Critical Retinal Disease Detection from Optical Coherence Tomography Images by Deep Convolutional Neural Networks and Explainable Machine Learning

by

Pranab Datta
16301177
Saniul Islam
16301020
Retuparna Das
16301205
Mihiran Uddin Zabir
16301198

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

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.
2. The thesis 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 thesis 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:

Pranab Datta
16301177

Md. Saniul Islam
16301020

Retuparna Das
16301205

Mihiran Uddin Zabir
16301198

Approval

The thesis/project titled "Critical Retinal Disease Detection from Optical Coherence Tomography Images by Deep Convolutional Neural Networks and Explainable Machine Learning" submitted by

1. Pranab Datta (16301177)
2. Md. Saniul Islam (16301020)
3. Retuparna Das (16301205)
4. Mihiran Uddin Zabir (16301198)

Of Fall, 2020 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on January 08, 2021.

Examining Committee:

Supervisor:
(Member)

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

Program Coordinator:
(Member)

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

Head of Department:
(Chair)

Mahbubul Alam Majumdar, PhD
Professor and Dean,
School of Data and Sciences
Department of Computer Science and Engineering
Brac University

Acknowledgement

Firstly, we want to thank our Almighty God for always keeping his blessing on us and delivering us to our destination. Many people play vital roles in making us capable of accomplishing the thesis at the appropriate time. We are grateful to them wholeheartedly. In truth, they are blessing in human disguise. They have helped us to be mentally prepared and encouraged us to reach the goals of our life. We convey our best regards to our most respected supervisor and supporting hand, Md. Golam Rabiul Alam, sir, whose contribution in our thesis, compassionate and understanding, helped us do some creatives with innovative ideas and always motivated us to do something new. We also convey best regards to our co-supervisor, Tanzim Reza sir, who did friendly behavior during the research, an instructor for giving good advice. Furthermore, we are grateful for graduating from BRAC University, a renowned university all over Bangladesh, pleased with BRAC University's environment where we understand the actual way of study. We are then like to thank the Department of Computer Science and Engineering for teaching us technical skills and providing us a research lab for the thesis purpose. Finally, we desire to devote our whole thesis to our families for their pure love, endurance during this time. So, we must say that we all are lucky enough to have these support systems to fulfill our thesis thoroughly within time.

Abstract

Retinal disease diagnosis by machine learning can be achieved using Deep Neural Network based predictors. Use of Explainable Artificial Intelligence (XAI) has the potential to explain the black box of those neural network models which are used in identifying critical retinal diseases. Due to lack of explanation in neural networks, Machine Learning based systems are not well trusted in medical field. People still have to rely on the doctor's clarification to come into any conclusion with medical issues. In our proposed model, we have used several Convolutional Neural Network (CNN) models leveraging transfer learning to identify some of the critical retinal diseases such as Choroidal Neovascularization (CNV), Diabetic Macular Edema (DME), DRUSEN from Optical Coherence Tomography (OCT) images along with the explanation. For our classification task, we have used four different CNN models namely which are ResNet 50, inceptionv3, xception, VGG 16. At first, A dataset of several thousand OCT images consisting of four classes: CNV, DME, DRUSEN, and Normal were collected. Afterward, The dataset were pre-processed and applied to our proposed CNN models to classification. We achieved the accuracy gradually 95.20 %, 94.00 %, 96.30 % and 93.30 % by performing the four deep learning model respectively Inception V3, ResNet50, VGG16, and Xception. Eventually, in order to understand the results produced by the black box models, we applied a method of Explainable AI named Layer-wise Propagation (LRP) for a better understanding of retinal disease detection by the CNN models. To add with, the LRP have analysed the models with back propagation and focused on the area of the input image based on the model's training parameters. To sum up, our proposed model has been able to perform critical retinal diseases detection as well as the explanation behind the identification.

Keywords: Machine Learning, Image classification, Artificial Intelligence, Explainable AI, Optical Coherence Tomography, CNN, Inception V3, Resnet50, VGG16, Xception, Black Box, LRP.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Artificial Intelligence has been dominating throughout the world. It has the potential to automate medical diagnosis. AI has been increasingly specific in defining disease diagnosis through the medical images and has been a more feasible source of diagnostic knowledge[2]. The more development in AI will assist more in the detection of diagnosis more effectively in the coming years. AI applications in the area of healthcare may even detect illness and prescribe with the best treatment choices [34]. Artificial intelligence is can be also applied in the field of critical retinal diseases. Among all the retinal diseases there are some critical ones which can hampers our vision very badly. For example, Macular insufficiency is a disease that impacts the section of your eyes called the macula [15]. Age-related Macular Degeneration(AMD)'s prevalent clinical trait is the involvement of DRUSEN [4], which is irregular extracellular accumulation of tissue. The later stage of AMD, is known as Choroidal Neovascularization(CNV). Diabetic Macular Edema(DME) is distinguished by cysts packed with fluids and hard exudates, as well as retinal thickening within the eye[11]. The main causes of macular disorders are cataracts, glaucoma, age-related macular degeneration, diabetic retinopathy, trachoma and different cysts of eye. [11]. The macular lesions are therefore essential to research medical and ophthalmic disorder detection and care.
Optical coherence tomography (OCT) is a three-dimensional imaging retina scanning test which have been used extensively in arts, literature, and research health applications including screening testing ophthalmology[5]. OCT non-invasive retina scanning, tests the acquired images of the formation of the eyes. Optical Coherence Tomography (OCT) is an emerging test technology for diagnostic applications that involves the rapid analysis with the scanning and interpretation of emitted light[6]. The OCT is more developed than the typical eye scanning system as it does not require the dilating drops. In the OCT scanning, patients had to hold their eyes to attach with the light to the inner side of the machine to detect the disease[7]. OCT can come up with early detection for various types of diseases because it is likely to be able to identify structural features that will appear before the onset of behavioral symptoms of retinal diseases[8]. OCT is a process of machine aided testing for the accurate diagnosis of eye diseases. .
Deep Neural Networks (DNNs) are becoming a reality in recent years as this procedure can be composed of many different fields, although they are administered mainly as black-box predictors[19]. Since the actions of artificial intelligence systems is important in innumerable circumstances and analysis of the neural network

has been established as an important active line of research. Though advances have been made in deep learning for OCT, such applications for ophthalmology include quantifying intra-retinal fluid and diagnosing eye disorder[6]. Convolutionary neural network is one of the standard deep learning models that can be generated automatically to train datasets for testing from a very large training dataset.[13].
In this paper, we have evaluated several models for image classification such as vgg16, resnet50, xception, inception and then applied explainable artificial intelligence on vgg16 and resnet50 by a OCT image dataset. Firstly, we arranged the dataset into 3 parts which are training images, validation images and testing images. Then, we have applied some preprocessing on the training images so that the images are generalized. After that, with the help of the compiler we have trained our data by the models one by one. We have got some different accuracy results on the models. Finally, we have implemented Explainable Artificial Intelligence (XAI) by Layer Wise Propagation (LRP) with the help of Deep Taylor Decomposition.
As we are using deep neural networks to simulate these factors, we don't get to know what is happening inside [19].We are using machine learning for creating the predictions by giving inputs and attempts to make decisions without having to explain the main cause. The machine reaches a decision or takes any measure although as a consequence of doing analysis, we do not know whether or how it has achieved the outcome[22]. The machine has taken more control over certain things. The region of AI (known as XAI or interpretable AI) tends to draw experts in the field.Though many recent works have been done to classify these diseases using deep neural networks there's hardly some of the work which describes those black boxes as deep neural networks are basically black box predictors. In this paper, we used Deep Taylor Decomposition(DTD) for LRP as explainable Ai to describe those predictors. In recent years, generalized linear networks have been influenced by the multiple approaches made toward the accuracy of machine learning; among them, LRP has got its mathematical foundation in Deep Taylor Decomposition (DTD) theorem.

1.1 Motivation

Our main motive is to increase the acceptability of machine learning specially in medical field. In the past, when anyone wanted to know more about healthcare, they would have to go ask a doctor. We explore various papers which are related to image classification and by doing these we approached to do the image classification by using deep learning models. By reading several research papers we determined to work with the Explainable Ai to remove the black box of neural network models for identifying retinal diseases. Our main aim is to create doctor's reliability in machine learning prediction. It will be very much effective for the doctors as well as the patients. To reduce diagnostic error, we need to persuade doctors how machine learning predicts more accurately in terms of medical dataset. Finally, we took a decision to work with image classification with the deep neural network and the explainable artificial intelligence for optical coherence tomography for fulfilling our purpose.

1.2 Problem Statement

Around 4.2 million people who are 40 years older are either who have poor eyesight or with the exception of those who have already been qualified as blind[53]. This Scenario is the same all over the world. Errors of diagnosis are associated with long lasting harm to individual patients that could lead even to deaths. A stats by John Hopkins states more than 250000 people die in the United states due to medical error while others claim the rate to be 44400 even deaths per year [23]. Not even in the United states, the graph is also even harsh for the third world countries[23]. Errors in diagnosis is not only leading people to death but also it is making permanent disability to the people. For identification, People have to depend on the doctor's decision over any medical problem. Though there are some implementations over various problems by machine learning, these implementations are not well reliable in medical fields. We have tried to implement such a system which can provide more reliability on machine learning systems by remove black box from the system. To overcome those problems, our proposed model can classify the retinal disease with the clarify of the prediction over that disease. It eases the doctor's life as well as the patient.

1.3 Research Objectives

In today's era most of the people face retinal disease. Our main aim is to classify retinal diseases such as CNV, DME, DRUSEN and NORMAL by using deep neural networks and explaining these neural networks by Explainable AI. The main motto of image classification is categorizing image into a specific class. As a result, it could be able to predict the images and also be capable of presenting the disease name. Beside that, we aimed to classify retinal disease with better accuracy than human experts' opinion on OCT images. As machines will predict the disease so it can assure more certainty. Moreover, explaining the reasons behind the prediction is the main motive. The main research objective are written below

- To identify critical retinal diseases.
- To classify the retinal diseases by using the deep neural network.
- To reduce the diagnosis error.
- To gain trustworthiness of a doctor towards prediction by using machine learning.
- To enlarge the admissibility of machine learning.

1.4 Research Methodology

The main objective of our research is to add more transparency in understanding the black box of neural networks. Nowadays, people are depending on machine learning for many purposes. Image classification is one of the most efficient contributions in machine learning through deep neural networks[13]. Here we came up with the idea of classifying different diseases in the retina of our eyes and explaining the reason

behind the predictions of our model. We have tried to differentiate some eye diseases such as CNV, DME, DRUSEN through our trained model. We have collected our dataset from Mendeley Data which was taken from various hospitals in California and Beijing. We have implemented an ensemble CNN classifier to identify the most usual retinal diseases. We tested multiple machine learning approaches such as VGG16, ResNet50, Xception, Inception, and then applied Explainable Artificial Intelligence to VGG16 and ResNet50 using the OCT dataset. First, we sorted the image dataset into three sets: preparation, validation, and testing. Then we've preprocessed the training photos, so they're as generalized as possible. After that, with the aid of the compiler, we successfully trained the models one by one. We have run the models, and the findings vary. Finally, we have used Explainable Artificial Intelligence with the aid of Deep Taylor Decomposition with LayerWise Propagation.

1.5 Thesis Orientation

This section of the paper is described in the order of the total tasks of our thesis. Here, in chapter 2, includes the background study which helps to get a clear understanding of our research motive. Chapter 3 includes the related work formed on our proposed model. Chapter 4 states the methodology along with that our datasets description are also analyzed for the entire proposed model. Chapter 5 encompasses the results and analysis of our proposed model. Finally, Chapter 6 ends with the plan and summarizes the entire proposed model of our system.

Chapter 2

Background Analysis

2.1 Retina

The retina, which is a part of the eye that sends sensory signals to the brain. The retina was created from embryological neural tissue in the infancy . [10]. Retina is a complex, translucent, multi-layered tissue with just one photoreceptor cell that is light-sensitive. The light that arrives from both directions to reach the photoreceptors has to travel through the overlying layers of the body which are of two forms, rods and cones[9]. The rods and cones are separate with having unique shapes and their exposure to different kinds of light. Rods operate at very low light levels [9]. Rod can only be activated by a few bits of light (photons) which can only be worked as night vision. Rods don't support the color vision, but we see it all in a grayscale at night[10]. Eyes are more evolved in animals who are involved during the day and important for vision. The more distinct cones in the eye, the greater the detail that an individual can discern. Rods are spread across the retina, but cones appear to cluster in two areas such as the fovea centers at, a pit at the rear of the retina, which has the largest number of cones in the body, and the surrounding macula lutea, a circular patch of yellow - pigmented tissue around 5 or 6 millimeters in diameter[39]. When the light passes through the eye, the cornea, and the lens, it is scattered and refracted to settle on the retina.[10][9]. The light-sensing molecules found in the rods and cones in the retina. Crucial cells above and within the retinal cell layers organize them for transferred them to the visual cortex, where it is further arranged and interpreted.[10].

2.2 Retinal Diseases

Retinal disease is a very serious problem which occurs due to several reasons. Due to retinal disease many serious parts of the retina can be hampered. Even many people lose vision or become blind due to retinal disease. It can take place in any retina layer such as pigment epithelium layer, rod and cone layer, inner nuclear layer etc. People who have diabetic for a long time have had their retina hampered due to a lot of high level sugar[11]. When the eye disease arose then the retina stopped the blood supply of the eye. There are innumerable cause of retina such as diabetic retinopathy, age-related macular edema, glaucoma, cataract, choroidal neovascularization (CVD), Diabetic macular Edema, DRUSEN, amblyopia, retinal

tears etc[11]. In our proposed model we have worked with 3 types of eye disease which are very dangerous for humans.

2.2.1 CNV

CNV which is the choroidal neovascularization is a related to a disease of subretinal part of eyes due to increase of new blood vessels. There are various reasons behind the CNV such as myopia, macular telangiectasia, multi focal choroiditis, choroidal Rupture etc[16]. The CNV is mainly the process of accumulate new blood vessels in the the subretina. One of the main reasons of CNV is the age related macular degeneration. Beside, the center of eye become larger as well as it hamper a lot of retina. Furthermore, the fungal infection of the eye is also the reason for CNV[14]. The CNV can be a dangerous reason for the partial or permanent vision of loss.

2.2.2 DME

DME stands for Diabetic Macular Edema[20]. Diabetic Macular Edema is mainly eye disease which occurs due to leakage of blood vessels fluid. While becoming the leakage of the blood vessels extra cholesterols and fats lipid grow in that part of the retina. DME can easily weaken the blood vessels in our eyes. As a result, it grows out of control in the retina. This diabetic vision is called retinopathy[41]. This is a pre stage of DME. In most of the DME do not show any symptoms. By doing eye tests it can be identified[41].

