# An interpretable diagnosis of retinal diseases using Vision Transformer and Grad-CAM

by

Mahdi Hasan Bhuiyan
20101541
Sumit Haldar
20101544
Maisha Shabnam Chowdhury
20101459
Nazifa Bushra
20101536
Tahsin Zaman Jilan
20101581

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

Department of Computer Science and Engineering
Brac University
January 2024

# Declaration

It is hereby declared that

1. The final thesis report for defence submitted is my/our own original work while completing degree at Brac University.
2. The final thesis report for defence does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The final thesis report for defence does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

**Student's Full Name & Signature:**

Mahdi Hasan Bhuiyan
20101541

Sumit Haldar
20101544

Maisha Shabnam Chowdhury
20101459

Nazifa Bushra
20101536

Tahsin Zaman Jilan
20101581

# Approval

The the final thesis report for defence of thesis titled "An interpretable diagnosis of retinal diseases using Vision Transformer and Grad-CAM" submitted by

1. Mahdi Hasan Bhuiyan(20101541)
2. Sumit Haldar (20101544)
3. Maisha Shabnam Chowdhury (20101459)
4. Nazifa Bushra (20101536)
5. Tahsin Zaman Jilan (20101581)

Of Fall, 2023 has been accepted as satisfactory in fulfillment of the requirement for the final thesis of B.Sc. in Computer Science on January, 2024.

**Examining Committee:**

Supervisor:
(Member)

Md. Ashraful Alam, PhD
Associate Professor
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Md. Golam Rabiul Alam, PhD
Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

# Abstract

Early detection of retinal diseases can help people avoid going completely or partially blind. In this research, we will be implementing an interpretable diagnosis of retinal diseases using a hybrid model containing VGG-16 and Swin Transformer and then visualize with Grad-CAM. Using Optical Coherence Tomography (OCT) Images gathered from various sources, a unique multi-label classification approach is developed in this study for the diagnosis of various retinal diseases. For the research, a transformer-like hybrid architecture will be used, which is Vision Transformer that works by classifying images. Recent developments in competitive architecture for image classification include the original concept of Transformers. The implication of this architecture is done over patches of images often called visual tokens. It can handle different data modality. A ViT employs several embedding and tokenization techniques. In order to accurately highlight key areas in pictures, the gradient-weighted class activation mapping, known as (Grad-CAM) technique has been used so that deep model prediction can be obtained in image classification, image captioning and several other tasks. It explains network decisions by using the gradients in back-propagation as weights. We used both VGG-16 that is a variant of Convolutional Neural Networks (CNN) and Swin Transformers in our model. We combined these two and introduced a hybrid model. After being tested, the VGG-16 component's output accuracy was 0.8888, while the Vision Transformer component's accuracy was 0.9139. Then the hybrid model was tested after some fine tuning and it performed extraordinarily. The output accuracy of the hybrid model is 0.988.

**Keywords : Deep Learning, Vision Transformers, Ocular diseases screening, Grad-CAM, Detection, Diagnosis, classification**

# Table of Contents

# Chapter 1

# Introduction

## 1.1 Background

Retinal disease refers to any condition that affects the retinal area of the eye which is the light-sensitive layer of tissues by which the brain receives electrical signals to generate images after the conversion of light. Retinal illness has traditionally received little attention in developing nation programs aimed at preventing blindness stated by Yorston & David [1]. Millions of people of all ages are affected by retinal illnesses, which are a serious health concern on a global scale. It is projected by Nazimul, et al. [3] that by 2030, there will be more than 42 million retinal disease victims in industrialized nations and over 82 million in developing countries who are older than 64. In older adults, blindness is primarily caused by age-related macular degeneration known as AMD. The World Health Organization (WHO) estimates that there were 288 million AMD sufferers worldwide in 2010; by 2020, this figure is projected to rise to about 600 million[19] .

Other retinal diseases such as Diabetic retinopathy, retinal detachment, retinitis pigmentosa, uveitis and others also affect a large number of people worldwide. Retinal disease can be fully cured if it gets diagnosed in an early stage. Furthermore, Retinal repair can stabilize eyesight and stop additional vision loss, even if a patient's vision may not fully return. It's really important that patients receive therapy as quickly as possible for their injured retinas. Thus, diagnosing retinal disease with the least amount of time became necessary. But for medical professionals, recognizing retinal illnesses can be difficult, especially when dealing with subtle or complex symptoms in photographs. Fortunately, Deep learning algorithms have been employed more frequently in recent years to accurately diagnose diseases by analyzing retinal images [14]. With the help of deep learning models, quicker diagnosis of retinal disease can be possible [11].

GradCam and vision transformer use is one possible strategy for diagnosing retinal diseases in place of traditional convolutional neural networks (CNNs). In contrast to convolutional neural networks (CNNs), transformers have proven to be an effective tool for vision tasks in recent research, showing competitive performance mentioned in the paper by Zhou, Daquan, et al. [20]. GradCam is a method that makes it possible to see which areas of an image are most crucial for a CNN model's prediction [6]. It has been demonstrated that the neural network architecture known as

"vision transformer" performs well on a range of image and vision-related tasks [41]. By combining these two techniques, it's possible to not only identify the presence of a retinal disease but also understand the most applicable areas of the image for the prediction, providing insights to medical professionals.

## 1.2 Problem Statement

For the early detection and treatment of illnesses such as (AMD) and diabetic retinopathy, reliable retina related disease identification from pictures is crucial. For medical specialists, identifying these disorders can be difficult, especially when dealing with the images' nuanced or intricate symptoms. While DL techniques have been visually constructive in identifying retinal diseases from images, they lack the interpretability needed for medical professionals to trust the predictions. To make it easy we will use deep learning models.

GradCam can be used to determine which parts of an eye image are crucial for the network to diagnose a particular condition, and vision transformers can be used to categorize an image into a particular disease or healthy condition based on the context of the image when it comes to identifying the retinal diseases. The presentation and explicability of the diagnosis system can be enhanced by combining the two methodologies. More specifically, on to the first stage, we are going to do image processing to enhance the different areas of retinal structure. Such as retinal vessels, optical disk, fovea. By enhancing there areas of the eye, it will be easier to diagnose the abnormalities and other types of symptoms for retinal disease.
To identify retinal disease, one of the most important things is to find the region of focus in the retinal image. A vision transformer can be used for this purpose, as it utilizes an attention mechanism. To lower the complexity of the transformer model, saliency mapping can be used to determine the region that should be focused on. On the other hand, GradCam will be used in comparison with saliency mapping to achieve the highest accuracy. The vision transformer is able to pinpoint in which region Gradient-weighted Class Activation Mapping (Grad-CAM) approach can be applied. For instance, Using GradCam, we may still utilize the DR model to extract some information from the AMD image even though CNV does not belong to any classes in DR. Though the accuracy of gradCam is low compared to saliency mapping. By combining both, improvement in the accuracy of GradCam can be achieved.

## 1.3 Research Objective

The objectives of our thesis are-

- Early detection and treatment: To facilitate early detection and treatment of retinal diseases such as AMD and diabetic retinopathy is the primary objective.

- Reliable retinal disease identification: Identification of retina-related diseases from images is crucial for medical specialists dealing with nuanced or intricate symptoms, so to develop a reliable method is necessary.
- Deep learning models for diagnosis: deep learning models such as vision transformers can be used for categorizing images into certain diseases or in healthy conditions based on provided information.
- Interpretability for medical professionals: The interpretability obstacles of deep learning techniques can be addressed by combining Grad-CAM to determine crucial parts of an image of eye for diagnosis purpose.
- Identification of region of focus: To identify the region of focus in retinal images, utilization of vision transformer with an attention mechanism addresses one of the crucial aspects of identification of retinal diseases.

## 1.4 Research Orientation

The prior study done by other researchers that is connected to our subject topic is shown in chapter 2. Then, in chapter 3, we went into detail about each and every algorithm, convolution layer, and activation function we had utilized. Chapter 4 once more explains how our thesis work was put into practice. We also demonstrated the distribution of our datasets before describing the pre-processing methods utilized on the dataset in question. In chapter 5, we demonstrated the outcomes of using the algorithms and then discussed the analysis of the outcomes. Finally, we presented the conclusion.

# Chapter 2

# Existing Works

Transformer is a kind of deep neural network that is primarily based on the self-attention mechanism. It was initially used in the field of natural language processing. Transformer-based models outperform convolutional and recurrent neural networks, among other types of networks, in a range of visual benchmarks stated by K. Han, et al in their study [40]. S. S. M. Sheet, et. al in the paper [33], for improving the condition of micro-aneurysms pixels, found Contrast-Limited Adaptive Histogram Equalization (CLAHE) more effective than others. They proved that despite having limitations in the training dataset, the network of RESNET50 offers good classification accuracy following the use of augmentation. The other state-of-arts method which was introduced previously by other researchers falls behind when placed in competition with the proposed design. It achieves classification accuracy of 95.63%, validation accuracy of 92.99% and 40 to 100% in sensitivity range using the VGG19 network and STARE database. Because of its superior quality, it can be used in electrocardiographs, ultrasonic imaging systems or other medical diagnostic equipments. But it had a drawback that as The supplied training sample was less so its classification covered only five retinal diseases.

The researchers S. A. Kamran, et. al [16] proposed a semi-supervised conditional GAN named VTGAN that can differentiate between healthy and malfunctioned retina along with creating FA images i.e the retinal vascular structure from fundus photographs at the same time. They used Frechet inception Distance (FID) to evaluate quantitatively and KernelInception Distance (KID) to quantify image features and to measure the similarity in structure. In another paper A. K. Bitto, et. al [21], to identify and classify conjunctivitis eye, cataract eye and normal eye, these three eye conditions they used the CNN based transfer learning (TL). As it consists of a large number of sample images in the provided dataset, it is more preferable in experimental works. Among the three deep CNN architectures, to diagnose eye disease, the greatest correctness of 97.86% and less time of 485 seconds is provided by Inception-v3. Secondly the accuracy of 95.68% and time of 1090 seconds is provided by ResNet-50 and lastly the accuracy of 95.48% and highest time of 2510 seconds is provided by VGG-16. This paper has a plan for the future goal of inventing a model for diagnosing eye disease with a real time approach like YOLO.
Then again the researcher's M. Subramanian, et. al of another paper [34] ,they used transfer learning for having a great scope of advantages such as it doesn't require relatively great size of training datasets, not required huge processing effort as only

the weights of few top layers are considered. As the project was made from scratch, it needed much computational power so they used models for extraction and fine tuning like VGG16, DenseNet201, InceptionV3, and Xception. This research can be further expanded for creating deep learning models which will require less parameters and less time.

