a total of 50 layers and was used architecturally for this study. Therefore, the images in the Diabetic RetinoScope dataset were then split into training, validation, and testing sets; the former contained 70% of the total images, while the latter contained 20% and 10%, respectively. Later, we labeled the training images with the severity labels found within the associated CSV files. All images were resized to 224x224 pixels, ensuring uniformity in the dimension of the images. We set the batch size for training the neural network to 8, with a learning rate of 1e-4. Further, we put in a total training epoch to 100.

In this work, a pre-trained model was used for feature extraction, following the ResNet50 architecture. Additional layers were added in order to adapt the model for the requirements of ViT on the dataset that we have created, namely Diabetic RetinoScope. First of all, we froze all the layers of the model by setting their trainable attribute to False. This was a precautionary measure in order not to update the weights of these layers. Next, we schedule the model for fine-tuning by strategically unleashing the last five layers such that only their weights can be updated in successive training iterations. In addition, fine-tune the whole model for more training. Then, test the model on the test set, and draw some metrics needed like a confusion matrix, other diagrams to verify its performance.

7.1 Result Evaluation

We see, among the pre-trained models Resnet-50 performed the best on our dataset (Diabetic RetinoScope). With a accuracy of 86.89% and F1 score of 88.52%, among all the pre-trained models. Secondly, Xception, Swin and InceptionV3 got a good enough result on our dataset also ranging near the mark of 95%. From this, we had the inspiration to build custom model based on Resnet-50 adding some prominent Transformer layers to the model. We named the model as, DRDetector. Which aims to have a higher accuracy than these pre-trained models. As we see the DRDetector model perform best among all the models that was run with our dataset(Diabetic

RetinoScope). As we have seen that the train and validation accuracy was gradually increasing for the DRDetector model, with some more fine tuning we will get higher accuracy result on our dataset.

Table 7.1: Performance Comparison of Different Models

MODEL NAME	**ACCURACY**	**PRECISION**	**RECALL**	**F1-SCORE**
DR Detector	89.40%	89.36%	89.40%	89.35%
RESNET-50	86.89%	87.11%	86.89%	86.93%
VGG 16	84.56%	84.59%	84.56%	84.52%
INCEPTION-V3	85.18%	85.22%	85.18%	85.20%
XCEPTION	85.84%	85.91%	85.84%	85.87%
VIT	85.04%	85.19%	85.04%	85.11%
DEIT	82.77%	82.69%	82.77%	82.72%
SWIN	86.13%	86.07%	86.13%	86.09%

Chapter 8

Conclusion

The key objective of the carried-out research work is to consider the identification and prognosis of DR, a very critical subject with the aim of mitigating visual impairment as well as ocular blindness in people suffering from diabetes. In that respect, a novel deep learning technique is designed for diagnosis and severity grading of the retinal conditions, which is to be named as DRDetector. With the pressing need for early detection of DR, we have performed extensive work in problem definition, data collection, and the development of a particular deep learning model. In particular, we have gainfully employed the capabilities of CNNs and ViT, especially ResNet, in extending the capability for automatic diagnosis of retinal diseases. This development has been achieved by merging the best data from two independent datasets, namely ”Diabetic Retinopathy Detection 2015” and ”Aptos 2019 Blindness Detection”, in our build dataset named ”DiabeticRetinoScope”. Further, our research goals are an improvement of the state of current severity grading of the retinal diseases for early detection and eventually the treatment of the patients. The deep learning technologies are being proved to bring a paradigm shift in the detection of DR by performance of in-depth evaluation and comparative analysis with traditional diagnostic methods. A literature review was performed and, after this review, we found numerous deep learning-based methods which contributed much to the advancement of the diagnosis of diabetic retinopathy (DR). The emphasis on employing a variety of techniques and architectures in these methods points toward possible automation within the healthcare domain and, correspondingly, fighting diseases such as DR. This work rests on a structured approach, whose steps include the following: data collection, preparation, model selection, model training, and model evaluation. For this work, our model is further improved by incorporating advanced CNN and ViT with pre-trained ResNet50 and transformer layers like MultiHead Self-Attention layer, Layer Normalization, and Feed Forward Network. This paper tends to enable early DR detection for which medical doctors can diagnose and perform more accurate and early interventions in curing this debilitating disease. It would be our vision that, in some time to come when advanced technology would meet clinical practice, vision loss on account of diabetic retinopathy would be prevented and appropriately managed. Generally, this study further consolidates deep learning among the vast set of conventional methods for diabetic retinopathy screening and diagnosis. This can eventually bring a complete transformation in the management of diabetic retinopathy and improvement in the outcome.

Bibliography

[1] T. Walter, J.-C. Klein, P. Massin, and A. Erginay, "A contribution of image processing to the diagnosis of diabetic retinopathy-detection of exudates in color fundus images of the human retina," *IEEE Transactions on Medical Imaging*, vol. 21, no. 10, pp. 1236–1243, 2002.

[2] C. E. Hann, J. A. Revie, D. Hewett, J. G. Chase, and G. M. Shaw, "Screening for diabetic retinopathy using computer vision and physiological markers," *Journal of Diabetes Science and Technology*, vol. 3, no. 4, pp. 819–834, 2009.

[3] M. D. Abràmoff, M. K. Garvin, and M. Sonka, "Retinal imaging and image analysis," *IEEE Reviews in Biomedical Engineering*, vol. 3, pp. 169–208, 2010.