2.2.3 DRUSEN

DRUSEN is an eye disease which is the small yellow fatty protein that aggregate beneath the retina. It can also be white inflation of extracellular substances that grow up in the brooks membrane and in the beneath of retina pigment epithelium of the eye[4]. Due to DRUSEN, extra fats are accumulated in the retina. It is not as risky for the people under the age of 40. However, after the age of 40, it becomes adverse for the retina as it can be also the cause of AMD[29]. The DRUSEN has two types such as soft drusen and hard drusen. Soft DRUSEN are very enlarged as they stay together in the retina.It is very dangerous for humans. These types are not very clear to identify.It has a lot of changes to occur the AMD (Age-related Macular Degeneration). Hard DRUSENs are very tiny in size. They are not as risky as soft DRUSEN. However, it is also harmful for the retina[4]. Hard DRUSENs generally do not show any large symptoms in the retina. But by doing eye tests DRUSEN can be seen. In the Optical coherence tomography images there are some dome shaped lumps that can be seen under the retina. These are mainly called the DRUSEN. Due to DRUSEN people don't become completely blind, however they can lose their partial vision[8].

2.3 Optical Coherence Tomography(OCT)

Optical Coherence Tomography(OCT) is an analytical image scanning method that simulates the curve observation of various disconnected transmitter of biological living cells which can be acquired via the retinal scanning test[5]. At the present

time, OCT is broadly seen in the clinical diagnosis. It is beneficial for the forecast diagnosis and intervention of patients to decrease the amount of people suffering to related retinal disease. By using OCT, the automatic images classification has become an emerging research field which has brought a lot of improvement for the humans along with the doctors[34]. In the retina scanning procedure a high resolution imaging method is used in the ophthalmology. In this test, we wanted to determine computer vision mechanisms to make the differentiate with the innumerable retinal disease from the OCT images. As the enhancement is being developed in deep neural networks has a significant impact on the image classification through using the oct scanning system[2]. In the optical coherence system, patients don't have to face a lot of struggle. They only have to hold their eyes to do the attachment with the light towards the inner side of the machine. In our proposed model, we utilize an OCT image dataset to identify retinal disease.

2.4 Image classification

Image classification is supervised learning systems where a group of selected classes are given to a machine as well as it also trains any specific model to identify the defined images. In the image classification machines can classify an image[1]. Image processing can also provide the probability of any specific image of being particular class. In early ages, image classification was dependent on only raw pixels of data. Machine can divide an image into a single image[1]. The main purposes of image classification is to differentiate various classes according to their features.[1]. There are two types of image classification one is supervised classification and the other one is unsupervised classification. In the supervised classification, human have to train the images and guided them as one's own specifications.Our proposed model is a supervised classification. It is mainly a human conducted procedure. On the other hand, unsupervised classification is the process of classifying each datasets as belonging to one specific class in that group[5]. It is mainly a machine automated procedure where according to the similarities machine assign them as a same group.

2.5 Different Methodology for image classification

There are many artificial neural network models for image classification such as Resnet, Xception, VGG16, VGG19, Inception V2, Inception V3, DenseNet, MobileNet, NasNet etc. These models are generally used by Keras in image classification.

2.5.1 Convolutional Neural Network

CNN is a neural network that has so far been the most popularly used for analyzing images. Although image analysis has been the most widespread use of CNN, it can be thought of CNN as an Artificial neural network that has some type of specialization for being able to detect patterns and make sense of them as well[18]. This pattern detection is what differentiates it from a standard multi layer perceptron (MLP). Through CNN a wide range of computer vision and image processing applications, image recognition ,image segmentation including object detection has

been successfully applied with the typical convolutional (Conv) layers, nonlinearity, pooling layers, and completely connected (FC) layer[21]. The most significant part of CNN is the convolutional layer. For each layer in convolutional function, two-dimensional (2D) kernels are taught to express spatial connectivity with the previous layer to get a collection of features[18].

Architecture of CNN

CNN can be described in two phases . There could be a training phase and inference phase.

- **Convolutional layers:** It takes filters which will be applied to the given input image for creating different activation features in that input image[31].
  (3D) CNV:

  1. input = width × height × depth
  2. kernels/filters= 3×3 ,1×1,1×7 (so on) × depth
  3. no. of kernels = N
  4. Output = width × height × no of kernels

- **Non Linear Activation Layer:** There are several ways of doing that, one of the ways is Rayleigh or there is a leaky value, there could be sigmoid, there could be tangent and what it is supposed to do is to bring the nonlinearity to the network[31].

- **Polling:** Pooling is also a non linear layer but its mainly used to down sample. As CNN is known for computing large and memory problems, with the help of pooling it reduces the complexity of computation and the complexity of memory hence it is quite handful[31].

- **Fully connected layer:** The goal of a fully connected layer to categorize and to detect the final output categories and that's why it is in the output stages. Every node at the input is connected to its coefficients in this layer. So, it is very data heavy all the coefficients are getting loaded and what it will do is after dealing all of them, it will create these bunch of output datas and it can be picked by probability distribution algorithms like soft max or support vector machine (SVM)[31].

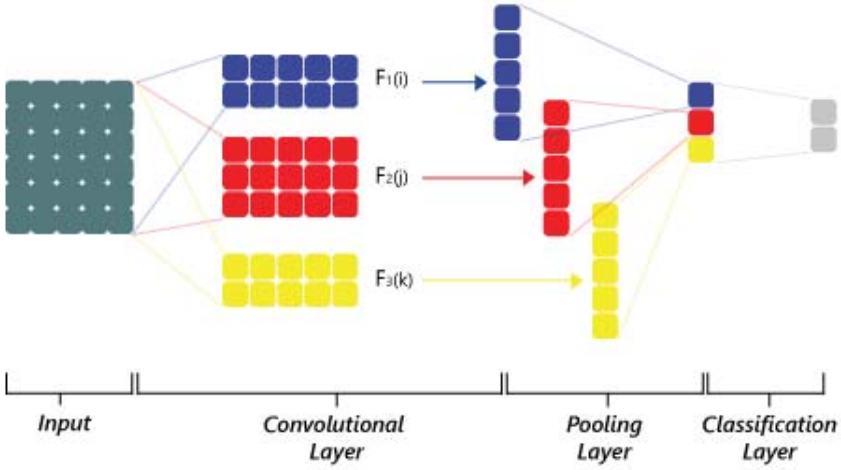

Figure 2.1: Architecture of Convolutional Neural Network (CNN)[54]

This figure 2.1 is showing that convolution layer of CNN which can generate feature map and pooling layer can decrease the parameter size.

Advantages of CNN

- It can share weights .One can have one layered CNN with 4 filters of size 2×2
- In terms of performance it outperforms other NNS on image recognition tasks
- CNN is also known for being very good at feature extracting. It can extract useful attributes from an already trained CNN with train weights

2.5.2 ResNet 50

Residual Networks (ResNet) is a part of Deep neural networks. However, we don't have to face any gradient problem in ResNet 50. Residual Network first develops the skip connection concept which means to join the input layers with the output of the convolution block[35]. For the skip connection ResNet can dissolve the gradient problem.

Architecture of ResNet

The ResNet50 model is made of 5 stages. Each layer has its individual convolution and identity block[35]. These identity and convolution blocks also have 3 layers[35]. Like other models, the ResNet model also has 23 million trainable parameters. ResNet50 is a convolution layer which has a kernel size of 7×7 as well as 64 Kernels along with the size of stride is 2 and this provides 1 layer[30]. In the ResNet architecture there have 3×3 max pool with the stride size of 2. In the second convolution layer, there are three time repeating of kernel size (1×164, 3×364, 1×1256). In the third convolution layer, there are 4 time of repetition of kernel(1×1128,3×3128, and 1×152)[30]. In this step, the 3rd convolution layer is composed of 12 layers. In the fourth convolution layer, there are 6 times repeating of 1x1256 along with two more kernels of 3×3256 and followed by 1×1,1024. In this

stage the ResNet 50 assembled with 18 layers. Finally, in the last convolution layers there consist of 3 times of repeating of (1×1512, 3×3512, 1x12048) kernels which create 9 layers. After completing these five layers, the ResNet model starts with an average pool with 1000 nodes. These 5 steps finish in soft max function by providing 1 layer. Consequently, the ResNet 50 model came up with (1+9+12+18+9+1)= 50 layers[30].

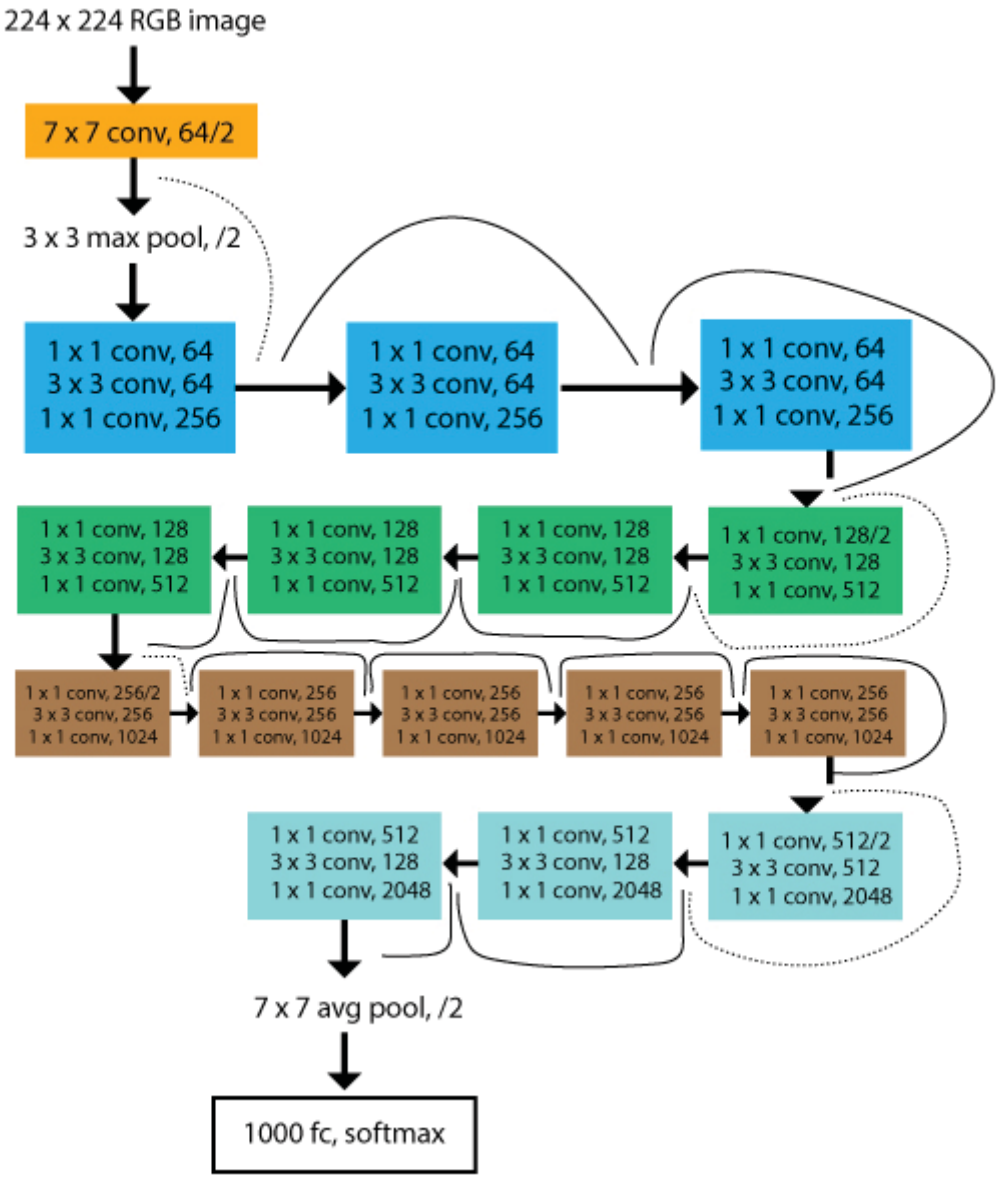

Figure 2.2: Architecture of Architecture of ResNet50

This following figure 2.2 is showing that resNet50 architecture where the initial input size is 224×224.In these five stages of ResNet50, working with the maxpool layer, average pool layer along with 1000 connected nodes.

Blocks of ResNet50

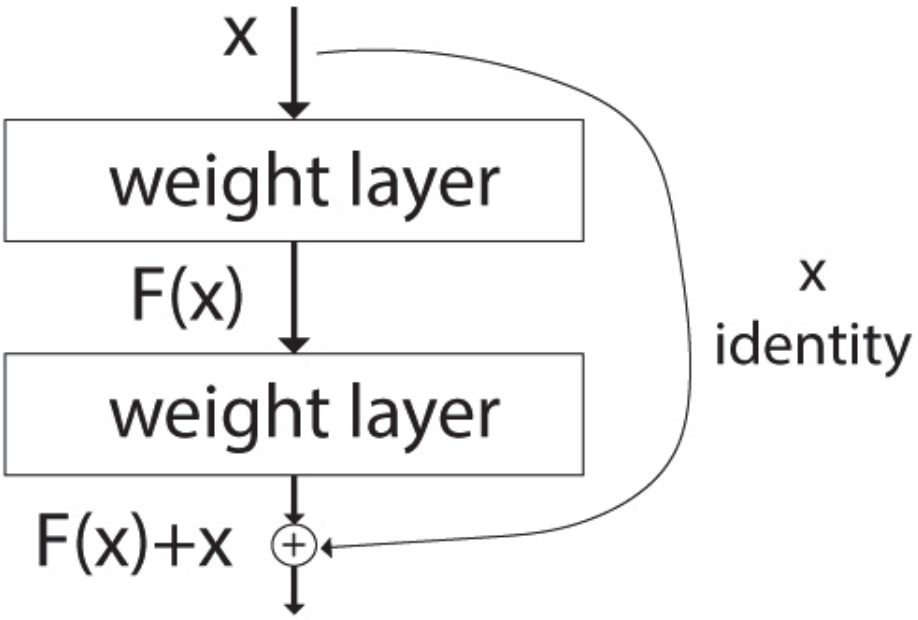

Figure 2.3: Blocks of ResNet50

In Figure 2.3, F(x) is a loss function,X is the starting layer and Y is the output layer. To make the input size with the output size,the initial layer skips some of the layers.

The ResNet50 follow this equation:

$$Y = F(x) + X \tag{2.1}$$

The loss function F(x) should be zero. For this reason, the initial input layer skipped other layers and finally we added the input layer with the loss function. In the residual network our main goal is to make equal value of input and output image size. That is why the loss function F(x) should be zero. Thus, the accuracy will also increase as the input and output value will be the same image size. There are two blocks of the ResNet50 model. One is an identity block and the other one is a Convolution block[50].

- **Identity block:** In the identity block the input size image is equal to the output size image. Here we do not need a have to a convolution layer. As in this block achieved our goal of ResNet50 as input image size is as same size of output size. We do not need to create an alternative path for getting the desired result. All the layer follow skip connection excluding the starting input.