Physicians and also the patient should be mindful of illnesses that have a propensity to simultaneously affect the patients neurological system and eyes. Thus, treating it in time with accuracy is important. Letting retinal disease untreated can result in blindness in a number of cases. The computer-aided automatic diagnosis model demonstrates excellent capabilities in processing medical images and may be considered to be a major strategy in current investigations. Eye problems, heart problems, rectum problems, brain tumor problems, and breast cancer problems are all examples of retinal disorders. Additionally, DL (Deep Learning) techniques have made significant advances to the accurate diagnosis and forecasting of diabetic retinopathy. First, the CLAHE filter produces sufficient vein enhancement and maintains its durability in noisy environments. CLAHE is typically used to enhance the contrast between the tinted dark areas (BVs, HMs, and MAs in the shade-corrected image) and the backdrop. The drawback mentioned by the author which is classification only included five categories of retina classes. Therefore, additional research is necessary to assess the models performance in identifying more illness instances. Blindness results from delayed or untreated infection. Therefore, it is essential to find retina infections early on. This paper introduces an updated CLAHE approach for increasing image brightness. Using transfer learning on the CNN (RESNET50) model, the enhanced-fundus images were examined to identify retinal diseases.

In paper of S. S. M. Sheet, et. al [33], physicians should be mindful of illnesses that have a propensity to simultaneously affect the patients neurological system and eyes. Thus, treating it in time with accuracy is important. Leaving retinal disease untreated can result in blindness in a number of cases. The computer-aided automatic diagnosis model demonstrates excellent capabilities in processing the medical images and may be considered to be a major strategy in current investigations. Eye problems, heart problems, rectum problems, brain tumor problems, and breast cancer problems are all examples of retinal disorders. The specialists in medical image processing have been concentrating more on the use of artificial intelligence (AI) in analysis, disease diagnosis, and prediction. Additionally, DL (Deep Learning) techniques have made significant advances to the accurate diagnosis and forecasting of diabetic retinopathy. According to current studies, there are two main stages of eye disorders. The first stage aims to identify diabetic retinopathy (DR) symptoms. The second stage is a brand-new field of study that addresses a new range of retinal diseases. The Contrast-Limited Adaptive Histogram Equalization (CLAHE) filter was successful in enhancing the micro-aneurysm pixels in order to improve the accuracy of diagnosis. First, the CLAHE filter produces sufficient vein enhancement and maintains its durability in noisy environments. The shade-corrected image's coloured dark parts (BVs, HMs, and MAs) are often enhanced using CLAHE so that they stand out more against the image's background.. The drawback mentioned by the author is that its classification only included five categories of retina categories. Therefore, additional research is necessary to assess the models per-

formance in identifying more illness instances. The increased performance of the suggested models may be included into medical screening equipment for viewing the retina, including the electrocardiograph, ultrasonic imaging system, and other medical detection devices. Blindness results from delayed or untreated infection. Therefore, it is essential to find retina infections early on. This paper introduces an updated CLAHE approach for increasing image brightness. By applying transfer learning on the CNN (RESNET50) model, retinal disorders were identified in the enhanced-fundus pictures.

We have to consider something as the golden standard for eye disease diagnosis, then it will be OCT which is Optical coherence tomography where the interior microstructure of the eye may be seen in vivo using noninvasive, high-resolution optical medical diagnostic imaging according to the paper of V. Annavarjula [5]. Most of the earlier studies acquired high accuracy based on CNN, but vision transformers don't rely on CNN and have different feature extraction methods from it. When the vision transformer was applied to categorization of remote sensing images and testing of several remote sensing image data sets, the accuracy rate was higher than CNN.In this paper, three types of images such as AMD , DME disease, and normal ocular fundus were collected then integrated OCT and vision transformer, and train it to improve the diagnosis. AMD and DME substantially impair the vision of the elderly. For patients, prompt diagnosis and treatment are crucial. In this study, they provided a CAD technique for classifying OCT fundus pictures utilizing vision transformers. After trimming, we demonstrated that the vision transformer had the quickest recognition speed without sacrificing recognition accuracy.

In this paper of M. Subramanian, et. al [34], by June 2021, there were 173 million COVID-19 instances worldwide, and the number is steadily rising. Given that the virus can cause serious sickness and even death, early detection is essential. Early diagnosis is still difficult, nevertheless, because of the scarce resources and data. The typical diagnosis method for COVID-19 is RT-PCR, however it has a number of disadvantages, including cost, risk to medical personnel, and a lack of readily available diagnostic test kits. X-ray and CT-based screening are two medical imaging procedures that are reasonably safe, quick, and more accessible. Deep learning and data science have demonstrated good outcomes in the diagnosis of COVID-19 from CT and X-ray images in several areas of medical imaging, including thoracic imaging. These include object detection-based approaches, approaches based on unsupervised learning, and the most popular method, deep convolutional neural networks (CNNs) with supervised learning.

This thesis of M. Wasse, et. al [36] covers the difficulties of early COVID-19 diagnosis and the limits of RT-PCR as a diagnostic technique. As an alternative way for detecting COVID-19, the report also discusses the use of medical imaging techniques, such as X-ray and CT-based screening. However, the precision of modern imaging techniques is highly dependent on the expertise of radiologists, particularly in underdeveloped nations. The research then explores the application of deep learning and data science in medical imaging, specifically thoracic imaging, for the diagnosis of COVID-19 using CT and X-ray images. To detect COVID-19 with variable degrees of accuracy, a number of research have employed diverse techniques, includ-

ing unsupervised learning, object detection, and deep convolutional neural networks.

The difficulties of early COVID-19 diagnosis and the limits of RT-PCR as a diagnostic technique re covered in the thesis of R. Fan, et. al [39]. As an alternative strategy for detecting COVID-19, the article advises using medical imaging techniques, including X-ray and CT-based screening. However, the precision of modern imaging techniques is highly dependent on the expertise of radiologists, particularly in underdeveloped nations. The research also investigates the application of deep learning and data science in medical imaging, particularly thoracic imaging, for the diagnosis of COVID-19 from CT and X-ray images. To recognize COVID-19 with variable degrees of accuracy, a number of studies have employed diverse techniques, such as unsupervised learning, object detection, and deep convolutional neural networks. In addition, a few object detection-based techniques, such as the YOLO-based object detection model, which achieved a detection accuracy of 90.67 percent, have been employed to detect COVID-19. Utilizing deep convolutional neural networks (CNNs) with supervised learning has been the most common method for completing this goal. As an illustration, a study by Mukherjee et al.

In this paper of Z. Ma, et. al [29] introduces the need for the advancement of chest X-ray image methods for diseases like Pneumonia. A new model scheme was proposed that is based on the new transformer's mainframe networks and is fitted to the characteristics of the CXR. It can significantly increase the CXR's ability to identify pneumonia. Grad-cam and a transformer were used to extract the identifying decision criteria from the chest X-ray pictures and locate the lesion locations. In the experiment, the Swin transformer model achieved better accuracy than the traditional CNN models. The result of this work concludes that the model of CXR image recognition is optimized by the application of a Swin transformer. In future works, the Swin Transformer might be utilized as a backbone architecture for a wide variety of vision applications, such as picture categorization and object identification.

The idea of Neural Network is used so that it can support doctors in the process of diagnosis of Cardiovascular diseases by increasing the findings of constricting arteries according to the paper of K. Przystalski, et. al [42] . A comparison between different variants of the Inception Network and Vision Transformer has taken place in a stenosis detection task. Examination is done on the effects of various artery stenosis-free percentages in fragments on model performance and demonstrates the importance of the data set configuration. Inception-V3 Network and Vision Transformer variations were used in the experiments and measurements described in the papers. In order to highlight the key elements of the photos, gradient-weighted Class Activation Mapping images are displayed. The ratio of arteries in the negative data set to samples in the positive data set had a substantial impact on the results. The obvious point that they should constantly be concerned about the caliber of the data set utilized for training is emphasized by this. Conclusions concerning cardiovascular function should also take into account artifacts like picture borders or ribs. Results demonstrate that transformer-based designs are often outperformed by convolutional neural networks. Some ideas pertaining to future study in this topic might include the analysis of coronary angiography data as a whole video.

To make the suggested pavement image categorization model easier to understand, LeViT, a brand-new Transformer technique, was introduced for automatically classifying asphalt pavement images, in this paper of Y. Chen, et. al [24]. CNN based automatic pavement distress detection techniques take a lot of time and computational power. Also they have poor interpretability. Consequently, motivated by the effective use of Transformer architecture for Natural Language Processing (NLP) applications, LeViT was introduced in the paper. Describing its accuracy and it's better performance compared with the six state-of-the-art (SOTA) DL models. In this paper, three separate pavement image dataset sources and in order to implement the specified processes, ImageNet-based pre-trained weights were obtained. Moreover, the suggested model for pavement image categorization was made easier to understand by integrating Grad-CAM and Attention Rollout as a visualization technique to examine the categorization results and determine what was learnt in each MLP and LeViT attention block.

In this study of S. Park, et. al [32], implementing the Vision Transformer (ViT) architecture, this research paves the way for the self-attention mechanism to heavily use unlabeled data through structural modeling. Using the current ViT may not be the greatest solution because of how the features are included (direct patch flattening or ResNet backbone), which were not intended for CXR. To address this problem, the authors use a novel Multi-task ViT that creates a corpus of low-level CXR features from a backbone network that extracts frequently-encountered CXR findings. The foundation network is first trained on big public datasets to recognize common abnormal outcomes like consolidation, opacity, edema, etc. The embedded properties of the backbone network are then used as corpora for both the diagnosis and the therapy, and a flexible Transformer model is built. Previously suggested models for COVID-19 classification have been shown to have poor generalization abilities across a wide range of external data. On the other hand, this model's performance was stable regardless of whether it was seen from the PA or AP viewpoint, and across a wide variety of external test datasets with varying characteristics. This finding has important implications for expanding the existing model's actual applicability in the clinical setting.

J. Cao, et. al in the paper [23], have created an upgraded eye disease diagnostic system with more intelligible structure. They have also had their model tested in clinical use. In a lot of their studies they have used heatmaps to visualize the work of the algorithm or to see how the algorithm is working. But unfortunately they were not able to explain that model's error. As different diseases or abnormalities in the eye may have similar area of abnormal tissue but the distribution of abnormality might differ, the legion atlas was made to describe more efficiently about the lesion in UWF images. Furthermore, they also tried to recognize 21 anatomy regions of optic nerve and macula fovea. They have also presented that IEDSS performed with excellence in identifying common eye conditions and also conditions outside of training class.