[4] C. Agurto, V. Murray, E. Barriga, *et al.*, "Multiscale am-fm methods for diabetic retinopathy lesion detection," *IEEE Transactions on Medical Imaging*, vol. 29, no. 2, pp. 502–512, 2010.

[5] V. Gulshan, L. Peng, M. Coram, *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA*, vol. 316, no. 22, pp. 2402–2410, 2016.

[6] H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding, and Y. Zheng, "Convolutional neural networks for diabetic retinopathy," *Procedia Computer Science*, vol. 90, pp. 200–205, 2016.

[7] M. Mateen, J. Wen, Nasrullah, S. Song, and Z. Huang, "Fundus image classification using vgg-19 architecture with pca and svd," *Symmetry*, vol. 11, no. 1, p. 1, 2018.

[8] Anonymous, "Aptos 2019 blindness detection," *APTOS 2019 Blindness Detection*, 2019. [Online]. Available: https://www.kaggle.com/competitions/aptos2019-blindness-detection/data.

[9] S. H. Kassani, P. H. Kassani, R. Khazaeinezhad, M. J. Wesolowski, K. A. Schneider, and R. Deters, "Diabetic retinopathy classification using a modified xception architecture," in *2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, IEEE, 2019, pp. 1–6.

[10] N. E. M. Khalifa, M. Loey, M. H. N. Taha, and H. N. E. T. Mohamed, "Deep transfer learning models for medical diabetic retinopathy detection," *Acta Informatica Medica*, vol. 27, no. 5, p. 327, 2019.

[11] W. L. Alyoubi, W. M. Shalash, and M. F. Abulkhair, "Diabetic retinopathy detection through deep learning techniques: A review," *Informatics in Medicine Unlocked*, vol. 20, p. 100 377, 2020.

[12] S. Mishra, S. Hanchate, and Z. Saquib, “Diabetic retinopathy detection using deep learning,” in *2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE)*, IEEE, 2020, pp. 515–520.

[13] L. Dai, L. Wu, H. Li, *et al.*, “A deep learning system for detecting diabetic retinopathy across the disease spectrum,” *Nature Communications*, vol. 12, no. 1, p. 3242, 2021.

[14] G. Mushtaq and F. Siddiqui, “Detection of diabetic retinopathy using deep learning methodology,” in *IOP Conference Series: Materials Science and Engineering*, IOP Publishing, vol. 1070, 2021, p. 012 049.

[15] E. AbdelMaksoud, S. Barakat, and M. Elmogy, “A computer-aided diagnosis system for detecting various diabetic retinopathy grades based on a hybrid deep learning technique,” *Medical Biological Engineering Computing*, vol. 60, no. 7, pp. 2015–2038, 2022.

[16] S. Bai, J. Li, and Y. Zhang, “Transformer-based model for detection of diabetic retinopathy in retinal images,” in *2022 International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Springer, 2022, pp. 435–445.

[17] X. Zhao, Q. Yang, and L. Zhang, “Comparative study of cnn and vision transformer for diabetic retinopathy classification,” in *2022 International Conference on Artificial Intelligence and Data Science (ICAIDS)*, IEEE, 2022, pp. 115–120.

[18] A. K. Anumol Sajan and M. S. M. Kurian, “Diabetic retinopathy detection using deep learning,” *Journal Name*, 2023.

[19] R. A. Dihin, E. Alshemmary, and W. A. M. Al-Jawher, “Diabetic retinopathy classification using swin transformer with multi wavelet,” *Journal of Kufa for Mathematics and Computer*, vol. 10, no. 2, pp. 167–172, Aug. 2023.

[20] V. Fernandes, “Using data-efficient image transformers for diabetic retinopathy severity classification,” *Research Gate*, vol. 6, no. 1, pp. 7–9, 2023.

[21] J. George and R. Varma, “Deep learning approaches for diabetic retinopathy: A comprehensive review,” *Applied Sciences*, vol. 13, no. 2, p. 853, 2023.

[22] I. Mohsen, L. Khedher, and A. Khelifa, “Multi-modal approach for diabetic retinopathy detection using cnn and vision transformer,” *Journal of Healthcare Engineering*, vol. 2023, pp. 1–12, 2023.

[23] W. Nazih, A. O. Aseeri, O. Y. Atallah, and S. El-Sappagh, “Vision transformer model for predicting the severity of diabetic retinopathy in fundus photography-based retina images,” *IEEE Access*, vol. 11, no. 3, pp. 117 546–117 561, 2023.

[24] A. Singh, V. Kumar, and R. Gupta, “A comparative study of vision transformer and cnn for medical image classification tasks,” *Journal of Medical Systems*, vol. 47, no. 2, p. 75, 2023.

[25] Y. Tahiri, A. Moussaoui, and K. Abdelkader, “A vision transformer for classifying diabetic retinopathy from fundus images,” *Journal of Biomedical Imaging*, vol. 2023, pp. 1–10, 2023.

[26] Y. Wang, Y. Zhao, and X. Liu, “Exploring the efficacy of swin transformer for medical image analysis: A case study on diabetic retinopathy,” *IEEE Transactions on Medical Imaging*, vol. 42, no. 5, pp. 1580–1592, 2023.

[27] T. Yang, H. Wang, and Y. Chen, “A hybrid cnn-vit model for automated diabetic retinopathy detection,” *Frontiers in Bioengineering and Biotechnology*, vol. 11, p. 1342, 2023.