- **Convolution block:** In the convolution layer the output size image follow the two procedures.

    1. **By padding the input volume:** Padding is such a way of putting zero into the input image due to the match according to the input size with output image size input. Padding can be two types such as valid padding and same padding. Padding can increase more accuracy of image analysis.
    2. **By perform 1×1 convolutions:** We use maxpool or average pool to reduce image size. A 1x1 convolution is a zero-dimensional operation that maps a one-dimensional input image to a one-dimensional output image. It is mostly used to decrease the variety of complexity channels, because it is inefficient to multiply the volumes with volumes of depths which are very large.

Advantages of ResNet model

- The ResNet model is a faster approach as it has less parameters as well as less operations
- It performs better than other models as the residual function modifies it's initial input layer to become the identical output. For this reason, the ResNet model succeeds the skip connection to get the accurate mapping with the appropriate value.

2.5.3 Inception V3

Inception V3 has 3 blocks such as block A, block B, block C. Each block has a convolution layer. The 3 convolution layers start parallelly in the model. In inceptionV3 our main aim is to reduce parameter as well as less operations[43]. In the each block of inception V3, have to apply three steps to get more accurate prediction.The first one is factorization into smaller convolution layer, second one is factorization into asymmetric convolution and last one is factorization into 7×7 convolution.In the first procedure factoring into smaller convolution,we split the convolution into smaller convolution layer to get less parameters[43]. For example,we are replacing 5×5 convolution in inception3,instead of that we are using 2 layers of 3×3 convolution layer. In the 5×5 convolution layer have (5×5) =25 parameters on the other hand, in the 2 layer (3×3)convolution layer have (3×3+3×3)= 18 parameters. So, here number of the parameter is reduce by 28. So, we are getting more accuracy through inception V3 by decreasing parameters. In the inception V3, have to do asymmetric convolution to do factorization also to get less parameters. After doing these have to concate all the filters of inceptionV3 model. Thus, inceptionV3 predicts more exact output[32].

Layer Architecture of Inception v3

There are 48 layers in a network in inception V3. Firstly,it receives the input size

of 299×299×3. There are three convolutional layers in inception V3[33]. After giving the input the first convolutional layer executes with 32 filters along with that stride 2 and also the filter size will be 3×3. In the second convolution layer, it has the same filter size, with stride 1[33]. In the third convolution layer,the filter will be 64 with stride 1. After that, by doing maxpool layer in inception V3, the window size become 3 × 3 and stride become 2. [33]After that getting the image size which is 73×73×64. Then, doing the convolution layer with 80 filters with stride 1. Afterwards, the convolutional layer the input size become 73×73×80.After that, applying a maxpool layer with a window size 3×3 and stride 2[33]. Then,have to run the inception block A with 3 times, inception block B with 4 times as well as inception block C 2 times. There are two dense layers and there are fully connected layers. Lastly, in the final layer of inceptionv3 consist of activation classifier, softmax classifier and the auxiliary classifier. The auxiliary classifier is used mainly to remove the gradient.

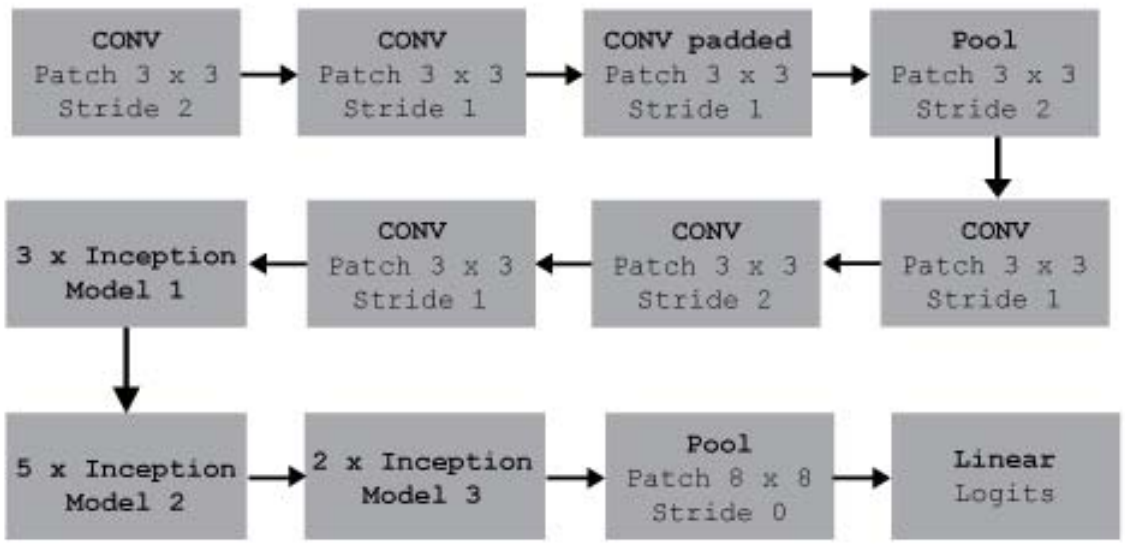

Figure 2.4: Architecture of Inception v3

This following 2.4 figure is showing the convolution layer with the different stride size in each layer.

Advantage of Inception V3

- The Inception model's main goal is to decrease the parameters so it gives more accurate prediction.
- It has less operations and functions.
- Inception V3 have 1×1 convolution layer which can reduce the dimension of the features

2.5.4 Xception

The end-to-end (e2e) pre-training approach named Xception is built to completely exploit deep convolutional neural networks[40]. It was developed by software engineers at Google. In 1984, Google implemented a concept known as Inception. The model generated an appreciation of convolutional neural networks just the same way as a halfway progress in the center of convolution and the depth wise distinct convolution operation (a depth wise convolution followed by a point wise convolution[40]. In the context of this, a depth wise divisible convolution may be viewed as an Inception module with a maximally gigantic number of pinnacles. This concept led the individuals to suggest a novel profound convolutional neural organization engineering propelled by Inception, where Inception modules have been supplanted with depth wise distinguishable convolutional layers The Convolutional Neural Network utilizes a term utilized by convolutional neural networks named that of an organisation[26].

As a consequence, it seems plausible that the preparation of interactions among channels, or spatial relationships among the elements of a convolutional neural net, may all be done without the intervention of the elemental channels[26]. Since there are more grounded variations of the fundamental concept of the Inception Principle, this latest paradigm is labeled "Outrageous Inception". (A technical term for plausible inference, since the theory is "Outrageous Inception".) The acts of the company as the train falls hurtling off the track are distinctly illustrated in the drawing that follows. The Xception architecture has 36 layers of convolutional layers that form the part extraction foundation[33].

There in the organization our fixation test employed an entirely visual experimental end result for measurement, and the test-run afterward connected on a convolutional architecture with a measured predicted result. Alternatively it should render a measurement layer totally used to each layer of relatedness, then make a decision on every given problem in the assessment section. Inside the 36 convolutional layers, there are 14 modules, each one with its own memory[3]. These memory modules are aligned strongly with what's following it, which alternates with position in the next module.

Based on their interpretation, the Xception model is a pile of divisible convolution layers, where each layer works on a convolutional background of less than one third associates[26]. This makes the engineering a lot easier to characterize.

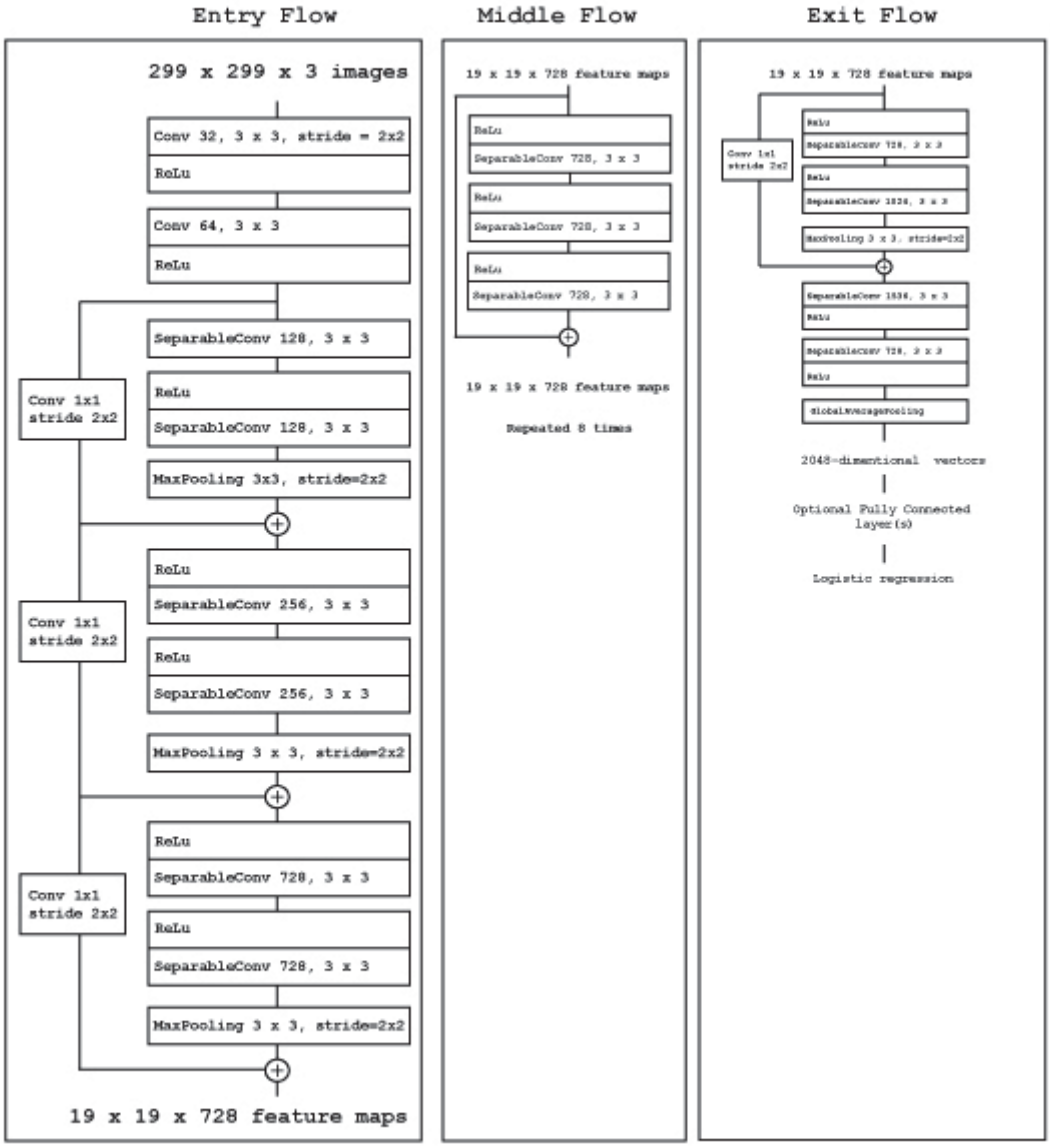

Figure 2.5: Architecture of Xception

The figure 2.5 is Xception model which input size is 299 X 299. Here,is also demonstrating the ReLu activation function with the different function with different convolution layers with the different stride size.

2.5.5 VGG16

A multilayer neural network model called VGG model proposed by K. Simeon and Abraham[44]. LUCC can achieve 87.5 percent in TOP-5 image recognition[44]. It was one of the high-quality models to compete in the ILSVRC-2014 contest. We can say that VGG16 is superior to AlexNet because it superimposes three filters on the previous layer, each of which has the size of 3×3 kernel[44].

Architecture Of VGG16

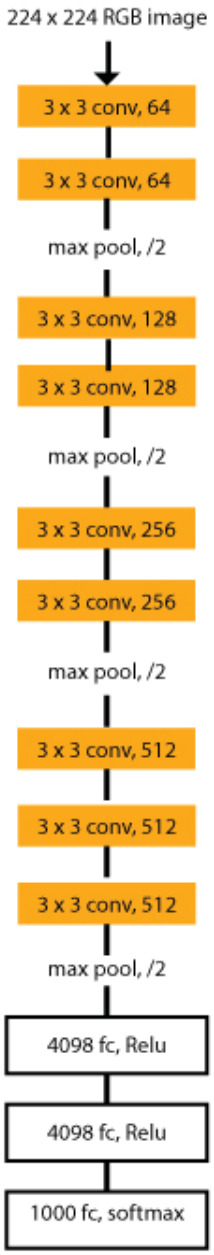

Figure 2.6: Architecture of VGG16

The figure 2.6 is showing the convolution layer with the reLu activation function along with 1000 fully connected layers.

The input to the conv1 layer is an RGB array of pixels of dimension 225 × 225. The features of the image passes through a stack of convolutive layers with Restricted Susceptible field: 3 × 3 [25]. By also using 1×1 filters, the device works as a conversion of the input channels [25].one pixel is set for Convolution stride; To the convolution ,the spatial padding is added. The layered input is such that after convolution, the spatial resolution is preserved, i.e. the 1-pixel padding around the original image is retained. Layers A maximum pooling layer followed by five further pooling layers is accomplished by spatial pooling. layers are followed by max-pooling. Pooling is done over a 2×2-pixel window of image samples, with stride 2. Three fully-connected layers adopt a stack of convolutional layers. The first of these has 4096 feature channels, the second has 1000 feature channels, and the third has 1000 class labels (one for each class)[44]. In the final layer soft max is used. Throughout the network, the topology remains constant.All secret layers with ReLU nonlinearity are employed. It is noted that none of the networks (except for one) employ LRN, and it is assumed that LRN will minimize the memory and computation requirements of ILSVRC.—cite46

Advantages VGG 16:

- TIt has strong architecture, so it is standard for solving any difficult problem.
- As pretrained VGG 16 is directly accessible from the internet,it is very easy to use.
- It has a large network architecture as a result mostly has implementation in image classification[25].

2.6 Transfer Learning

Transfer learning is a type of supervised machine learning methodology that uses sections of previous models on learning tasks.Transfer learning is valuable when there isn't enough data to do your classification problem through a neural network and there is a large existing database that can be transferred to your problem.Transfer learning uses an existing model to solve a new problem instead of teaching a new model for that purpose. Just as an analogy, if you trained a basic classifier to predict if a picture contains a backpack, you might use the information of the classifier to predict the existence of other objects. Researchers have used neural networks in computer vision fields such as medical diagnostics and intelligence research. In the transition learning, the first and middle layers are introduced and the latter layers are retrained. This allows data used for the original mission being done.Through transfer learning, people aim to transfer as much material as possible from the training assignment to the new task. The information will come in different ways, it all depends on the problem and the data. Models help one to classify objects that do not exist in the real world.[12]

2.7 Explainable Artificial Intelligence(XAI)

Explainable AI (XAI) refers to knowledge that is programmed in a way that can be easily understood.Mainly, AI systems collect, evaluate and make decisions with a large number of data without having a legible output[46]. Explanatory AI, however, allows us to analyze predictions and understand decisions of AI systems by showing them using graphs, explanations and routes[42]. XAI plays an important role in deep learning and is frequently discussed in relation to this technique (impartiality, accountability and transparency in machine learning).It removes the drawbacks of deep learning methodologies[46].