By using artificial intelligence to study the features of the lesion, several researchers have tried to create models that can automatically detect patients with COVID-19 stated by the paper of B. Wang, et. al [35] . Convolutional neural networks (CNNs)

were primarily employed in the majority of these research to automatically identify patients on chest CT scans. To address the issue of the receptive field, experiments depending upon the Vision Transformer (ViT) design have recently been released. In order to create a new model for COVID-19 case identification, the Swin Transformer was used as the primary and vital network. Its performance on the pneumonia dataset is evaluated and contrasted with other models for the first time in this study. Its findings give medical professionals a foundation for selecting a top-notch pneumonia detection algorithm.The COVID-19 detection model (STCovidNet) presented in this study was trained and assessed using MUST-COVID-19 and uses the Swin Transformer blocks. The advised course of action produces the best outcomes and follows standards for medical judgment. This research suggests that STCovidNet is a viable architecture for locating COVID-19 situations.

Diagnosis using AI based architecture to identify kidney disease was made necessary by the global shortage of nephrologists, the public health concern of renal failure. This research paper of M. N. Islam, et. al [26] offers a much better accuracy of Swin transformer and the VGG16 model to diagnose kidney tumors, cysts and stones. They have showcased Six ML models of their own creation. Three of them are similar versions of Swin, EANet, CCT and the other three are similar to DL models VGG16, ResNet and Inception v3. They had been modified at the final layer for its best use. The models showed a great performance VGG16 and CCT but comparatively the accuracy of the swin transformer was far more superior and accurate. Furthermore, to build an automated AL diagnostic system and contribute to both the AL and the medical community, this research paper deals with three common types of renal diseases. Those are kidney tumors, cysts and stones. Thus, to perform the research, this paper has gathered 12 thousand CT of whole abdomen and urogram images.

As the paper say of G. Cai, et. al [22], the classification of skin diseases utilizing multimodal data, including photographs and clinical metadata, is presented in this thesis using a new neural network. Two encoders for the images and metadata are part of the network, which is referred to as a multimodal Transformer, together with a decoder for fusing the multimodal data. While the metadata are incorporated using a Soft Label Encoder, the image encoder uses a pre-trained Vision Transformer (ViT) model to extract deep features from the images. In order to successfully combine the image and metadata information, the decoder has a Mutual Attention block. The network outperformed cutting-edge techniques in tests conducted on both a private dataset of skin diseases and the ISIC 2018 benchmark dataset.

Figuring out the diseases associated with retina with the help of OCT images is nowadays the most fruitful CAD in the field of retinopathy and CNN helped researchers the most, providing great results with the OCT images according to the study of Z. Jiang, et. al [28] . In the encoder component construction the multi-head observation is a self-observation formation which permits the method to centre on several feature of info. The multi-head attention calculation approach uses scaled dot-product attention. Reshaping and normalization of the images are the foremost steps for data processing. The encoder component of VIT consists of six uniform encoders stacked in a pile. The model performs the classification job following the

output of the MLP head framework as well as the output of a layer of entire connection layer. The model's exactness on the test set is 99.69%. The key channels are chosen and the channels with zero or tiny coefficients are removed to create an effective classification model by promoting the invariance of networks in the vision transformer. The model presented in this work has the highest classification accuracy, outperforming previous studies, demonstrating that the vision transformer is more capable of recognizing objects in OCT fundus pictures than the CNN model and conventional machine learning methods.

The optical coherence tomography (OCT) images are classified using a hybrid ConvNet-Transformer network (HCTNet) in this study of Y. Ma, et. al [30] to aid in the detection of retinal disorders. The HCTNet produces features that make network training easier using a small-degree feature mechanism for extraction based on a remnant conv layer. Additionally, it features two parallel branches, one built on the Transformer model and the other on the ConvNet model, both of which are intended to take advantage of the global and local context of OCT pictures, respectively. The retrieved global and local variables are then used with an adaptive re-weighting a method for predicting the classification of OCT pictures in testing datasets. It was discovered that the HCTNet outperformed various classification techniques, including a pure vision Transformer (ViT) model and multiple ConvNet-based techniques, on two publicly available retinal OCT datasets. On the two datasets, Overall and average correctness for the HCTNet were 91.56 and 86.18 percent, correspondingly.

# Chapter 3

# Data Collection

## 3.1 Data Description :

A dataset refers to a collection of data. Here, we employed an image classification dataset, which is a collection of digital images used to assess, train, and evaluate the efficacy of machine learning algorithms. It is the dataset sample image from which the algorithms learn to assess and can effectively carry out the decision-making process. We made use of a dataset that was comprehensive and accurate in terms of the images. The dataset was taken from this website [12].

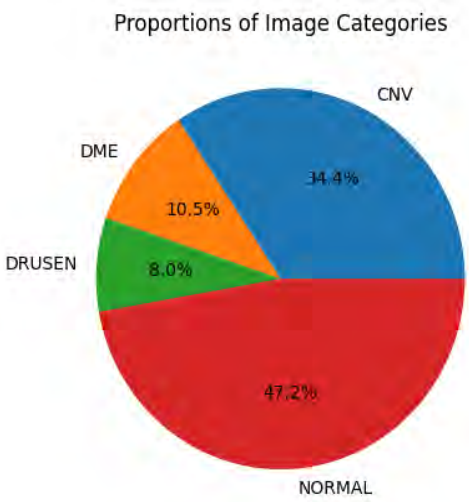


Figure 3.1: Dataset class proportions

It's called "Large Dataset of Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images" . This dataset is the 3rd version, where the first version was published on 2nd of June, 2018. The institutions engaged in this collection are University of California San Diego and Guangzhou Women and Children's Medical Center and its authors are Daniel Kermany, Kang Zhang and Michael Goldbaum. It has another folder in it named CellData, where there are two folders named code and OCT. Then the photos are divided into two folders: training and testing sets of independent patients. Images are labeled with (disease)-(randomized patient ID)-(image number assigned by this patient) and organized into four directories: CNV, DME, DRUSEN, and NORMAL.

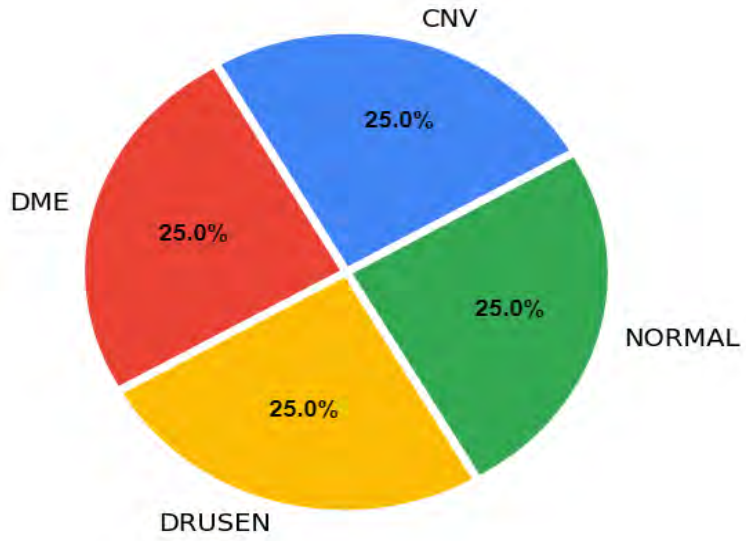


Figure 3.2: Dataset class proportions after balancing

**CNV(Choroidal Neovascularization)**: It is caused when some patients who have dry age-related-macular degeneration (AMD) gradually turns into wet AMD, this causes to grow abnormal blood vessels into the retina and leak fluid which eventually makes the retina wet[2]. By using the OCT, a cross-section picture of the retina can be obtained where if any fluid has leaked or any grayish macular lesion is observed with subretinal fluid, cystoid macular edema, exudation and/or hemorrhages, then CNV can be determined.

**DRUSEN**: They are the disease which is made up of lipids and proteins, can be of various sizes, which deposits under the retina as a yellow substance. When under the retinal pigment epithelium (RPE), many homogenous spherical yellow-white punctate accumulations then it can be identified as cutaneous drusen and if larger yellow-white dome-shaped deposit mounds are found under the RPE then they are called soft drusen [10].

**DME (Diabetic Macular Edema)**: It is the eye condition which is occurred in diabetic patients when due to poor glucose control, high blood sugar is observed which ultimately damages small blood vessels in the body including eye. In the OCT, if retinal swelling of the maula with reduced intraretinal reflectivity as DRT, intraretinal cystoid spaces of low reflectivity and highly reflective septa separating cystoid-like cavities in the macular area as CME, and a shallow elevation of the retina and an optically clear space between the neurosensory retina and retinal pigment epithelium as SRD is observed [4] then, DME can be identified.

**NORMAL**: When the retinal layers are well defined i.e. without significant abnormalities, this condition is known as normal. There is absence of irregularities nd notable fluid accumulation. Along with this, the foveal contour and central retinal thickness is assessible. When compared to the diseased eyes, the retinal layers, nerve fiber layer, ganglion cell layer, photo receptor layer are clearly visible in normal eyes, but in case of diseased eye conditions, irregularities, abnormalities in these layers

are clearly visible which ultimately results from AMD, DRUSEN, DME etc.

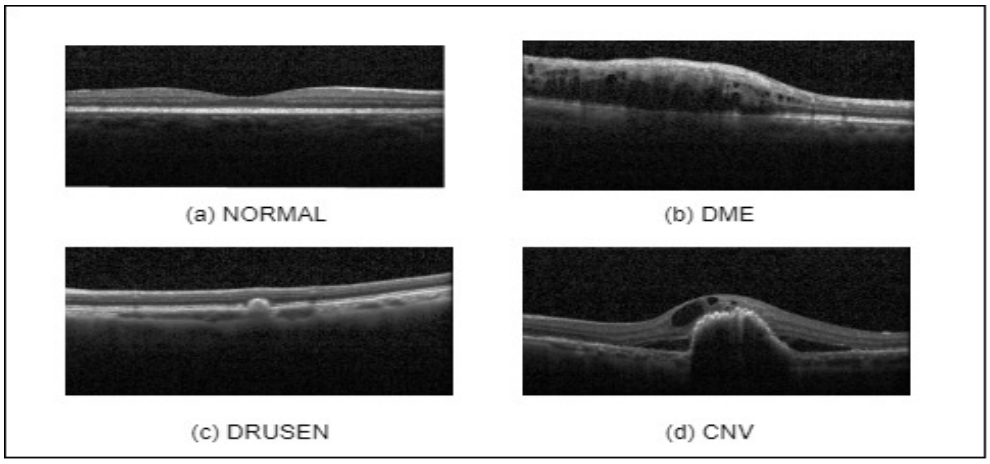


Figure 3.3: Different classes of retinal image

The train folder's subfolder CNV contains 37,205 images, DME contains 11,348 images, DRUSEN contains 8,616 images and NORMAL contains 51,140 images. And the test folder's subfolder CNV, DME, DRUSEN and NORMAL all contain 250 images.