2.8 Different Methodologies of XAI

There are several methodologies of XAI such as Counterfactual Method,Local interpretable model-agnostic explanations(LIME), Generalized additive model(GAM), Rationalization, Layer-wise relevance propagation (LRP) etc.

2.8.1 SHAP

SHAP is a tool used for representing individual projections. Shapley's value is based on the game theory optimum Shapley values. The purpose of SHAP is to explain how to forecast an instance x based on how each function contributes to the prediction[52]. The SHAP retains several approaches for global explanations for Shapley values. SHAP has many separate aggregation algorithms aimed at effective collection and aggregation of Shapley values. Shapley values decide how much everybody can get out of each function. The only solution that satisfies four properties is the Shapley Value. If the game satisfies these, it satisfies these[52].

2.8.2 Counterfactual Method

Counterfactuals are widely used in XAI applications to understand human decision-making and actions.It is useful in helping to provide coded representations of complex structures to developers and end users[32]. Counterfactual explanations can be used to explain forecast of different tasks very accurately[51].

2.8.3 Local Interpretable Model-agnostic Explanations (LIME)

Plan A is a Model-Agnostic Explanation of Molecules that can be represented locally on the level of various cultural backgrounds and languages. By the way, each aspect of the name represents something completely desirable in explanations. A local embedding is one in which various instances of the same classifier will have different embedding since they are selected in a specific way. The main distinction between a normal embedding and a local embedding is that you can only adjust how many nodes are in a local embedding[52].

2.8.4 Generalized Additive Model (GAM)

A GAM additive model is a statistical model in which the linear predictor depends on some smooth functions of the independent variable linearly.The linear predictor vector relies linearly on the coefficient vectors of the independent variables or a GAM[52]. In the statistical research area of statistics, generalized additive models are widely used to combine process characteristics of general linear models with additive models. As a consequence, it is a function which relates the bivariate response Y to the asuming variables Xi[52]. Let Y be the "X" for which the probability density(pdf) is defined under the relation function g for the corresponding probability mass function (pmf). Then, by the product law, the probability density function (pdf) of Y is specified both for the "x" and the link function g, by the exponential family distribution. And within this clever and fascinating algorithm is Self-Driving Vehicles, Machine Forensic Methods, and other so-called explainable artificial intelligence.

2.8.5 Rationalization

In Artificial Intelligence, an approach for rationalizing autonomous system behavior as if a person were conducting the same thing the system is actually performing. A framework used for neural machine translation that explains how an autonomous agent may articulate its internal state-action representations in a direct and functional way for an individual or particular machine receiving the transmission. Researchers tested their strategy in the Frogger game world and taught agents to rationalize its behavior decisions using natural language and based on various graphical representations of the game worlds. A natural language credential is assembled by an algorithm using an audio recording of me in a puzzle game. We propose two studies to assess the usefulness of rationalization as an approach to reason generation to get two groups of participants to produce theories and consider their impact on a final decision. Human subjects were tested in an analysis whereby they were required to explain (labeling) actions conducted by an entity intelligent enough to solve problems and define optimal tactics to achieve a target. Findings from the experiments showed that (human subjects) often given neural machine translation interpretations that represented behavior of intelligent agents, and the translations for these translations were more satisfactory to the subjects than other explanations[52].

2.8.6 Layer-wise Relevance Propagation (LRP)

Layer-wise Significance Propagation (LRP) is a process that gives both prediction and clarification in complex deep neural networks.It works by backwards propagating to find out the prediction in the neural network .It is a very quickly and reliably method that use a series of purposefully constructed rules for propagation.There are several forumalas and application of LRP such as LRP rule, LRP rule,LRP b rule etc.LRP also used in wieght mapping connection.[51].

The LRP algorithm uses back propagation to reach the input layer to the output layer. Initially, the last layer is considered as the output layer.Later on,by using the redistribution rules the corresponding function is back propagated to the input layer[51].

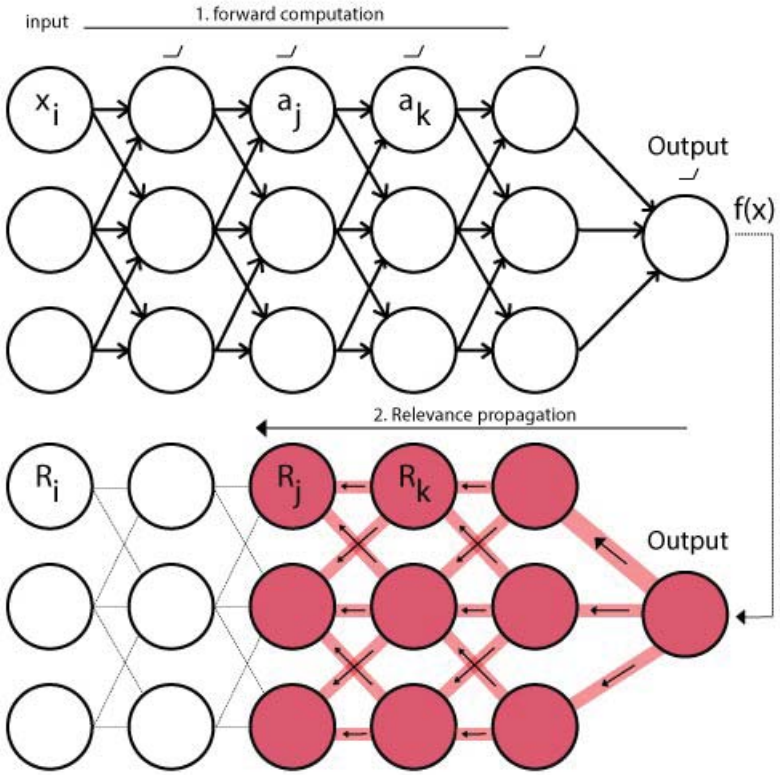

Figure 2.7: Procedure of LRP

Here Figure 2.7 is representing f(x) is the output and R is the function of j,k layers.

LRP decomposes the prediction of a deep neural network where in the initialized output is f(x). It's main function is to decompose every neuron which is directly connected to the output with the relevance propagation. Every Neuron has a certain relevance from top to down. The relevance is related to neuron inhibition and while doing simulations of the output should be a positive number[51].

Chapter 3

Releated Work

The research from [48] stated a classification model of OCT images by using neural networks. They used a dataset of four classes (normal, drusen, diabetic macular, choroidal neovascularization). As there are 3 types of pathologies visible in OCT they classified the design in 3 parts CNV, DME and DRUSEN. The classification stage was divided into two sub steps, one is training of deep learning known as Convolutional Neural Network and the other one is validation of the model[20]. For the classification, the data set was labeled by OCT with 84495 images, where each image went step by step classification consisting of several layers of trained classifiers. The second layer of classifiers consists of four ophthalmologists and the 3rd level is checked by 2 retinal specialist. As its been classified as three 3 parts where CNV deals with symptoms of central scotoma and metamorphosis. DME deals with retinal thickening involving of the macula which most common cause of vision loss for the patients suffering from diabetes. DRUSEN are classified deposits of extruded mitochondria that appear in the upper part of the optic nerve [20]. The proposed system was divided into three stages. At first division and normalization of the dataset and during the 2nd part training was carried out and finally the validation was done. The runned a k fold algorithm to separate the test images .And those images are applied in the test data prediction to collect the accuracy. As CNN has been used here, the layers they worked on are Convolutional layer, pooling layer and connected layer. Convolutional layer is composed of a set of kernels which is capable of learning according to training. The kernels are convolved with input data for getting a feature map.The actual values of the kernels change throughout the training, causing the network to learn to identify significant regions to extract characteristics from the data set. And the pooling layer is used to reduce the spatial size of resulting convolutional metrics[48].

One of the research from [45] stated that to detect Diabetic Macular Edema(DME) and age-related macular degeneration through OCT (optical coherence tomography)dataset. There are two primary drawbacks of their image classification process. Firstly, the dataset of the OCT is limited to a small number. They stated that, to use the characteristics in the network to reduce the vulnerability of the datasets as well as increase the versatility among the different datasets. For the classifiaction they used FastNIMeans and bilateralFller methods on the optical coherence tomography datasets. Due to the complete the preprocessing, they cropped the images.They used two types of datasets one is from an University another is from a

Hospital. Their datasets are categorised as three types AMD, DME and NORMAL. In the preprocessing, as they used two different datasets they used fine tuning to reduce different size of the images.In their verification set,they used 20 images and classified them as AMD in the class 0, DME in class 1 and NORMAL in class 2. They used the SVM method for the getting output. They got a good accuracy 95.5 percent in dataset 1 as well as got 98.6 percent for the dataset2 by the using the VGG 16. They also showed that DenseNet model has the advantages of feature extraction as well as the increase the feature regenerate. For these reason they also work on the DenseNet model. They showed comparison between their work with other models such as VGG19, VGG 16,Inception V3. They proposed that CliqueNet and DPN model can be more productive than the other models, their accuracy are also very high range than most other models in the deep learning.[45]

In one paper, the authors applied fuzzy image processing techniques and graph-cut methods to divide the OCT(Optical Coherence Tomography) into five distinct layers[37]. These methods will also grant the basics with the early diagnosis of eye disease. In their paper, they claimed that there are four major reasons for blindness such as age-related macular diseases, glaucoma, diabetic retinopathy, and cataracts. In this paper, they propose an method for OCT image segmentation with the combination of inexpensive methods. Firstly, they started with ROI(Region of internet) with the all layer information. After that, they transform all the images by using a fuzzy histogram to get the better result of each layer. Then they clustered the image using fuzzy C-means they use the value of the brightness of pixels to get the data to integrate into maxflow framework. They mainly showed the five retina layer segments on an OCT image. The five sections are 1.NFL 2.GCL+ IPL+INL 3.OPL 4.ONL 5. IS+OS+RPE. They proposed a method using two main parts. One is preprocessing and another one is segmentation. In the prepossessing step to get the maximum value of bright layer and minimum value of dark layers they used fuzzy histogram hyberbolisation. In the segmentation part, they used the Max-flow optimisation to solve the individual variable position. In this paper, they used the Lagrangian multiplier to optimise the flow. Up to find the convergence they used the repetition of multiplier and maxflow. They applied the MatLAB to determine the accuracy of the proposed model. Finally, they showed their output with a comparison that regions of RPE performance were more accurate than their proposed model[37].

One of the research from [27] stated that they utilized deep learning to classify AMD or Normal from the OCT images. In the manual image classification needs to depend on the feature extraction however by using the OCT images can be categorized by following the different methods of machine learning such as SVM, Random Forest, KNN, MLP etc. They worked with the raster macular OCT scanning. After that,they also combined the patient's medical report with the OCT scanning image.They divided their Images into two Division one is for validation and the other one is for training. For training purposes they chose the images randomly. For the classification purposes they used the VGG 16 model as well as used the Xavier algorithm for the classified image classification. In their methodology, their image batch size was 100. They also worked with the gradient descent optimization. They showed the main goal by doing classification of AMD versus normal from OCT im-

ages. However, their training model was stopped after the 8,000 epoch because of the overfitting. In this paper,they only worked with those patients who fulfill their study Criteria as well as only worked with real world images[27].

The authors from [47] stated their purpose to build a classification model that was not too difficult to differentiate type of cancerous cells.In the region they scanned the breast tissue,those can be illuminated for use in an intra-operative environment for margin evaluation. During imaging, the average specimen scale was between 1.0 and 4.0 cm2. They collected low-density frozen tissue samples from 23 patients and were analysed using a custom ultrahigh-resolution OCT system. They used deep learning algorithms and customized hybrid forms of machine learning. Two-dimensional CNN that maps each 2D B-scan to a 1D mark vector by using 2D convolutional filters, the tissue types for each trunk line is affected by its surrounding lines.Before the OCT picture volumes were preprocessed, the CNN analysis was carried out. The B-scans of a single volume from 1024 £ 800 pixels to 256 £ 200 pixels is down-sampled, of pixels.Using an 11-layer architecture composed of serial 3 £ 3 convolutional filters added to the CNN was introduced. The down-sampled B-scans) with growing channel sizes With increasing convolutional depth, from 4 to 64. Convolutional filters are implemented using two strides per position in the depth-to-shallow axis to compress the picture depth, to hold the picture big, a stride of one was added in the left-to-right direction. In this research, the deep learning algorithm achieved 94 percent accuracy, 96 percent sensitivity, and 92 percent precision in a binary classification of detecting cancerous versus non-cancerous tissue in OCT pictures of breast specimens[47].

The research group from [38] stating about the diagnosis of retinal disorders using neural networks. They have used optical coherence tomography images as the dataset. In their paper they used CNN to classify OCT image. They believed Ophthalmologists can readily classify the chief areas of risk in the eyes, popular characteristics of corneal disorders frequently seen in an OCT picture[19]. First they developed a lesion detection network (LDN) to produce an autoencoder. A single part of the entire OCT picture (the soft attention map) the focus map has shifted. The data is then used in a classification network to assess the metric. The classification network will use the details from local lesion-related areas to get still more information. The network training phase and increase the precision of OCT imaging. In two clinical acquired OCT datasets, their experimental findings have been performed[38].

Another research from [24] proposed that residual network is effective to optimize as well as give better accuracy than VGG. They also represented the CIRAR-10 with the explanation of their layer and they got a better reformation of COCO object identify dataset.In this paper,they mentioned that by increasing the layer in the deep neural network will have better performance, along with that it takes less training defeat. However, it becomes very difficult for optimization. In their paper,they also mentioned that while the layer becomes less test error on the other hand, while the increasing layer of deep network becomes 56, it takes more test error than previous layer.And,they also started that overfitting was not the reason behind it. They mentioned that the input layer has to learn the identity function

to get the high accuracy. However, some layers didn't learn this as they were initially zero,some layers are Gaussian with standard deviation which are also near to zero. In the convolution layer learning the identity function is hard. This paper mainly focused on the initial input initial layer should become a default function that does not become a zero function. They also proposed the skip connection of identity function to get the same input and output size.In the residual network main goal is to get high accuracy. They used "relu"which is an activation function used for vanishing gradients. They also showed a comparison of other models such as. VGG19 has a lot more parameters and operations than a residual network. The ResNet has less complexity than VGG19 as it has fewer parameters layers. ResNet has less filter then VGG19. In their paper they showed that 34 layers plain less performance rather than VGG19 as it had more layers. However, the residual networks have identity blocks; they skip the weight layer and provide higher accuracy. In the result part, they showed that by increasing the layer the training and validation error become higher. In contrast, by using a residual network it can reduce training error. In their paper,they showed that ResNet152 with a higher layer than VGG19 performs better and has less parameters. Finally, they showed an accumulation of their all models with their parameters and training error[24].

The research group from [51] proposed method is XAI with the model LRP(layerwise relevance propagation) for the object localization as well as the class differentiation of the LRP. The deep neural network has become very effective in the research field for research purposes. They stated that LRP sustains a well built link with the output predictor. In their paper, they stated that they made a series of LRP to get a comparison between present LRP and earlier LRP's functions and operation. They used the computer vision datasets for the object localization with the help of the ImageNet and pascalVoc.They used the ReLU activation function as well as convolution and pooling layers. They used Layerwise Relevance propagation for the various classes by the following the method of the VGG-16 model. They also showed that neural networks have predictability power which created the Deep Taylor Decomposition (DTP).They used the basic attribution rule of LRP to find out the output prediction. The LRP rule is used for combining the different decomposition functions. They also used the LRPb2 decomposition rule to know the information of higher output layer neurons[51]. They used XAI for the current better practice for the LRP. They showed the comparison LRP with the other variation of the LRP by using the VGG-16 model from the keras model zoo.In the validation step, they used their dataset for the predicted localization of the object area as well as image background. For the attribution localization they use the inside- total relevance ratio as a quantitative measure. Their work showed a good accuracy which is 90.1 percent for the ImageNet test sel. They presented in their paper about the decomposition of the Layerwise Relevance propagation with other methods. They showed that LRP CMP doesn't provide the attribution map. However it gives a solution to reduce the problem of gradient shattering which is connected to the object localization as well as the differentiation of the different class[51].

The research group from [49] stated Graph Neural Networks (GNN) are the cooler way to graph organizeds all of the credit. For neural networks that tightly intertwine the input node graph with the neural network structure, it is like an explanation

can solve, but is not applicable. The core feature of these tools have been a secret a-la Minority Report for the consumer so far. In this paper, we show that GNNs can be naturally explained with the aid of higher-order expansions, such as defining groups of vertices that jointly contribute to the prediction. In a very straightforward change of rhetoric, as indicated by their research, they found that the attribution of causal patterns can be more reliably retrieved using a compilation of basic rules for changing the sentence embedding weights. Out of potential candidates, the output is a filtered multitude of steps or "walks" towards the input graph, the result is a large enough set of steps that led to the right response. Our novel explanation method (GNN-LRP), which we denote by GNN-LRP, is applicable to a wide range of graph neural networks and lets us extract practically relevant insights on sentiment analysis of text data, structure-property relationships in quantum chemistry, and image classification[49].

Chapter 4

Proposed Model

4.1 Dataset Description

The dataset of our proposed model consists of thousands of validated OCT images.We worked with the image classification by deep learning models with the explainable artificial intelligence for Optical Coherence Tomography images. These OCT images were collected by patients from the an eye institute which is a retinal research Foundation, in Beijing[47] .There are total 104,312 number of images in our dataset. The OCT images were being selected from the duty of patients daily checkup in that hospital. In our proposed model, we divided our datasets into 4 directories such as CNV, DME, DRUSEN and NORMAL. In our datasets we don't have any parameter accoridng to age or gender. We examined local electronic past medical records for classifying the Choroidal neovascularization, Diabetic macular edema, DRUSEN, and normal images of OCT [47].

DATA	Train Set	Test Set
CNV	32,206	250
DME	11,349	250
DRUSEN	8,617	250
NORMAL	51,140	250
TOTAL	103,312	1000

Table 4.1: Dataset Description

4.1.1 Figure of DATASET

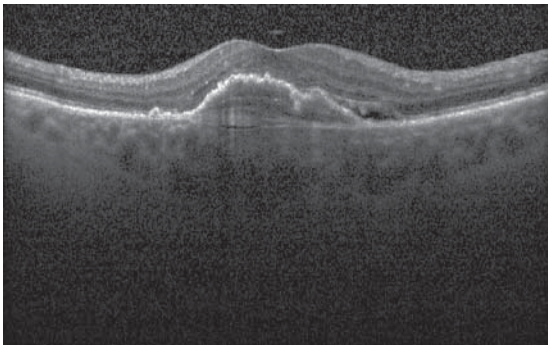

Figure 4.1: CNV Image 1

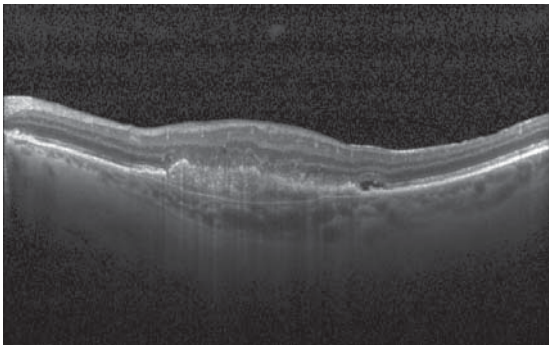

Figure 4.2: CNV Image 2

These figures are choroidal neovascularization (CNV) as here neovascular membrane is covered with the subretinal fluid.

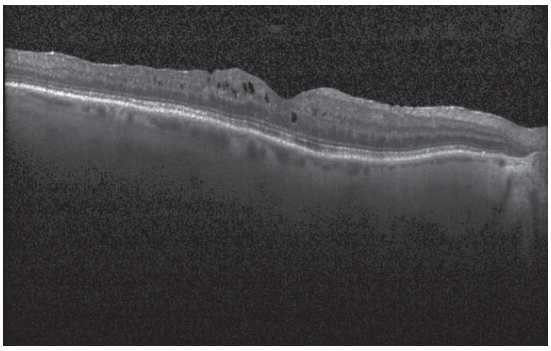

Figure 4.3: DME Image 1

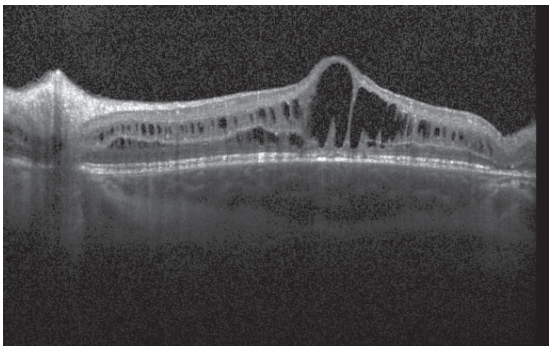

Figure 4.4: DME Image 2

These figures are Diabetic macular edema (DME) because the thickening of the retina is connected with the intraretinal fluid.

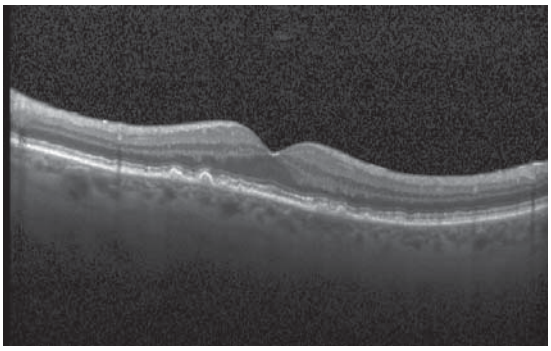

Figure 4.5: DRUSEN Image 1

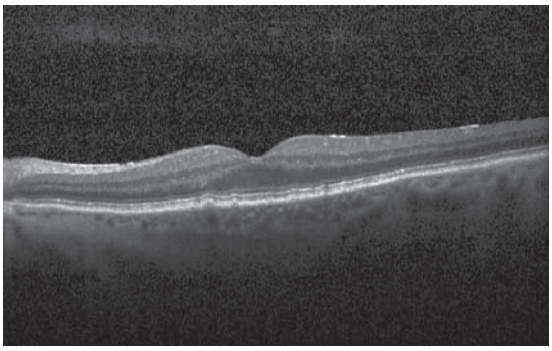

Figure 4.6: DRUSEN Image 2

These figures are DRUSEN as it is showing deposition of lipid that accumulate under the retina.

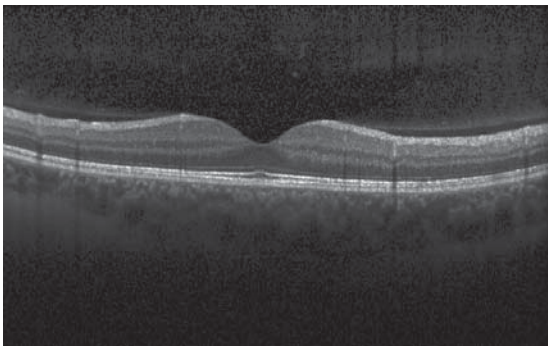

Figure 4.7: NORMAL Image 1

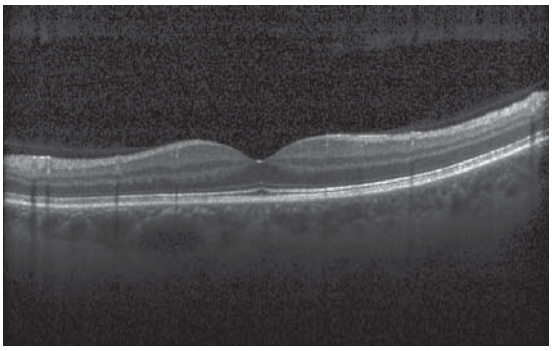

Figure 4.8: NORMAL Image 2

These are Normal images of the retina as they do not have any retinal fluid.

4.2 Data Pre-processing

For data pre-processing, we created a validation dataset in a separate folder named 'val' from splitting images from the training image set. Same as the training and testing set, the validation data also consists of CNV, DME, DRUSEN, NORMAL images separately. Validation data plays a vital role in getting unbiased predictions while training a model. Generally, a testing dataset can be biased or not balanced properly in terms of it's classes. In that case we can create a validation set as our needs. That's why we have taken 11826 images in four different classes in our validation data. Our validation data consists of 3220 images in CNV, 1134 images in DME, 861 images in DRUSEN and 5114 images in NORMAL which is 10 percent of training data. We have used the image dataset from directory function of Keras to load our dataset through Directory argument.

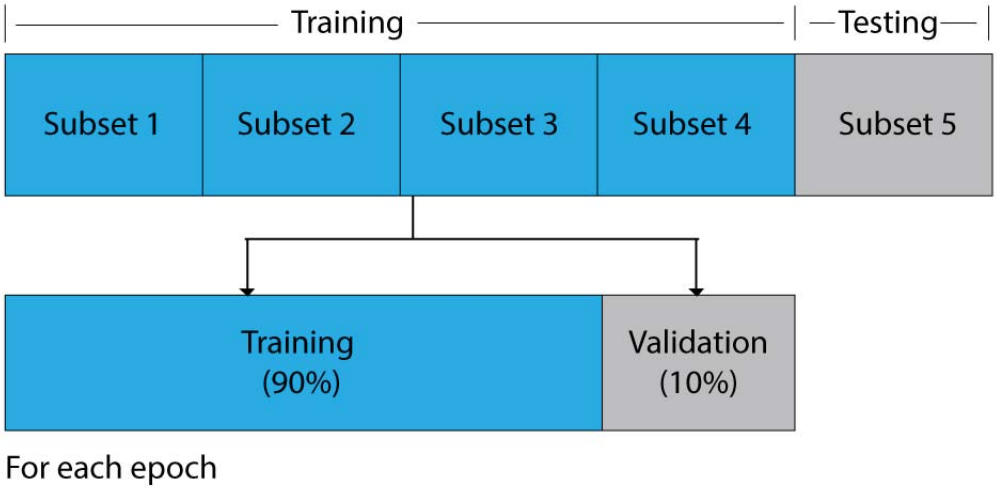

Figure 4.9: Formation of Dataset

4.3 Experimental Setup

We have used the noble approach "conda create -n research tensorflow-gpu==1.12" in anaconda prompt to create a virtual environment which have collected all necessary dependencies like cuda toolkit 10.1, cudnn 7.6.5, python 3.6, tensorflow gpu 1.12 etc. Then we have installed Matplotlib 3.3.1 for and Keras 2.2.4. The whole experiment was done in Windows 10 operating system with Ryzen-5 3600 CPU and NVIDIA 1660 Super GTX.

4.4 Data Augmentation

Image Augmentation is one of the key parts of preprocessing which helps to generalize the whole data in a standard size. To implement this on our dataset, an image data generator function of Keras framework has been used. Firstly, we have applied the rescale argument in train, val and test set to normalize our data. As images are having 255 pixels which is complex for neural networks, we applied 0.0039 (1/255) in the rescale argument. Secondly, training images were randomly rotated 30 degree through rotation range argument as some of the images were curved. Some of our images were not in proper shape of representation, That's why we have used width sifr range function to shift random images horizontally. Adding that, some of our images were in the upper part of a frame and some were in the lower part of a frame. Therefore we have used the height size range function to shift random images vertically. There were few images which were zoomed out and some images were too much zoomed in condition. For this, we kept the zoom range between 0.9 to 1.25 through the zoomrange function of keras framework. Generally, To generalize our left eye and right eye's scanned images, we have flipped random images. Furthermore, we didn't flip our images vertically as it might misbehave during the training period.

4.5 Initializing Training parameters

We have initialized batch size as 20 for the training procedure. Thus, our dataset has been trained by 20 images in one step. In general, taking a small integer value as batch size is mostly convenient and preferable. On the other hand, to fit images in the model, we have taken the input shape of the images as 224 pixels x 224 pixels. Defining input shape parameters is a very important job of the training procedure. The accuracy of a model's output largely depends on it. Due to training the dataset on a model, we have decided this input shape parameter for best results and better accuracy. Taking the input shape parameter in a very large amount can be time consuming and unqualified for a partial amount of the training dataset if any image's pixels are not as large as the input shape value. In contrast, if the input shape parameter is too small then it will largely downgrade the accuracy of the model's output. That is why, keeping in mind all the possible scenarios, we have decided that particular number as input shape. We have used different step sizes to accomplish training, validation and testing. The step size for training, validation and testing was 4823, 591 and 50 respectively. Step size plays a significant role in convergence to a specific axis during training period. That's why we followed the

traditional approach of training data divided by batch size to determine the ideal step size.

4.6 Preparing CNN Model

As OCT dataset is relatively small which takes less training time and gets higher accuracy, we are proposing a few deep neural network models to classify different classes of our dataset. VGG16, ResNet50, Xception, Inception V3 are very promising convolutional neural networks which were pre-trained on the ImageNet dataset. Initially, the convolutional layers were loaded with the pre-trained weights. These weights were calculated and saved for redundant processes reduction and boosting training. Then we initialize a new model for normalization along with adding loaded models output. After initializing the model, we added a Flatten layer which flattens the output of the model as it is from a convolutional layer. In this layer, the output parameters of the upper layer are multiplied and stored in the flatten layer. Then, we have a dropout layer with the parameter 0.5. Adding a dropout layer can prevent the neural network model from over fitting [48]. The function of the dropout layer is to drop units along with it's connection from the network. After that, a fully connected dense layer is added with relu activation function. Here, 1024 neurons are implemented for this layer. A dense layer of neurons feeds all outputs from the previous layer to each separate neuron, and in turn, each neuron sends its output to the next layer.Relu which is an activation function, that outputs determines the input directly if it is positive, otherwise, the output will become zero. Lastly, we have added a new dense layer with softmax activation function as the final layer. The softmax function is used to activate a multinomial probability distribution in the output layer of the neural network models.