## 3.2 Data Pre-Processing :

For data pre-processing ,first we loaded the dataset into the jupyter notebook. Then we made sure that our dataset is balanced. Imbalance of the dataset can cause biased model training. The model might become biased towards the majority class since they are exposed to more instances of that class during training. In our dataset, different classes have different numbers of images. So it's necessary to balance the dataset.There are many ways to handle an imbalanced dataset. Two of the most popular are Random Undersampling and Random oversampling. We could not use Random undersampling as it removes instances from the majority class. In that case , our dataset might have fewer instances to train the model . So we used Random oversampling to balance the dataset.

In this method, random pictures from the minority class were duplicated until the class distribution is more balanced. Then we used Data Augmentation for pre-processing. It helps to increase the diversity of the data by implementing various transformations like rotating , flipping , scaling , translating etc. This helps to enhance the dataset. We splitted the dataset into 3 subsets. Train , Test and Validation. 80% of the dataset was used to Train the model and 10% for testing while the other 10% for validation. Then We resize all samples for training and testing to $78 \times 78$ resolution.

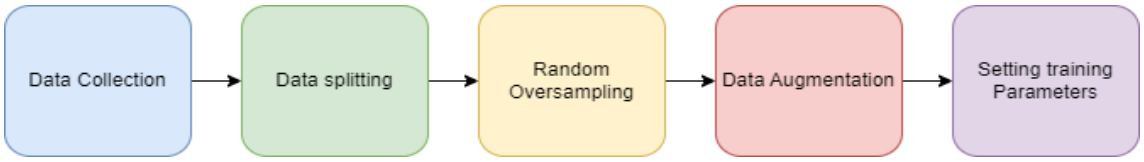


Figure 3.4: Class proportion after balancing

## 3.3 Experimental Configuration

The experiment was conducted on hardware containing Core i5 12th Generation processor with a RTX 3050 Ti GPU and 16 GB of Ram. We used a jupyter notebook to run the model.

| **Parameters** | **Value** |
|---|---|
| Image Size | 78 |
| Patch Size | 6 |
| Weight Decay | 0.0001 |
| Projection Dimension | 64 |
| Number of Patches | $\left(\frac{78}{6}\right)^2 = 169$ |
| Number of Encoder | 8 |
| Transformer Unit | (128, 64) |
| Heads | 4 |
| MLP Head | (2048, 1024) |

Table 3.1: Model Parameters

# Chapter 4

# Methodology

To begin with, we have selected the VGG16 algorithm first which is a Convolutional Neural Network supporting 16 layers. So that we can compare this result with the Swin Transformer which is a variant of Vision Transformer. In order to train the VGG16 model , we have used the Tensorflow library .We combined the Adam optimiser with accuracy measures, sparse categorical cross entropy loss, and the model. We have run this model up to 20 epochs with a batch size of 20 with a validation split of 0.1 .

Then we trained a Transformer model using our retinal image dataset(Swin Transformer). We fed the dataset images to Swin transformer model to see the results.During the acquisition procedure, several fundus photos looked to have a slanted retina, undesirable distortion, and a fuzzy impression. These circumstances necessitated the optimization of image features in order to improve the trained model's detection of illness classes.

In the domain of image classification, CNN based architectures are known to prioritize learning from the samples of majority class in imbalanced datasets which leads to biased classification and also overlooks some important informations. To overcome these issues we built a hybrid model combing VGG-16 and Swin Transformer. This hybrid model effectively overcomes the limitations of tradition CNN architectures and provides a more comprehensive understanding of global information in retinal diseases images. For this,We used the same data as inputs for both models and concatenate their results afterwards. In this study, a GradCam model based on visual transforms is proposed. It consists of residual, spatial feature fusion, upsampling and downsampling blocks for generators, and transformer encoder blocks for discriminators. In order to produce vivid fluorescein angiography images from normal and pathological fundus shots for training, we incorporate various losses. We go over the corresponding loss functions and their weight multipliers for every different type of architecture that makes up the suggested model.

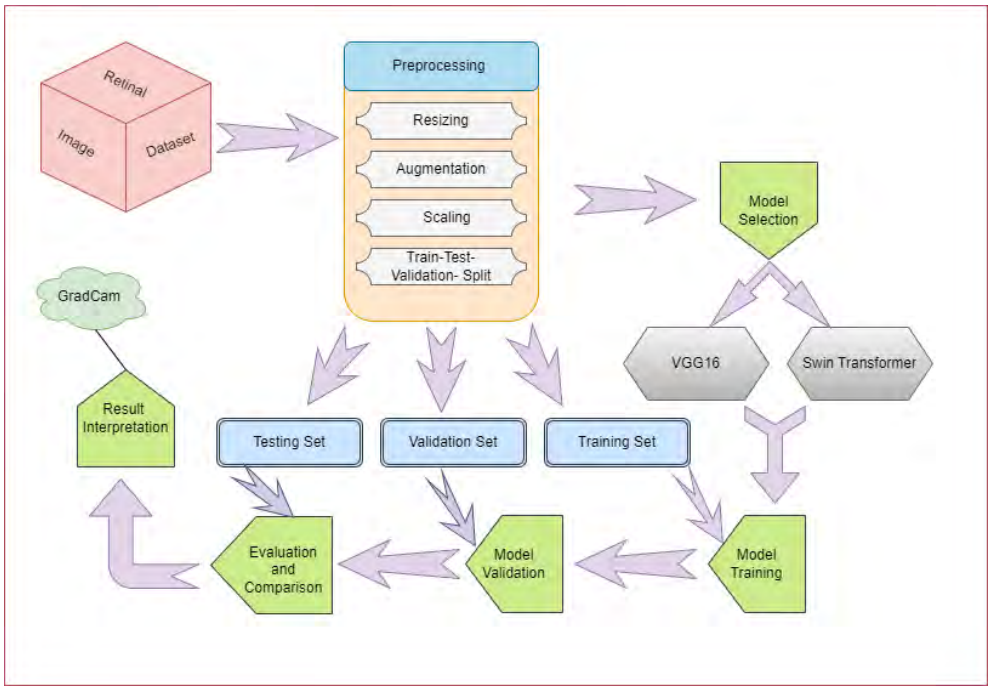


Figure 4.1: Methodology flowchart

# Models

## 4.1 CNN

CNN is not a Vision Transformer model. CNN stands for Convolutional Neural Network and a deep learning algorithm. Deep Learning algorithms are created in a manner that emulates the cerebral cortex's functionality in humans[8]. CNN is a neural network with convolutional (and other) layers. A convolutional layer consists of multiple filters that perform convolutional operations. Image classification, object detection, and image segmentation are just a few of the many computer vision applications that employ this algorithm. CNN's architecture is inspired by the organization of the human brain's visual cortex.A CNN comprises multiple layers, including convolutional layers, pooling layers, and fully connected layers, at a high level. CNNs learn local patterns and characteristics from the input data via convolutional operations and capture essential information via pooling operations. Fully connected layers facilitate the learning of high-level representations and the generation of predictions based on the extracted features. Activation functions induce non-linearity, and during training, the model adjusts its internal parameters to minimize the gap between predicted and actual outputs. CNNs have achieved remarkable success in a variety of computer vision tasks by automatically learning relevant features from unprocessed visual data.

## 4.2 Vision Transformer (ViT)

ViT which stands for Vision Transformer is a model that focuses on image classification that leverages a Transformer-like architecture across the patches of a picture. Fixed-size portions of an image are split up and then linearly embedded. To acquire shapebias-based semantic characteristics that are resilient to texture alterations, position embedding is essential [31]. Following the addition of positional embeddings, the resulting vector sequence is fed into a standard Transformer encoder. Vision Transfer has emerged as the competitive alternative to CNN (Convolutional Neural Network). CNN was used broadly in different image classification tasks until ViT models outperformed CNN by almost 4 times in terms of computation accuracy and efficiency. Patch embedding is identical to a convolution with kernel size and stride equal to patch size [15]. Recently competitive performance in benchmarking for several computer vision implementations,for example, With the use of Vision Transformer, picture classification, object identification, and semantic image segmentation were obtained.

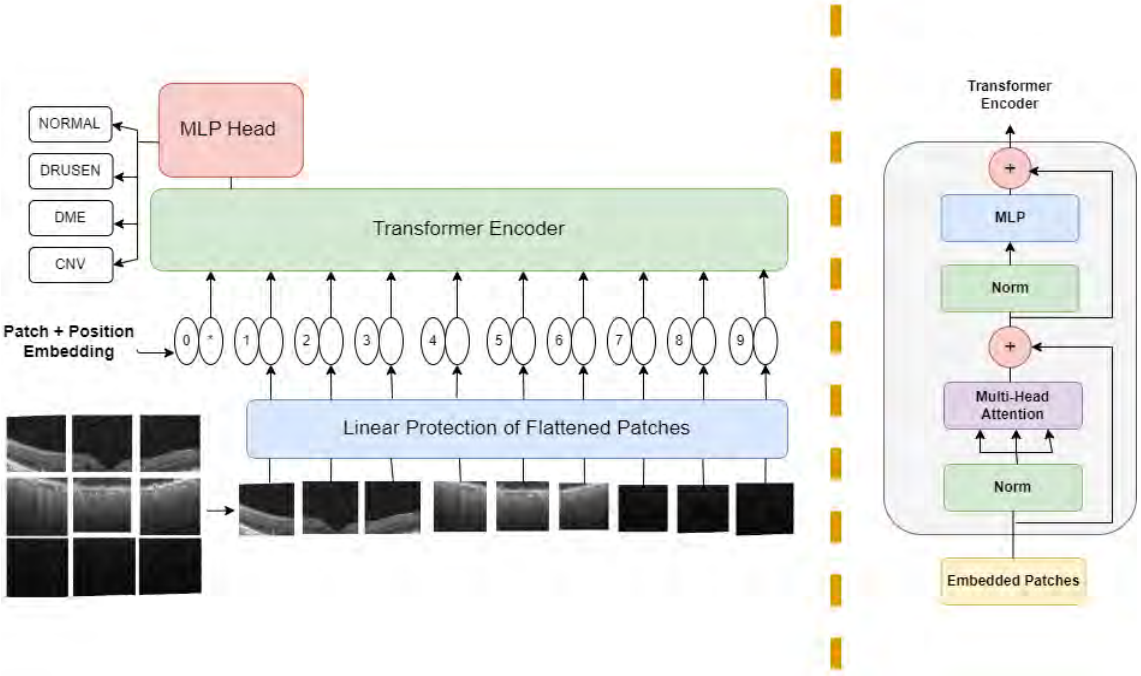


Figure 4.2: Vision Transformer Architechture

## 4.3 GradCAM

GradCAM is a deep neural network approach that highlights important areas in an image to visually understand CNNs and clarify the model's prediction emphasis[7]. It creates a heatmap that indicates important picture regions by utilizing gradients of the projected class in relation to the feature maps of the final convolutional layer. GradCAM is implemented step-by-step and results in a heatmap that reveals important regions that contribute to predictions. It does this by using forward propagation across the network, gradient calculation by backpropagation, global mean pooling for significance the weights, and a weighted mixture of feature mappings[9].

Because of its adaptability to different CNN architectures, users are given more visibility into the model's decision-making process, which improves interpretability and increases confidence in predictions by displaying heatmaps on input pictures. This is a really useful tool for explaining, debugging, and analyzing models, which helps us understand neural networks and their workings better.

## 4.4 Swin Transformers

Vision Transformer's variation is called Swin Transformer. As a result of self-attention processing occurring only within each local window, it creates deep hierarchical maps by joining picture patches in deeper layers and has linear calculation complexity for input image size[18]. It can therefore serve as a general-purpose backbone for applications such as dense recognition and picture categorization. The Swin Transformer is more computationally accurate and efficient than the Vision Transformer. Due to which, it is used as the main element in various vision-based model architectures in recent days. with high resolution images ViT kind of suffers as its computational complexity is quadratic to the image size. But the Swin Transformer can overcome these problems as it can work with high resolution pictures.

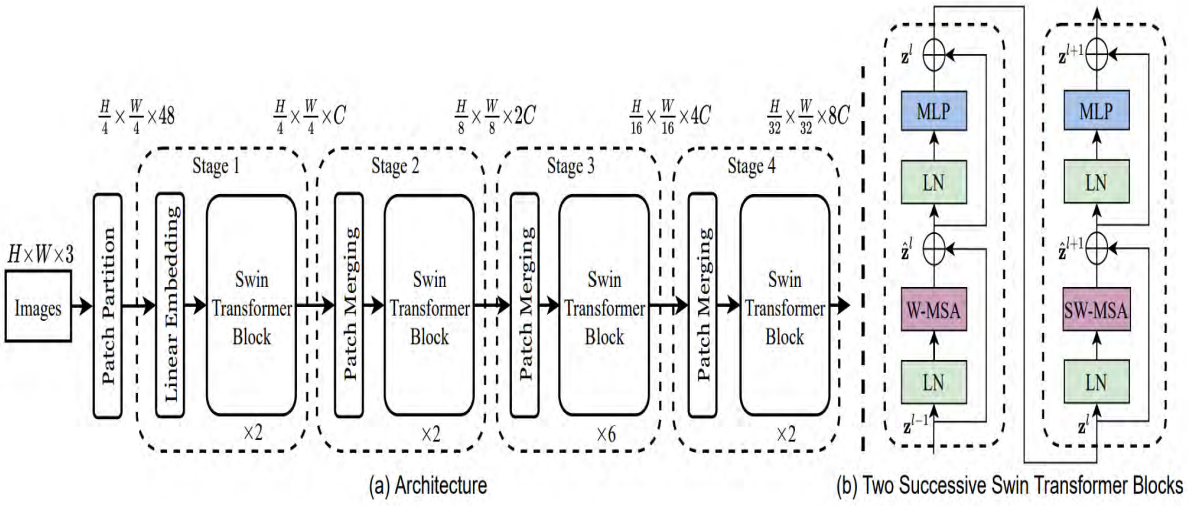


Figure 4.3: Swin Transformer Architechture

## 4.5 The Model based on VGG-16 and Swin Transformer

Convolutional Neural Networks (CNNs) have limitations when it comes to obtaining overall structural information in the field of picture categorization. Specifically, in imbalanced datasets, their propensity to give priority to learning from majority class samples might lead to biased classifications and perhaps miss important insights from minority class samples. Our work presented an improved deep learning technique that combines the advantages of the VGG16 and ViT models to overcome

these drawbacks. This fusion method, named VGG16-ViT, effectively overcomes the limited perception capabilities of traditional CNNs to provide a more comprehensive understanding of global information in bone tumor pictures. Using the unique benefits of both VGG16 and ViT architectures, our model significantly enhanced classification job performance, reduced the effect of data imbalances, and achieved higher classification accuracy. These outcomes highlight the model's potential for improving biomedical image categorization techniques.

**VGG-16 architecture :**

The deep design of VGG16, which consists of three completely linked layers after 16 convolutional layers, is what makes it unique. VGG16 is established as a strong and all-encompassing convolutional neural network with this setup. Through a succession of convolutions, the depth of the architecture allows it to learn complicated hierarchical characteristics from input pictures, improving its capacity to recognize and extract elaborate patterns and representations. More complex feature discovery and representation learning are made possible by this multi-layered structure, which has layered convolutional and dense layers[38]. These layers also greatly enhance the model's ability to understand complex features and hierarchies from the input data.

$$a^{[l]} = g^{[l]}\left(w^{[l]}a^{[l-1]} + b^{[l]}\right)$$

**Swin transformer with SWA or Sliding Window Attention :**

Vision Transformer, or ViT, is a specialized architecture for visual processing that is based on the self-attention mechanism. Its basic architecture consists of a set of Transformer modules, each of which combines feedforward neural networks with multi-head attention processes.

$$h^{[l]} = MultiHeadAttention\left(x^{[l]}\right) + x^{[l]}$$

$$x^{[l]}x^{[l+1]} = LayerNorm\left(h^{[l]}\right) + FFN\left(h^{[l]}\right)$$

$$h^{[l+1]} = MultiHeadAttention\left(x^{[l+1]}\right) + x^{[l+1]}$$

$$y = softmax\left(x^{[l]}\right)$$

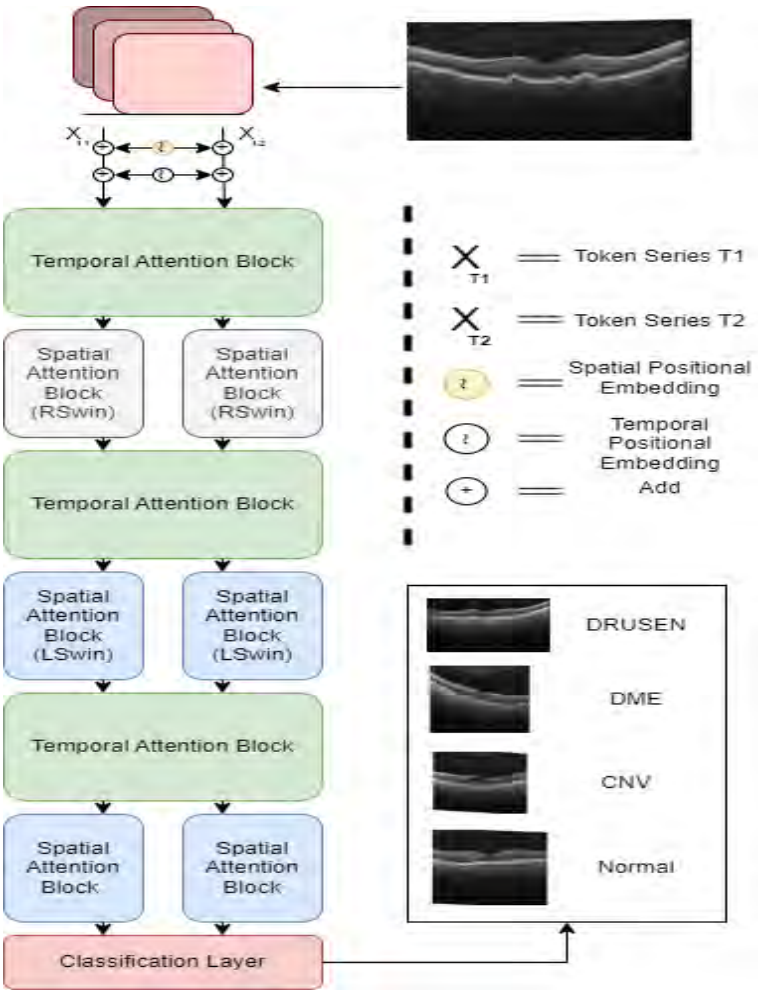


Figure 4.4: Combine two architectures, Vision Transformer and CNN respectively

These Transformer modules are the fundamental building blocks of ViT's architecture, where complex calculations are carried out via multi-head attention techniques. Through the use of these methods, the model is able to extract features comprehensively by processing and distilling information from several parts of the input data concurrently. Furthermore, ViT effectively refines and modifies the retrieved features to provide a more organized representation when combined with feedforward neural networks, facilitating improved comprehension and interpretation of visual input. ViT is a flexible tool for a range of visual identification tasks because of its modular design and use of feedforward networks and multi-head attention to capture complex visual patterns and connections inside pictures. To improve the local longitudinal feature extraction from the axis slices of OCT image information, we have developed an advanced methodology based on fusion of features techniques. Our methodology is based on the deliberate use of an improved sliding-window attention mechanism (SWA). To handle tokens collected from the OCT picture series, this SWA is meticulously built to emphasize their spatial correlations. The fundamental aspect of our SWA is its distinct ability to coordinate an indirect combination of characteristics that are included within a certain window range. This unique characteristic not only makes feature fusion easier, but it also broadens the use of feature fusion in the field of OCT imaging. Correlations between traits that occur at similar spatial distances greatly exceed the significance of those that reside at divergent spatial spans, making this extension especially important.

Our suggested technique places a great deal of weight on the integration of sliding-window attention mechanisms and temporal attention within OCT datasets. These processes serve as sophisticated instruments that improve the model's perceptual skills, enabling it to detect even the smallest changes in local characteristics—a feat that is possible even in the confined context of repetitive training sessions. This enhanced perception acts as a pillar, enabling the model to take advantage of these minute adjustments with more efficacy, which raises the bar for both prediction accuracy and process efficiency.