4.7 Training Dataset

After initialization of training parameters, the compiler has been generated for training purposes.

4.7.1 Adam Optimizer

The Adam optimization has been applied as an optimizer for the compiler with a learning rate of 0.0001. An optimiser leads to quicker performance. The optimizer we use for determining, can update the weights and learning speeds of the neural network to lower the losses. Algorithms or techniques for optimization are responsible for minimizing losses and producing the most detailed results possible. The most simple and used optimization algorithm is Gradient Descent. It is used extensively in algorithms for linear regression and classification. Stochastic Gradient Descent is the part of Gradient Descent which attempts to more consistently change the model's parameters and model parameters are altered for each training example after loss calculation. Adam is the best optimization algorithm for stochastic gradient descent which integrates the best properties of the Adaptive Gradient Algorithm (AdaGrad) and Root Mean Square Propagation (RMSProp) algorithms to have an

optimization algorithm that can work with sparse gradients on complex problems [17].

4.7.2 Learning Rate

In neural network training, the learning rate is generally a customizable parameter with a minimal positive value between 0.0 and 1.0. The model's learning rate controls how easily it adapts to new problems. It will take more training epoch in lower learning rate thus higher learning rate takes less training epoch. If a learning rate is too low that can lead the way to getting stuck while if the learning rate is too high that can lead the model to finding a weak solution in a short given time. We have applied different learning rate to get the best output and finally we have decided to have a lower learning rate (lr=0.0001) so that the model can adapt possible solutions more precisely.

4.7.3 Categorical Crossentropy

To implement an optimization algorithm, it is necessary to evaluate the error repeatedly throughout the current state of the model. This includes the option to use a loss function, which can be used to evaluate loss of the model, so that at the next measurement weights can be changed to minimize losses. Generally neural network models learn a mapping from input to output by looking at examples, so the choice of loss function for a prediction problem must fit the context of the problem. As we have four different categories to classify, we have used categorical crossentropy loss function to estimate the loss of the model.

Finally we have operated our training with given training and validation's directory and step size. The whole training operation was done using keras framework in 10 epochs.

4.8 Analyzing through XAI

Neural networks have advanced the state of the art in many areas in recent years, such as image classification, object detection, speech recognition etc. Neural networks are usually still regarded as black boxes, despite their performance. Their internal roles are not well known and the basis is unknown for their predictions. Explainable Artificial Intelligence (XAI) mainly explains the reasons how the decision of the black box of neural networks is made. It helps user of Machine learning to gain more trust on Artificial Intelligence models. Several approaches, such as LRP, Deconvnet, PatternNet Attribution, Guided Backprop, Smooth Grad, Saliency, Integrated Gradients etc. were introduced in an effort to better understand neural networks.
We have analyzed our trained models of VGG16 and InceptionV3 through Explainable Artificial Intelligence for understanding how the models have implemented training on our dataset. This implementation is based on Keras which requires keras-backend to execute on models. That is why, the training by different models on our dataset is also done using keras-backend.
After finishing with training the dataset with neural network models, we have loaded

the trained model and initialized the preprocess input function corresponding to the model. The preprocess input function is mainly used for adapting the input images to the format specified by the model.
After that, we have stripped the final layer from the model which includes the softmax activation function and added a new dense layer. The final layer of the trained neural network model includes the predictions of the dataset which can be defined as the output layer for the model. As XAI only works with the training parameters, there is no need for the output layer for analyzing the model. Beside that, the analyser for investigating the images only accepts and runs on a model which only consists of training layers without any prediction layer[36].
Then we created an analyzer for analyzing the images. Here, we have used DeepTaylor as our analyzer. For each neuron, DeepTaylor calculates a rootpoint that is similar to the input, but whose output value is 0, and applies these differences to recursively estimate each neuron's attribution. By using this analyzer, we have analyzed our training images. As a result, we have got a new image as an output for each image which gives a better understanding of the training process. These output images are only focused on the particular segment of the image where the neurons of the model have focused for training.

4.8.1 Deep Taylor Decomposition

Deep taylor decomposition mainly focused on getting relevant information from input. This the input f(x) is mainly the output of our trained model from figure 2.7. We are doing relevance propagation through deep taylor decomposition. As we are doing back propagation to open the black box, so we need to find out the relevant weights of the next layer through deep taylor decomposition.

Deep Taylor Decomposition Function :

$$R_j = \sum_k \frac{a_j w_{jk}}{\sum_{0,j} a_j w_{jk}} R_k \tag{4.1}$$

Here, R is the Relevance function , j is the lower layer and k is the next layer of j. a_j is activation of layer j. We have chosen $Lrp - \gamma$ to analyze our model as this is the most suitable method to extract the upper layer of the neural network[28].

4.8.2 Steps for XAI implementation

- **Step 1 :** Load an input image used
- **Step 2 :** Plotting the input image
- **Step 3 :** Load the trained model
- **Step 4 :** Load the preprocess function
- **Step 5 :** Strip the final layer containing softmax activation
- **Step 6 :** Create an analyzer by Deep Taylor

- **Step 7 :** Preprocess the input image
- **Step 8 :** Applying analyzer for the input image
- **Step 9 :** Aggregate analyzed image along color channels and normalize to [-1, 1]
- **Step 10 :** Plotting the analyzed image

Chapter 5

Result and Analysis

Since our research is about medical diagnosis classification, accuracy of our results will play a vital role in terms of patient's treatment. The accuracy and prediction are therefore significant indexes that are taken care of. We have applied Inception V3, Resnet50, VGG16, Xception model to our dataset and got prediction accuracy. There were 1000 images in the test dataset which were taken to evaluate these models.

5.1 Performance Metrics

$$\textbf{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.1}$$

Here, TP is true positive, TN is true negative, FP is false positive and FN is false negative. In our case, TP and TN shows that patients were diagnosed correctly and FP and FN represents that patients were diagnosed inaccurately. Inception v3, Resnet50, VGG16, Xception all show good accuracy on OCT dataset. The result is given below.

Model	Train Accuracy	Validation Accuracy	Test Accuracy
Inception V3	97.29%	95.39%	95.20%
Resnet50	96.08%	94.19%	94.00%
VGG16	97.27%	94.36%	96.30%
Xception	97.07%	94.46%	93.90%

Table 5.1: Performance Table

One of the research groups who officially released the third version of OCT dataset got 93.4% test accuracy from this dataset[45]. But we have got higher test accuracy by applying the fore-mentioned models.

5.2 Training and Validation's Accuracy and Loss

As we performed different neural network model's to our dataset, we have got their training and validation's accuracy and losses during training and validation period.

5.2.1 Result of Resnet50

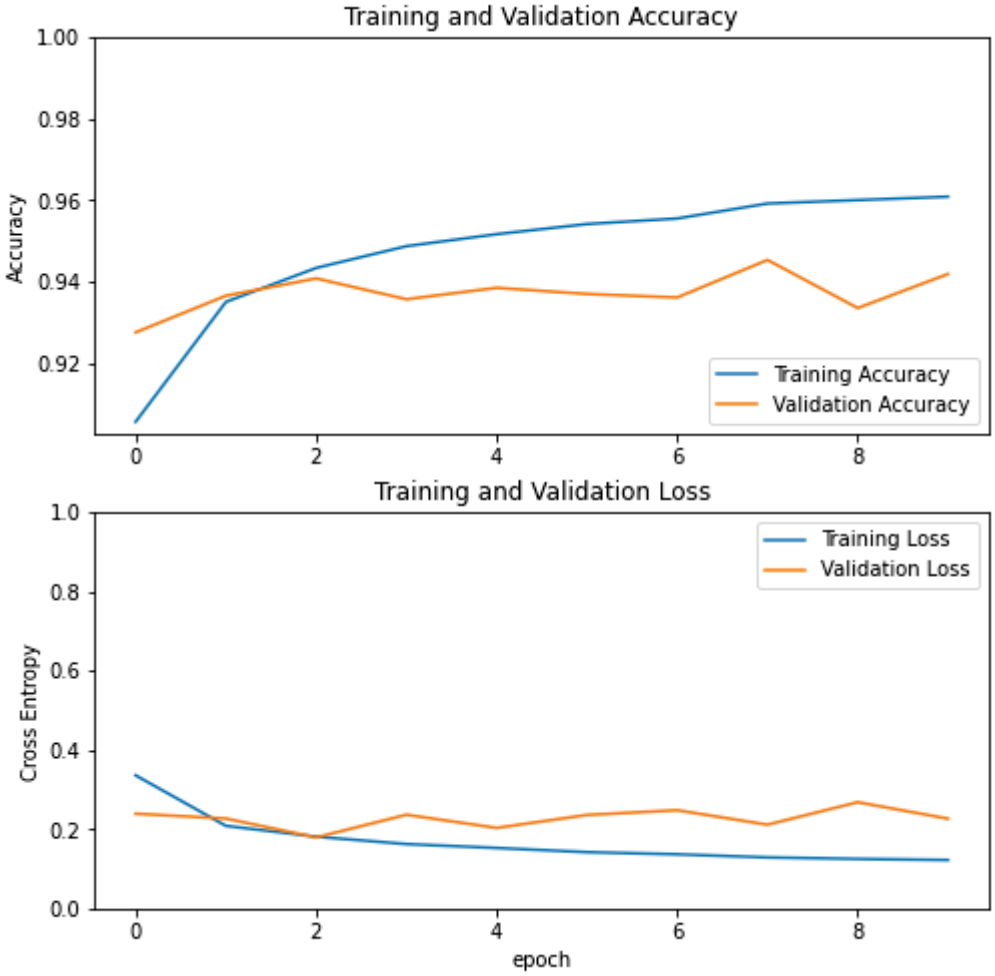

Figure 5.1: Training and Validation's Accuracy and Loss of Resnet50

The figure 5.1 shows that training accuracy is increasing but the validation accuracy dropped after 7th epoch and increased after 8th epoch. Training loss is decreasing quite well but validation loss has fluctuated a few times and ended in 0.2265.

5.2.2 Result of Inception V3

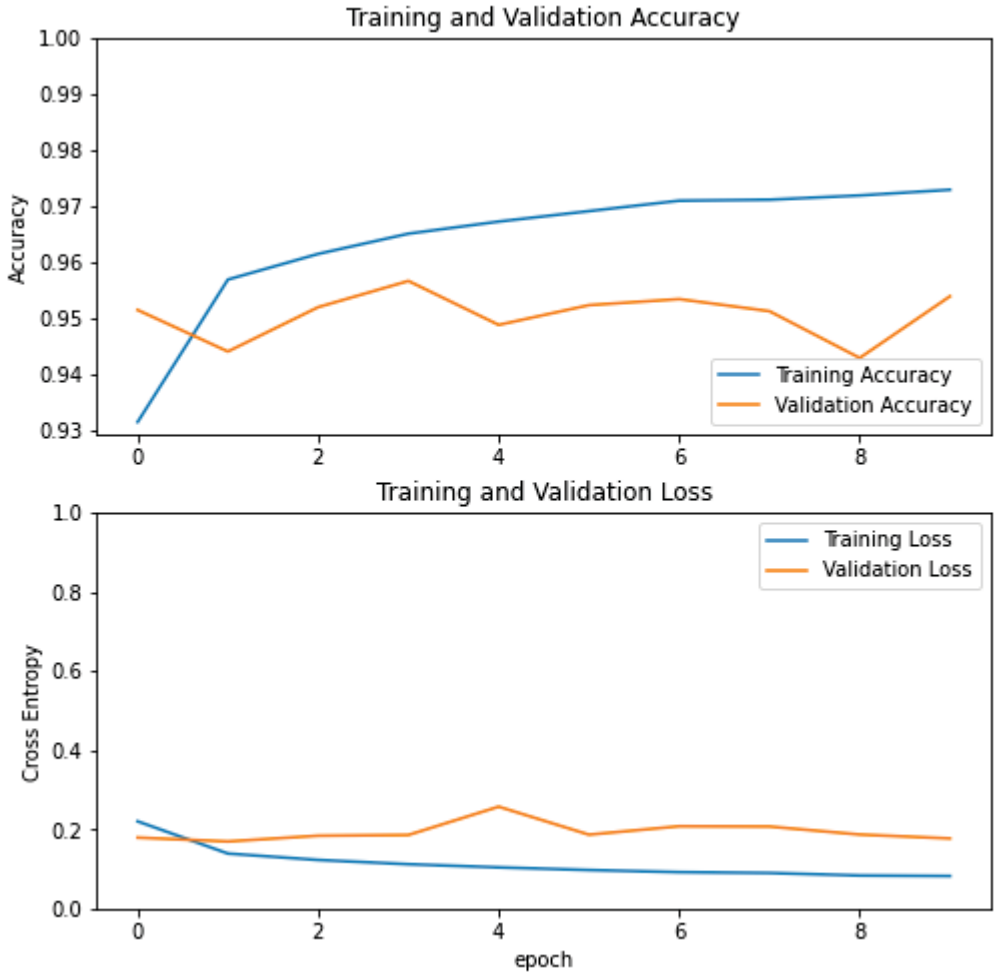

Figure 5.2: Training and Validation's Accuracy and Loss of Inception V3

The figure 5.2 represents that training accuracy has changed 2% after the first iteration and stated increasing consecutively whereas validation accuracy has fluctuated but stayed between 94% to 96% . The training loss was decreasing continuously and validation loss was 0.1770 lastly.