The whole architecture of the model was built through a series of steps. Firstly, the image data was split into several smaller patches to prepare them for further handling. After this, these patches were made flat for easier processing, and onto each of these flattened patches, information about their location and category labels was added. These modified patches were then fed into the encoder part of the Transformer network. Finally, the results obtained from this stage were sent to the MLP (Multi-Layer Perceptron) module. Here, a process involving weighted summing was carried out to classify and generate the final outcomes.

**Classification assessment factor**

Sensitivity, specificity, and accuracy are the three main quantitative assessment metrics in binary classification. F-value comparison is used to assess the performance of separate VGG16 or ViT networks as well as their combination. The accuracy formula indicates the fraction of accurate categorization based on the accuracy rate of the test set. The specificity formula shows the specificity rate for properly categorized negative samples, whereas the sensitivity formula shows the sensitivity rate for correctly recognized positive samples. Furthermore, the model's ability to learn

is strengthened by the incorporation of transfer learning, as stated in the F1 formula.

$$\textbf{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \times 100\%$$

$$\textbf{Sensitivity} = \frac{TP}{TP+FN} \times 100\%$$

$$\textbf{Specificity} = \frac{TN}{FP+FN} \times 100\%$$

$$\textbf{F1} = \frac{2\times TP}{2\times TP+FP+FN} \times 100\%$$

# Chapter 5

# Result Analysis

## 5.1 Demonstrating performance of the models

Initially we tried to see how a CNN network performs in our dataset to classify the retinal diseases. We used VGG16, the working mechanism of this model is clearly mentioned above in detail. After applying this model we obtained an accuracy of 0.8888 or 88.88%. It indicates that it performs fairly well. We ran our code for 10 epochs for this model. We can see the accuracy and loss graph below.

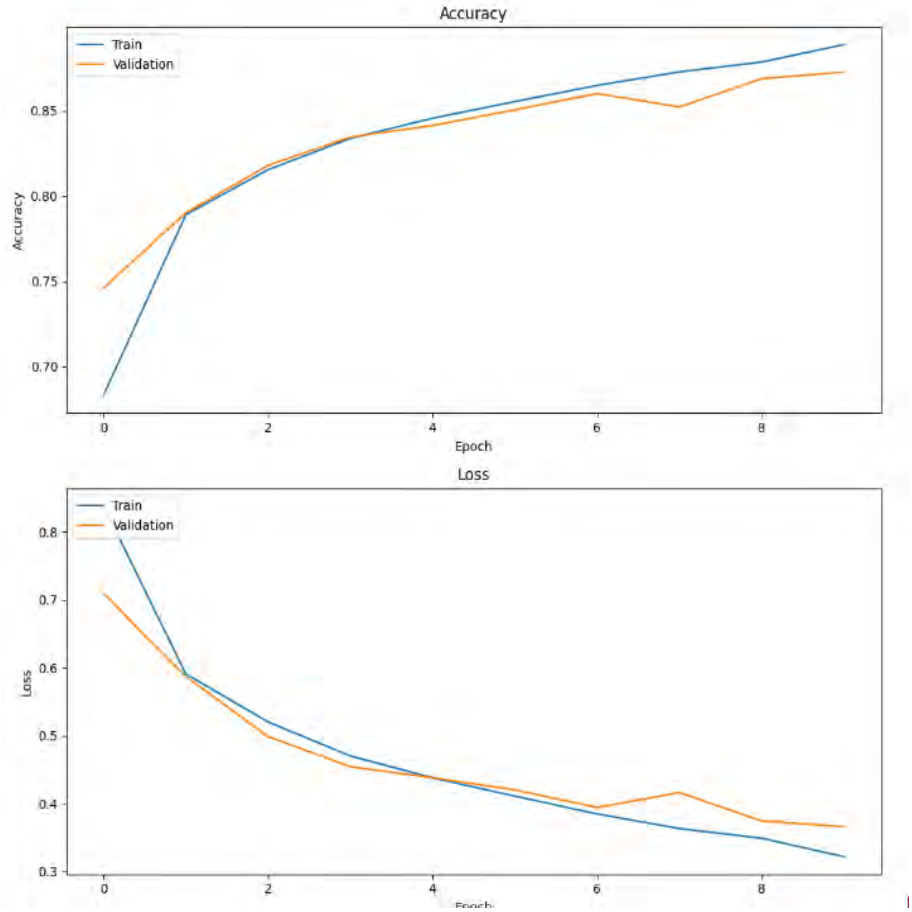


Figure 5.1: Accuracy & Loss gained from CNN(VGG16)

Secondly, we wanted to see how the ViT (Vision Transformer) model works on our

dataset to correctly identify the class. For this experiment we used the base ViT model. Similar to VGG16, we have described this model in detail step by step in the above discussion. For this model also, we ran the code for 10 epochs. This model performed significantly better than the classical CNN model. We gained an accuracy of 0.9139 or 91.39% using ViT base model. From the below accuracy and loss graph we can clearly visualise the performance of the ViT model.

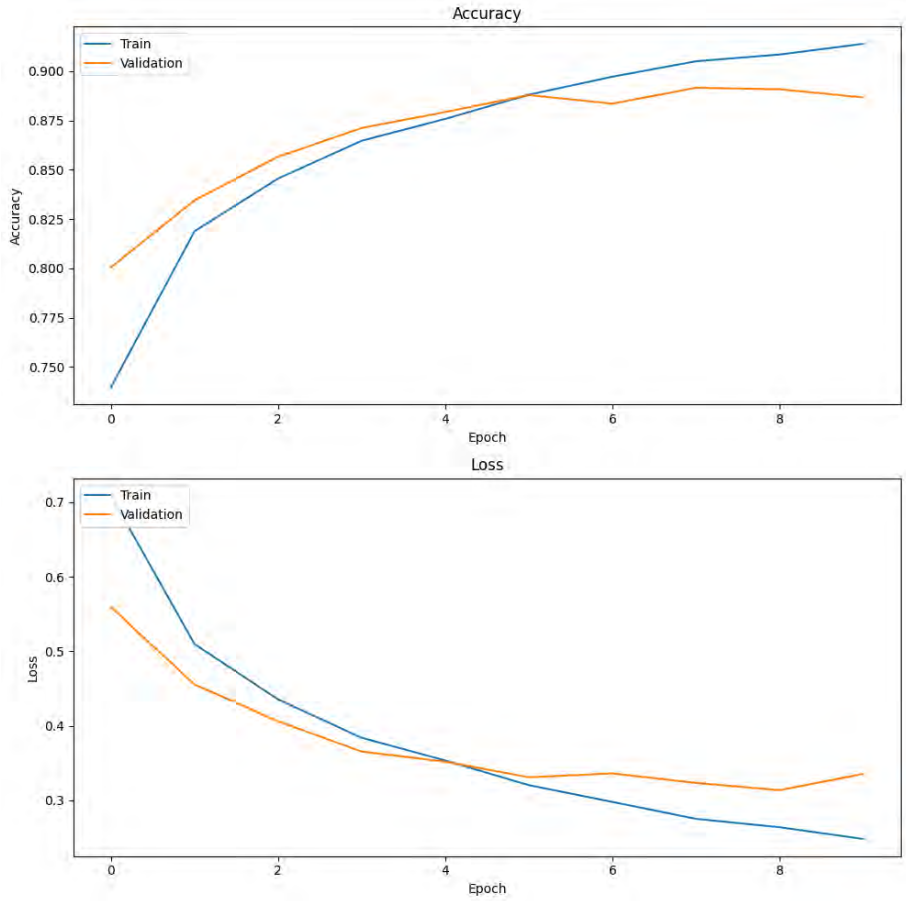


Figure 5.2: Accuracy & Loss gained from ViT(Swin Transformer)

Finally for our proposed model we combined CNN and ViTmodel, and to be specific we focused more on fine tuning in the Swin transformer model. Our proposed model introduces a novel approach by combining the strengths of Convolutional Neural Networks (CNN - VGG16) with the Vision Transformer (ViT) variant Swin Transformer architecture for the classification of retinal diseases—Normal, CNV (Choroidal Neovascularization), DRUSEN, and DME (Diabetic Macular Edema). This custom model harnesses the spatial hierarchies learned by CNNs and the attention mechanisms intrinsic to ViT, synergistically enhancing the overall performance in discerning intricate features within retinal images.

By fusing the feature extraction capabilities of CNNs with the self-attention mechanisms of ViT, our model aims to capitalize on the complementary nature of these two architectures. According to study [25], ViT is more reliable and achieves comparable accuracy to CNN. This amalgamation is designed to capture both local and global contextual information, allowing for a more comprehensive understanding of

the diverse patterns associated with retinal diseases. The integration of CNN and ViT components within our model seeks to exploit the distinctive advantages of each architecture, thereby presenting a robust and effective solution for accurate classification of Normal, CNV, DRUSEN, and DME conditions in retinal imagery.

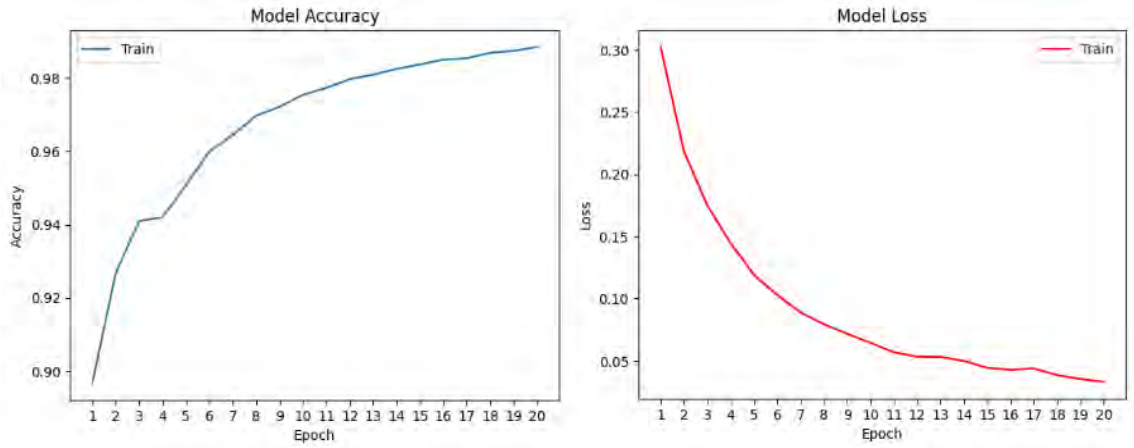


Figure 5.3: Accuracy & Loss gained from our proposed model

## 5.2 Classification report & confusion matrix

The classification report we have got here is like a detailed performance report card for our model that helps identify different eye conditions. These conditions include Choroidal Neovascularization (CNV), Diabetic Macular Edema (DME), Drusen (DRUSEN), and a normal, healthy state. When we look at the report, we see how good our program is at making accurate guesses. For each eye condition, it tells us how often it's right. For example, it's about 96% to 100% right for most conditions. This means our program is doing a really good job at figuring out what's going on in the eye pictures. We also have numbers that show how often our program catches all the cases of each condition. These numbers are quite high, ranging from 95% to 100%. It means our program is really good at not missing any cases. Then, there's something called the F1-Score, which combines both accuracy and completeness. It's like a balance measure, and for our program, it's between 97% and 99%. That's really good. The report also mentions how many examples our program looked at for each condition. The overall accuracy, or how often our model is correct, is at 98%. So, in simple terms, our program is correct about 98 times out of 100. Finally, there are some averages that tell us, on average, how well our program is doing. These averages are also really good, showing that our program is balanced and reliable across all the different eye conditions. In conclusion, this report gives us a good understanding of how well our program is performing, and it's reassuring to see high numbers for accuracy and completeness in identifying different eye conditions.

| Class | Precision (%) | Recall (%) | F1-Score (%) | Support |
|---|---|---|---|---|
| CNV | 96 | 100 | 98 | 242 |
| DME | 100 | 98 | 99 | 242 |
| DRUSEN | 98 | 95 | 97 | 242 |
| NORMAL | 98 | 99 | 98 | 242 |
| **Accuracy** | | | 98.80 | 968 |
| **Macro Avg** | 98 | 98 | 98 | 968 |
| **Weighted Avg** | 98 | 98 | 98 | 968 |

Table 5.1: Classification Report

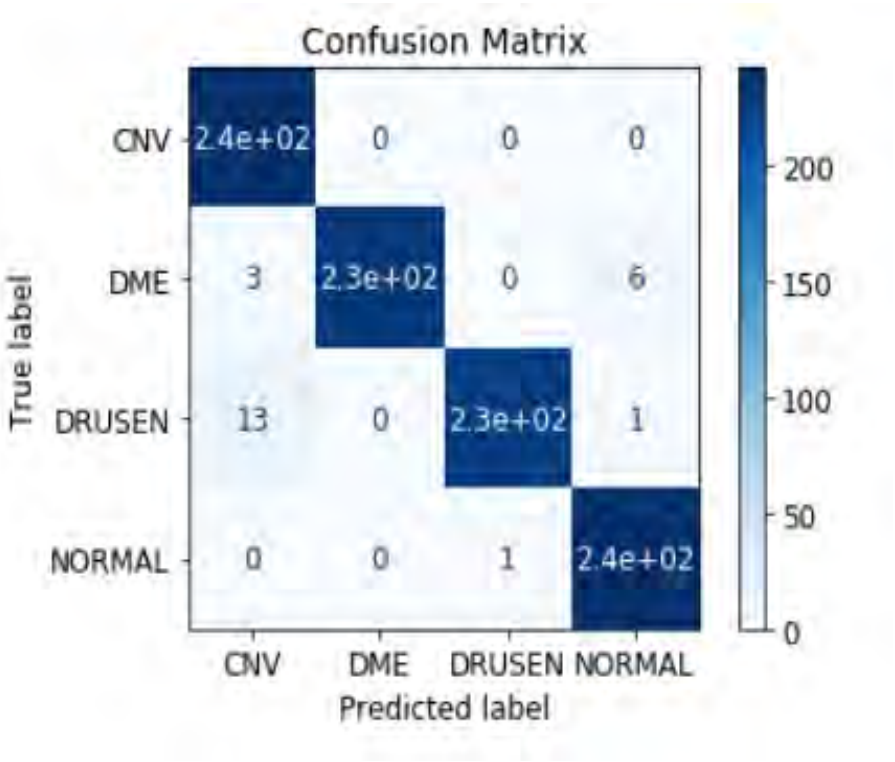


Figure 5.4: Confusion matrix of our proposed model

## 5.3 Comparison of results

This study introduces the custom model that we already described above, designed for the task of image classification. Our primary focus lies in evaluating the performance of ours against established benchmarks set by existing models. The table showcases the accuracy results, where the proposed model demonstrated a remarkable accuracy of 98.8%. This notable achievement positions ours as a state-of-the-art solution, outperforming other well-established models reported in the literature, including Swin-ViT, ResNet, VGG16, VGG19, and VIT. The superior accuracy of our model highlights its effectiveness in image classification, surpassing the capabilities of existing models as reported in prior works. This contribution underscores the potential significance of OURS as an advanced and competitive model in the field.

| Model | Accuracy (%) |
|---|---|
| ResNet [13] | 97.3 |
| VGG19 [17] | 97.8 |
| VIT [37] | 82.0 |
| VGG16 | 88.88 |
| Swin-ViT | 91.39 |
| **Ours** | 98.8 |

Table 5.2: Comparing different models' Accuracy

## 5.4 Grad-CAM technique for visualization

The obtained outcomes reveal the commendable performance of our proposed model in classifying rational images into distinct categories, namely Normal, CNV (Choroidal Neovascularization), DRUSEN, and DME (Diabetic Macular Edema). Consequently, it is imperative to explore the regions within the input images that significantly contribute to the model's efficacy in discerning these diverse classes, leveraging our proposed model. To achieve this objective, we adopt the well-established Grad-CAM methodology, which allows us to scrutinize each model layer and feature map layer. It is a method for increasing the transparency of judgments made by a broad class of Convolutional Neural Network (CNN)-based models by creating "visual explanations" [6]. This in-depth analysis is crucial for comprehending the impact of input values on the model's classification decisions.In a similar vein, previous research efforts, such as the study conducted by Jahmunah et al. [27], utilized Grad-CAM to develop interpretable deep learning models, specifically for detecting myocardial infarction from ECG signals. In our current study, we employ Grad-CAM on the output of our proposed model, generating activation maps that are superimposed onto Fractal Dimension (FD) images corresponding to the Normal, CNV, DRUSEN, and DME classes.

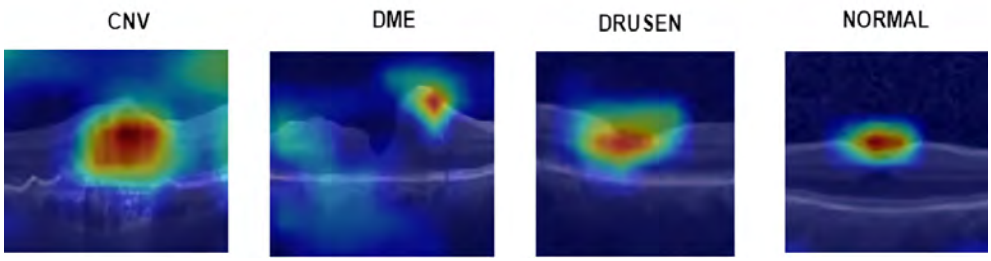


Figure 5.5: Grad-CAM technique for visualization

It is noteworthy that in Grad-CAM images, the color spectrum delineates the relevance of different regions during the classification process, with red signifying the most influential areas and blue indicating less critical regions. As depicted in the resulting figures, our proposed model demonstrates distinct focus areas for each class. For instance, in the case of the CNV class, the model predominantly concentrates on specific FD regions, showcasing high red color density. Similarly, for the DRUSEN class, our proposed model relies heavily on particular FD features, indicated by pronounced red coloration. The DME class, on the other hand, exhibits a

notable emphasis on distinct FD regions for classification decisions. These findings underscore the unique textural structures associated with each class, emphasizing the discriminative power of our proposed model in distinguishing Normal, CNV, DRUSEN, and DME categories based on their fractal characteristics.

# Chapter 6

# Limitations

While our study presents a novel approach to early detection of retinal diseases using a combination of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) with Grad-CAM interpretability, there are several limitations that should be acknowledged:

- Dataset Information: Although we utilized a comprehensive dataset from various sources, the paper does not delve deeply into the detailed characteristics of the dataset. A more extensive exploration of the dataset's specifics, such as its size, diversity, and potential biases, could provide additional context for our findings.
- Imbalance Handling Approach: We addressed class imbalance through random oversampling, a common technique. However, the specific impact of this method on our model's generalization is not thoroughly discussed. Future work could further investigate alternative methods for handling imbalanced datasets.
- Transferability Across Datasets : The proposed model's performance might be different when applied to different retinal disease datasets due to variations in data characteristics as well as disease manifestations. The model may not generalize well to diverse populations or imaging conditions
- Comparative Analysis: While we compare the performance of CNN and ViT components within our model, we do not extensively compare our approach with existing state-of-the-art methods in the domain of retinal disease detection. A more detailed comparative analysis would offer insights into the relative strengths of our proposed approach.

# Chapter 7

# Conclusion

In the conclusion, the combination of VGG16 and Swin transformer using the Grad-CAM for visualisation in diagnosing the retinal diseases offers a promising avenue for enhancing precision in treatment and stabilizing vision. Such improved imagine of relevant retinal regions provides important data for medical practitioners. The synergetic blend of VGG16's feature extraction capabilities and Swin transformer's self-attention mechanism aimed to produce a thorough understanding of intricate aspects seen in retinal pictures, leading to a reliable and effective solution in identifying the eye conditions. The effectiveness of VGG16 and Swin transformer models for identification of retinal diseases surpassed the other models as VGG16 achieved 88.88 percent accuracy after 10 epochs while ViT surpassed it with 91.39 percent and lastly Swin transformer achieved an accuracy of 94.61 percent.. Our proposed model demonstrated a robust performance in classifying Normal, CNV. DRUSEN, and DME conditions , achieving 98 percent accuracy combining VGG16 and swin transformer with a focus on fine tuning and yielding improved accuracy and loss metrics.