5.2.3 Result of Xception

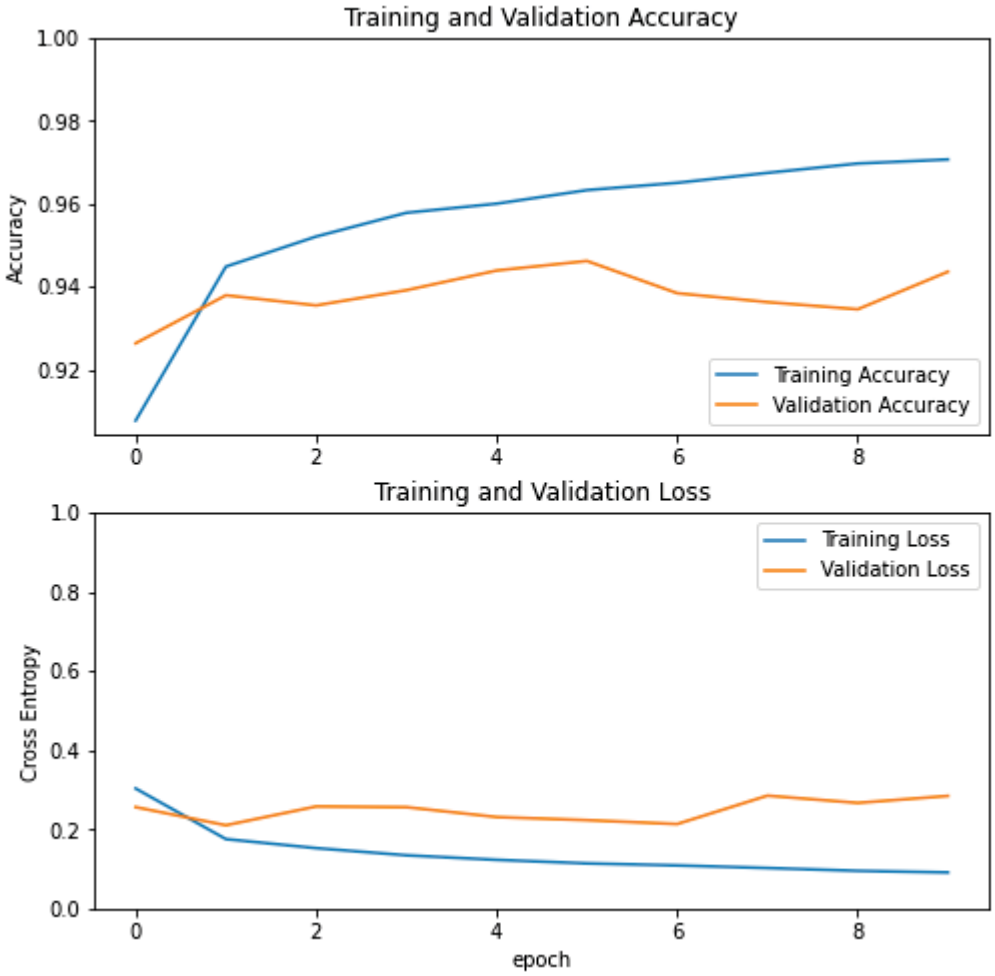

Figure 5.3: Training and Validation's Accuracy and Loss of Xception

The figure 5.3 shows that training accuracy is increasing whereas validation accuracy had decreased after 6th iteration but started increasing and ended in 94.46%. The training loss was decreasing continuously and validation loss was 0.2844 which is not a good state.

5.2.4 Result of VGG16

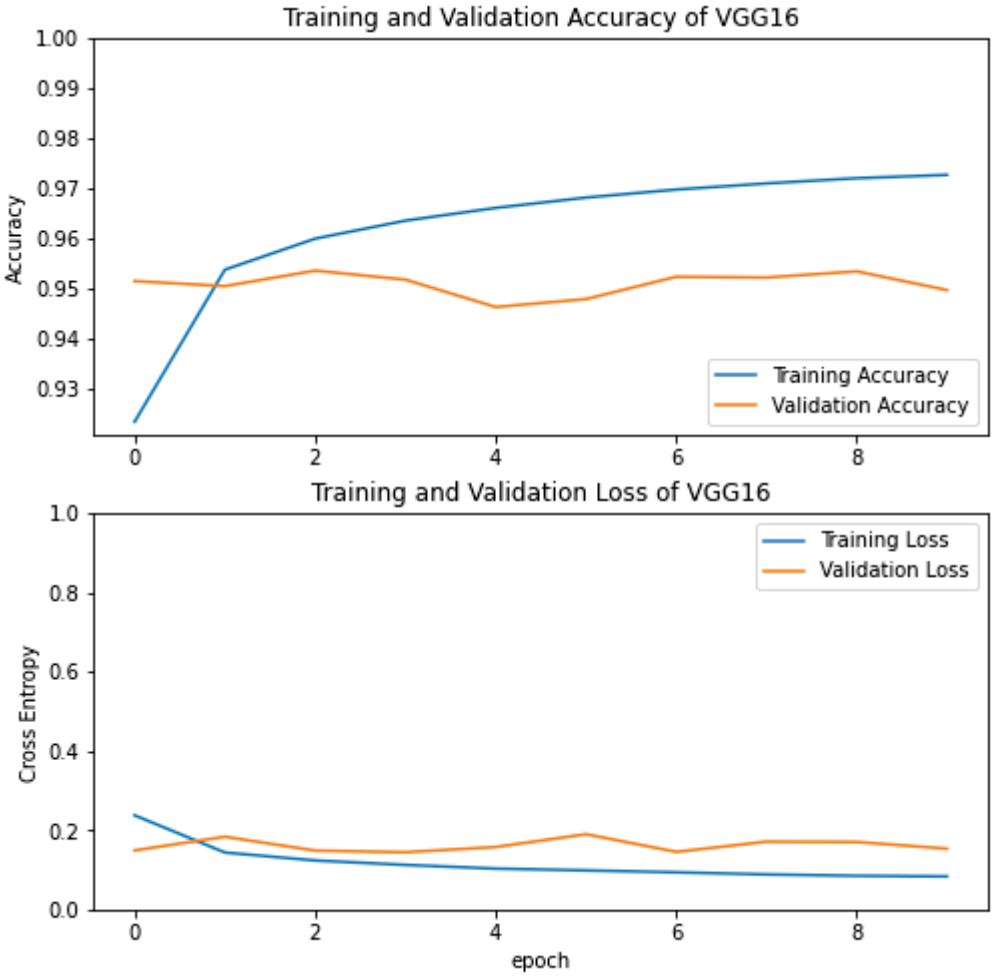

Figure 5.4: Training and Validation's Accuracy and Loss of VGG16

The figure 5.4 represents that training accuracy is increasing over iteration whereas validation accuracy has gone downwards once in 5th iteration. Therefore it started increasing from the next iteration.and validation accuracy started between 94% to 95%. The training loss was decreasing continuously and validation loss was 0.1534 in the last iteration.

Therefore, Inception V3 performed best during training and validation period. But it's validation accuracy is fluctuating where the second best model VGG16 is performing quite well. It's validation is increasing with the training which means this model is learning well during training.

5.3 Test Accuracy Analysis from Confusion Matrix

Here we came up with the confusion matrix to analyze our test accuracy. Confusion matrix is the best way to examine test performance which will show how well our model is performing with our test data. It will represent which data belong to which class in the true label and our model's predicted label.

5.3.1 Confusion Matrix of Resnet50:

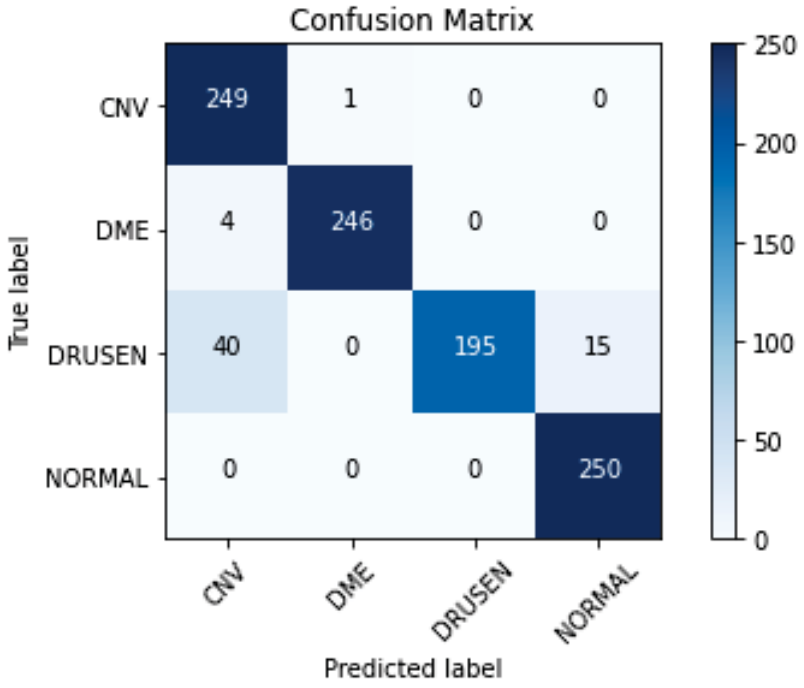

Figure 5.5: Confusion Matrix of Resnet50

Figure 5.5 represents the ResNet50 trained model can detect 249, 246, 195, 250 correct images of CNV, DME, DRUSEN and NORMAL respectively from every 250 images per class which provides 99.6%, 98.4%, 78% and 100% test accuracy.

5.3.2 Confusion Matrix of Inception V3

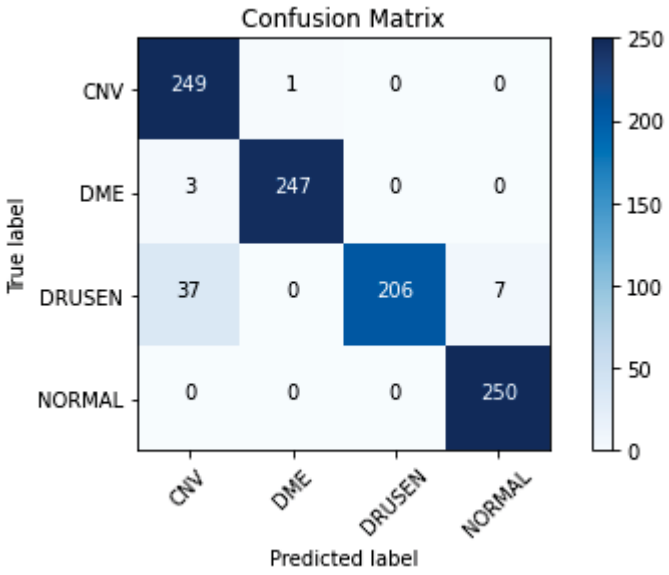

Figure 5.6: Confusion Matrix of Inception V3

Figure 5.6 shows tha Inception V3 trained model can detect 249, 247, 206, 250 correct images of CNV, DME, DRUSEN and NORMAL respectively from every 250 images per class which provides 99.6%, 98.8%, 82.4% and 100% test accuracy.

5.3.3 Confusion Matrix of Xception

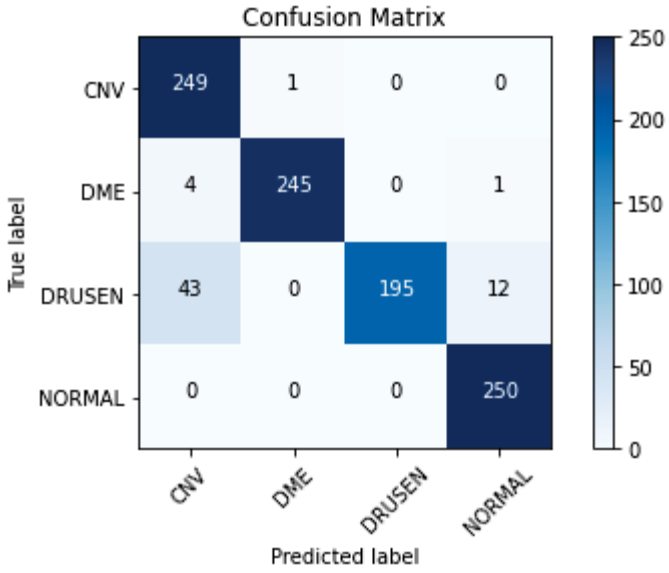

Figure 5.7: Confusion Matrix of Xception

Figure 5.7 represents the Xception trained model can detect 249, 245, 195, 250 correct images of CNV, DME, DRUSEN and NORMAL respectively from every 250 images per class which provides 99.6%, 98%, 78% and 100% test accuracy.

5.3.4 Confusion Matrix of VGG16:

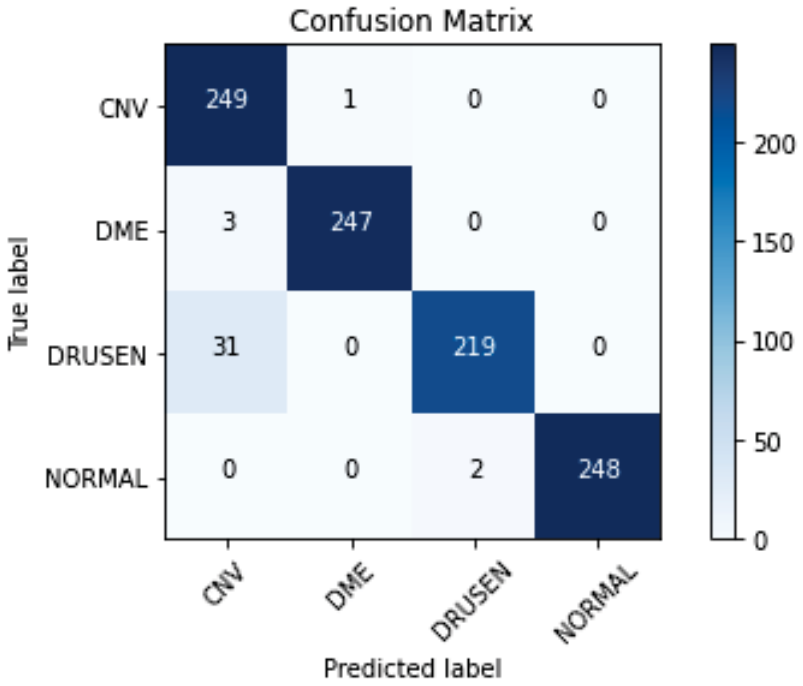

Figure 5.8: Confusion Matrix of VGG16

Figure 5.8 shows the VGG16 trained model can detect 249, 247, 219, 248 correct images of CNV, DME, DRUSEN and NORMAL respectively from every 250 images per class which provides 99.6%, 98.8%, 87.6% and 99.2% test accuracy.

Therefore, we can see that each and every model performs well in our dataset which is getting about 93 to 96 percentage accuracy. Here VGG16 and Inception V3 are performing best in the OCT dataset that are 95.40% and 95.20% respectively. Adding that, As we can see all the models are struggling with identifying DRUSEN because of fuzzy images. VGG16 is performing comparatively better in recognizing DRUSEN which has correctly identified 219 DRUSEN images.

5.4 Analysis of XAI

Finally, we come to our analytical section where we will analyze which part of an image was trained well in our model. By using previous version of OCT dataset, one of the research group got higher accuracy through CliqueNet[45]. But they did not analyzed their model. We have performed the VGG16 and InceptionV3 model to analyze our model through LRP-γ.