However, further study and validation are imperative so that these strategies can be effectively implemented in clinical settings. Thus, further research in this field holds potential for automated screening techniques which will greatly improve retinal disease identification, which would eventually end in better patient results. Hence, ongoing efforts to optimise sensitivity and specificity, together with further research and verifications are important for establishing the efficacy of the VGG16 and Swin transformer with GradCam method for identifying retinal diseases. With addressing the limitations such as dataset bias and size, future work should prioritize acquiring larger and more diverse datasets contrasting the strategy with other state-of-the-art approaches and adding advanced deep learning strategies to boost classification performance. Further research involving a larger patient population will offer important insights about the practical utility of this strategy in clinical settings which ultimately facilitate early detection and precise identification for retinal diseases. In the end, we can conclude by saying that our proposed model demonstrates a potential advancement in retinal disease identification, optimizing the strengths of VGG16 and Swin transformer. Contribution can be done by early detecting and prognosis of retinal diseases as the performance metrics and comprehensive evaluation confirms that it has the potential to be used practically in clinical settings.

# Chapter 8

# Future Work

The combination of VGG16 and Swin Transformer, coupled with Grad-CAM visualization, has shown promising outcomes in the diagnosis of retinal diseases, offering enhanced precision in treatment and vision stabilization. To build upon these achievements, several avenues for future research and improvement can be explored.

**1. Enhanced Interpretability and Visualization Techniques:** Investigate advanced visualization techniques beyond Grad-CAM to further enhance interpretability and provide more detailed insights into the decision-making process of the combined model.
**2. Model Optimization and Hyperparameter Tuning:** Conduct a systematic exploration of hyperparameter space and optimization techniques to fine-tune the performance of the hybrid model. This could involve leveraging automated hyperparameter tuning methods or exploring novel optimization algorithms.
**3. Ensemble Strategies:**Explore the potential benefits of ensemble methods by combining predictions from multiple models, including variants of VGG16 and Swin Transformer. Ensemble techniques can often improve overall performance and robustness.
**4. Domain Adaptation and Generalization:** Assess the adaptability of the proposed model to different retinal datasets and explore strategies for domain adaptation. Investigate how well the model generalizes to diverse patient populations and clinical settings.
**5. Clinical Validation:** Conduct extensive clinical validation studies to assess the real-world effectiveness of the combined model. Collaborate with medical professionals to validate the model's diagnostic accuracy and reliability in actual clinical environments.
**6. Comparison with State-of-the-Art Approaches:**Benchmark the proposed hybrid model against the latest state-of-the-art approaches in retinal disease identification. Evaluate its performance against a diverse set of baseline models and algorithms to establish its competitiveness.

By addressing these future research directions, the proposed model's capabilities can be extended, and its practical utility in clinical settings can be more comprehensively evaluated. Ongoing efforts in these areas will contribute to the advancement of automated screening techniques, ultimately leading to improved patient outcomes in the early detection and precise identification of retinal diseases.

# Bibliography


[1] D. Yorston, "Retinal diseases and vision 2020," *Community Eye Health*, vol. 16, no. 46, pp. 19–20, 2003.

[2] A. D. Kulkarni and B. D. Kuppermann, "Wet age-related macular degeneration," *Advanced drug delivery reviews*, vol. 57, no. 14, pp. 1994–2009, 2005.

[3] H. Nazimul, K. Rohit, and H. Anjli, "Trend of retinal diseases in developing countries," *Expert Review of Ophthalmology*, vol. 3, no. 1, pp. 43–50, 2008.

[4] X. Zhang, H. Zeng, S. Bao, N. Wang, and M. C. Gillies, "Diabetic macular edema: New concepts in patho-physiology and treatment," *Cell & bioscience*, vol. 4, no. 1, pp. 1–14, 2014.

[5] V. Annavarjula, *Computer-vision based retinal image analysis for diagnosis and treatment*, 2017.

[6] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[7] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626. DOI: 10.1109/ICCV.2017.74.

[8] J. Wu, "Introduction to convolutional neural networks," *National Key Lab for Novel Software Technology. Nanjing University. China*, vol. 5, no. 23, p. 495, 2017.

[9] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018, pp. 839–847. DOI: 10.1109/WACV.2018.00097.

[10] C. A. Curcio, "Soft drusen in age-related macular degeneration: Biology and targeting via the oil spill strategies," *Investigative ophthalmology & visual science*, vol. 59, no. 4, AMD160–AMD181, 2018.

[11] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, *et al.*, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature medicine*, vol. 24, no. 9, pp. 1342–1350, 2018.

[12] D. Kermany, *Large dataset of labeled optical coherence tomography (oct) and chest x-ray images*, Jun. 2018. [Online]. Available: https://data.mendeley.com/datasets/rscbjbr9sj?fbclid=IwAR3BJNfLwy4Iu5S0SpiWccgxEIS-Eq3qaQdP_F72-qUtkFcLphJNF18zOsE.

[13] D. Wang and L. Wang, "On oct image classification via deep learning," *IEEE Photonics Journal*, vol. 11, no. 5, pp. 1–14, 2019.

[14] M. Badar, M. Haris, and A. Fatima, "Application of deep learning for retinal image analysis: A review," *Computer Science Review*, vol. 35, p. 100 203, 2020.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[16] S. A. Kamran, K. F. Hossain, A. Tavakkoli, S. L. Zuckerbrod, and S. A. Baker, "Vtgan: Semi-supervised retinal image synthesis and disease prediction using vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3235–3245.

[17] J. Kim and L. Tran, "Retinal disease classification from oct images using deep learning algorithms," in *2021 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, IEEE, 2021, pp. 1–6.

[18] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1833–1844.

[19] J. W. Miller, L. L. D'Anieri, D. Husain, J. B. Miller, and D. G. Vavvas, *Age-related macular degeneration (amd): A view to the future*, 2021.

[20] D. Zhou, B. Kang, X. Jin, *et al.*, "Deepvit: Towards deeper vision transformer," *arXiv preprint arXiv:2103.11886*, 2021.

[21] A. K. Bitto and I. Mahmud, "Multi categorical of common eye disease detect using convolutional neural network: A transfer learning approach," *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 4, pp. 2378–2387, 2022.

[22] G. Cai, Y. Zhu, Y. Wu, X. Jiang, J. Ye, and D. Yang, "A multimodal transformer to fuse images and metadata for skin disease classification," *The Visual Computer*, pp. 1–13, 2022.

[23] J. Cao, K. You, J. Zhou, *et al.*, "A cascade eye diseases screening system with interpretability and expandability in ultra-wide field fundus images: A multicentre diagnostic accuracy study," *EClinicalMedicine*, vol. 53, p. 101 633, 2022.

[24] Y. Chen, X. Gu, Z. Liu, and J. Liang, "A fast inference vision transformer for automatic pavement image classification and its visual interpretation method," *Remote Sensing*, vol. 14, no. 8, p. 1877, 2022.

[25] S. Cuenat and R. Couturier, "Convolutional neural network (cnn) vs vision transformer (vit) for digital holography," in *2022 2nd International Conference on Computer, Control and Robotics (ICCCR)*, IEEE, 2022, pp. 235–240.

[26] M. N. Islam, M. Hasan, M. Hossain, *et al.*, "Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from ct-radiography," *Scientific Reports*, vol. 12, no. 1, pp. 1–14, 2022.

[27] V. Jahmunah, E. Y. K. Ng, R.-S. Tan, S. L. Oh, and U. R. Acharya, "Explainable detection of myocardial infarction using deep learning models with grad-cam technique on ecg signals," *Computers in Biology and Medicine*, vol. 146, p. 105 550, 2022.

[28] Z. Jiang, L. Wang, Q. Wu, *et al.*, "Computer-aided diagnosis of retinopathy based on vision transformer," *Journal of Innovative Optical Health Sciences*, vol. 15, no. 02, p. 2 250 009, 2022.

[29] Y. Ma and W. Lv, "Identification of pneumonia in chest x-ray image based on transformer," *International Journal of Antennas and Propagation*, vol. 2022, 2022.

[30] Z. Ma, Q. Xie, P. Xie, F. Fan, X. Gao, and J. Zhu, "Hctnet: A hybrid convnet-transformer network for retinal optical coherence tomography image classification," *Biosensors*, vol. 12, no. 7, p. 542, 2022.

[31] X. Mao, G. Qi, Y. Chen, *et al.*, "Towards robust vision transformer," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 042–12 051.

[32] S. Park, G. Kim, Y. Oh, *et al.*, "Multi-task vision transformer using low-level chest x-ray feature corpus for covid-19 diagnosis and severity quantification," *Medical Image Analysis*, vol. 75, p. 102 299, 2022.

[33] S. S. M. Sheet, T.-S. Tan, M. As'ari, W. H. W. Hitam, and J. S. Sia, "Retinal disease identification using upgraded clahe filter and transfer convolution neural network," *ICT Express*, vol. 8, no. 1, pp. 142–150, 2022.

[34] M. Subramanian, M. S. Kumar, V. Sathishkumar, *et al.*, "Diagnosis of retinal diseases based on bayesian optimization deep learning network using optical coherence tomography images," *Computational Intelligence and Neuroscience*, vol. 2022, 2022.

[35] B. Wang, D. Zhang, and Z. Tian, "Stcovidnet: Automatic detection model of novel coronavirus pneumonia based on swin transformer," 2022.

[36] M. Wassel, A. M. Hamdi, N. Adly, and M. Torki, "Vision transformers based classification for glaucomatous eye condition," in *2022 26th International Conference on Pattern Recognition (ICPR)*, IEEE, 2022, pp. 5082–5088.

[37] L. Cai, C. Wen, J. Jiang, *et al.*, "Classification of diabetic maculopathy based on optical coherence tomography images using a vision transformer model," *BMJ Open Ophthalmology*, vol. 8, e001423, Dec. 2023. DOI: 10.1136/bmjophth-2023-001423.

[38] W. Chen, M. Ayoub, M. Liao, *et al.*, "A fusion of vgg-16 and vit models for improving bone tumor classification in computed tomography," *Journal of Bone Oncology*, vol. 43, p. 100 508, 2023.

[39] R. Fan, K. Alipour, C. Bowd, *et al.*, "Detecting glaucoma from fundus photographs using deep learning without convolutions: Transformer for improved generalization," *Ophthalmology Science*, vol. 3, no. 1, p. 100 233, 2023.

[40] K. Han, Y. Wang, H. Chen, *et al.*, "A survey on vision transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 87–110, 2023. DOI: 10.1109/TPAMI.2022.3152247.

[41] A. Parvaiz, M. A. Khalid, R. Zafar, H. Ameer, M. Ali, and M. M. Fraz, "Vision transformers in medical computer vision—a contemplative retrospection," *Engineering Applications of Artificial Intelligence*, vol. 122, p. 106 126, 2023.

[42] K. Przystalski, M. Jungiewicz, P. Wawryka, and K. Sabatowski, "Vision transformer in stenosis detection of coronary arteries," *Available at SSRN 4175204*,