5.4.1 Analysis of CNV

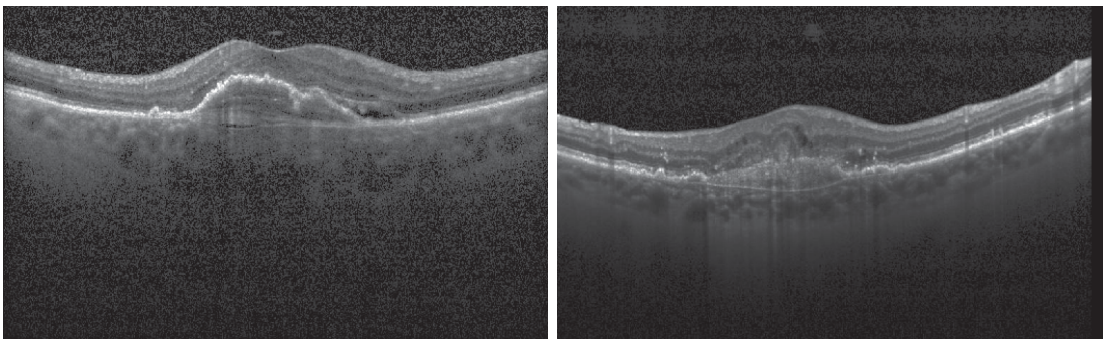

Figure 5.9: Input Images of CNV

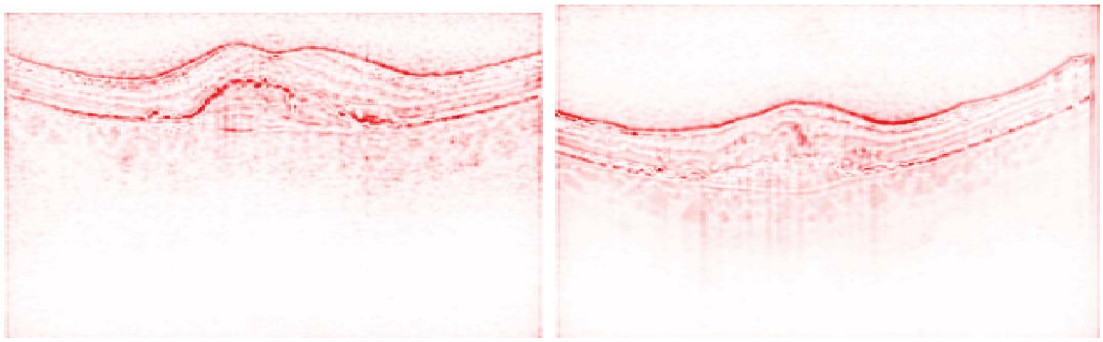

Figure 5.10: Analysis of CNV images from VGG16

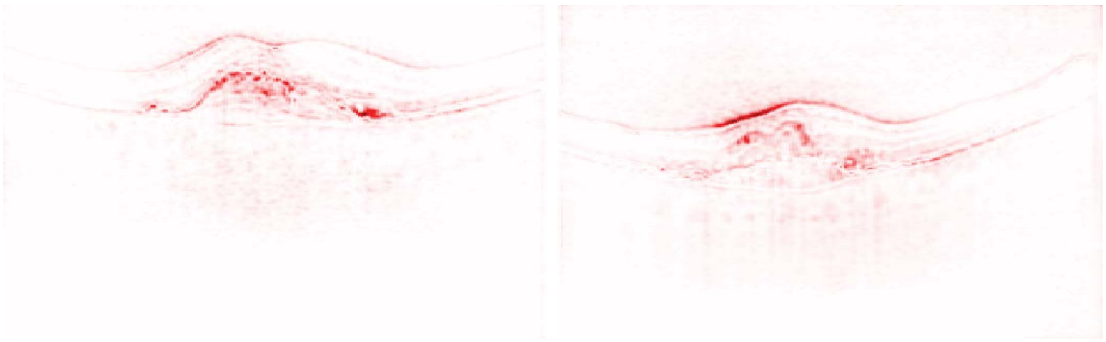

Figure 5.11: Analysis of CNV images from Inception V3

Here Figure 5.9 is input images of CNV, figure 5.10 is analyzed images from our trained model VGG16 and figure 5.11 is analyzed images from our trained model Inception V3.

5.4.2 Analysis of DME

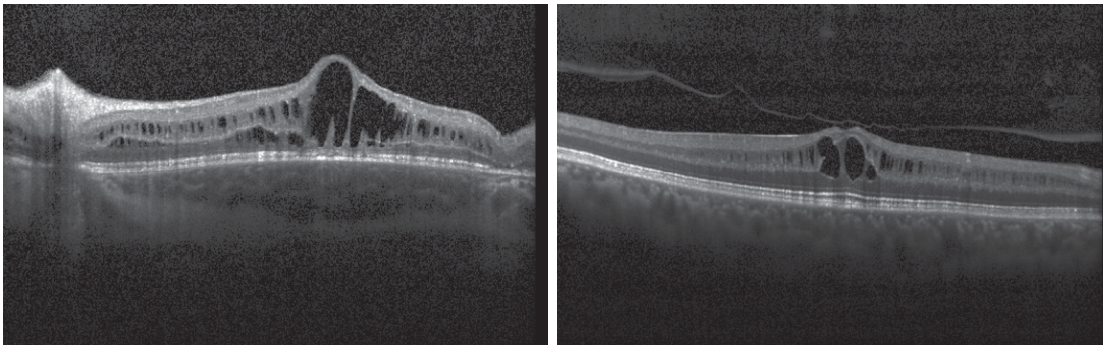

Figure 5.12: Input Images of DME

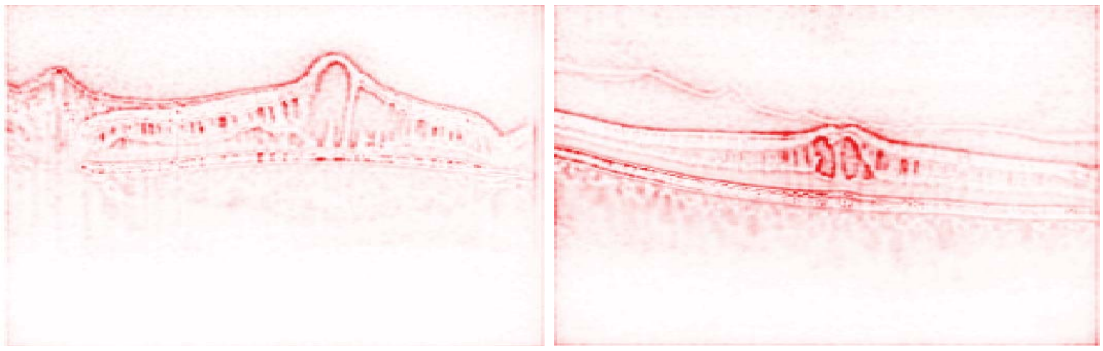

Figure 5.13: Analysis of DME images from VGG16

Figure 5.14: Analysis of DME images from Inception V3

The figure 5.12 is input images of DME, figure 5.13 is analyzed images from our trained model VGG16 and figure 5.14 is analyzed images from our trained model Inception V3.

5.4.3 Analysis of DRUSEN

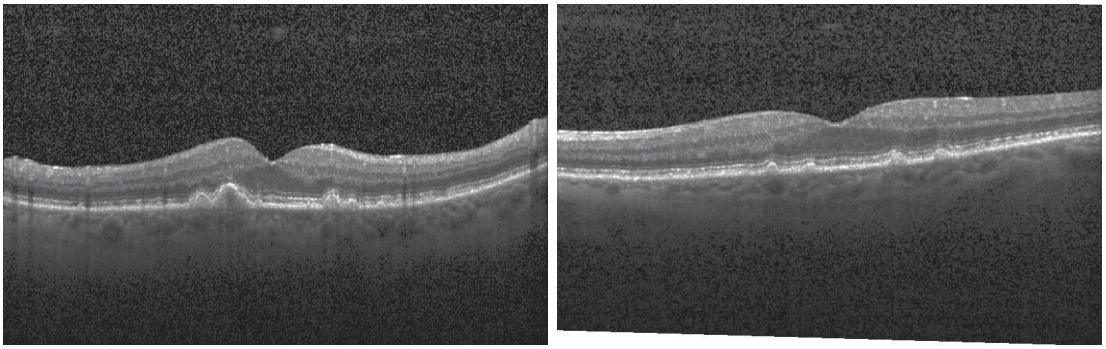

Figure 5.15: Input Images of DRUSEN

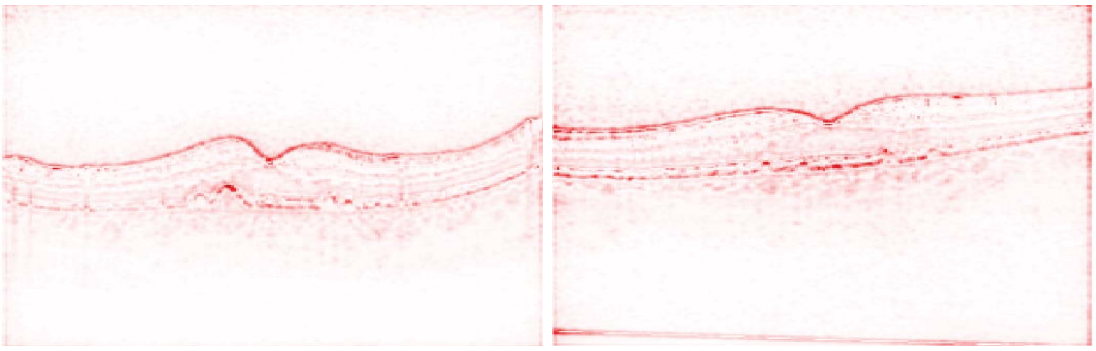

Figure 5.16: Analysis of DRUSEN images from VGG16

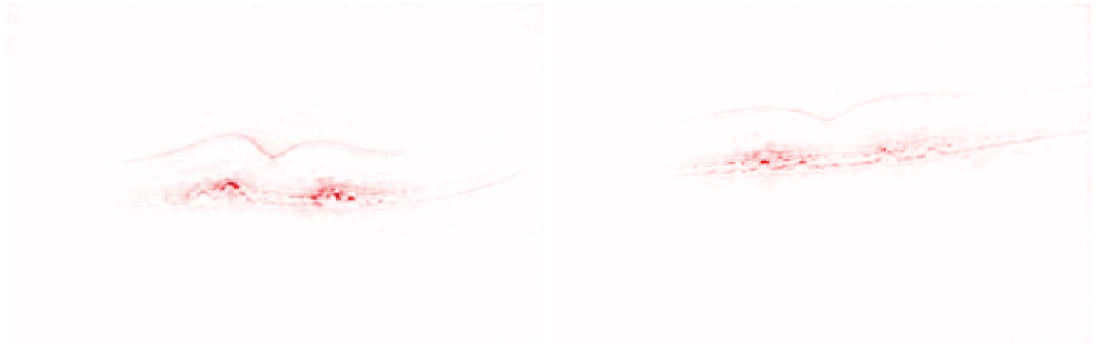

Figure 5.17: Analysis of DRUSEN images from Inception V3

Here, Figure 5.15 is input images of DRUSEN, figure 5.16 is analyzed images from our trained model VGG16 and figure 5.17 is analyzed images from our trained model Inception V3.

5.4.4 Analysis of Normal

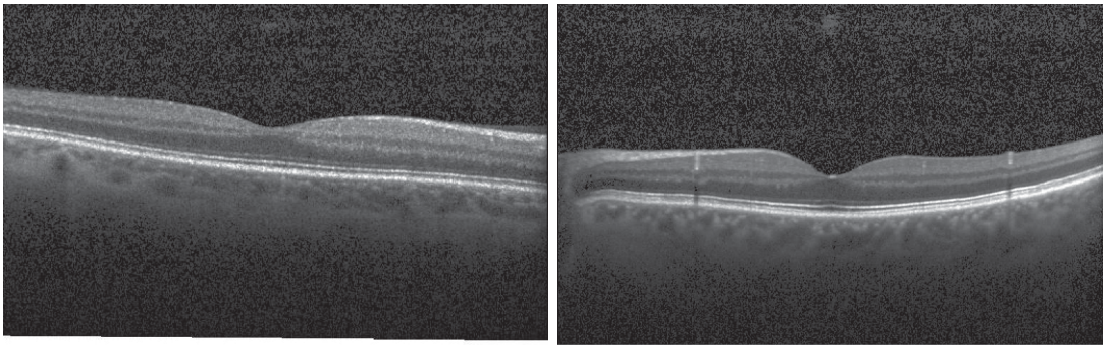

Figure 5.18: Input Images of NORMAL

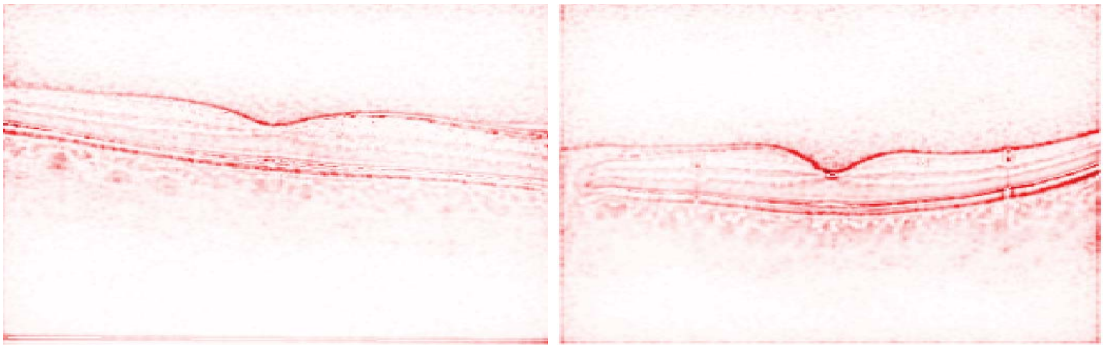

Figure 5.19: Analysis of NORMAL images from VGG16

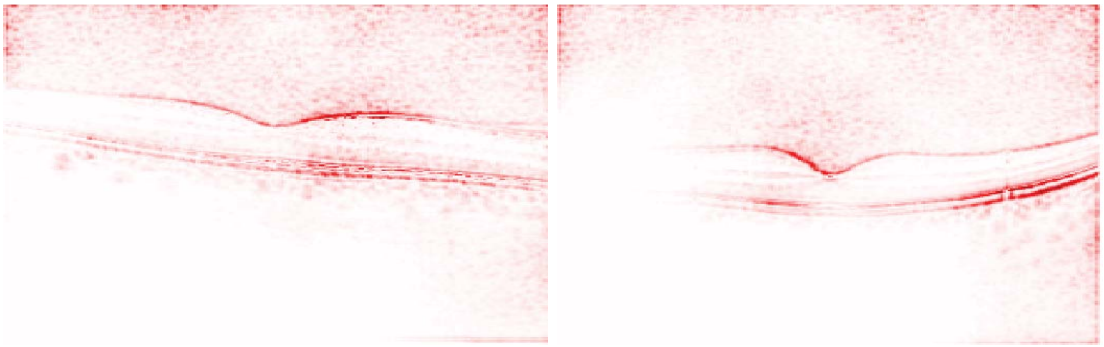

Figure 5.20: Analysis of NORMAL images from Inception V3

The figure 5.18 is input images of NORMAL, Figure 5.19 is analyzed images from our trained model VGG16 and Figure 5.20 is analyzed images from our trained model Inception V3.

Therefore, we can see that VGG16 performed well to analyze the key part of training features. By comparing the analysed image with the input we can understand that the VGG16 have focused on the whole shape of the retina and Inception V3 have focused on the basic part of the retina to classify .

5.4.5 Comparison Between Analysis methods of XAI

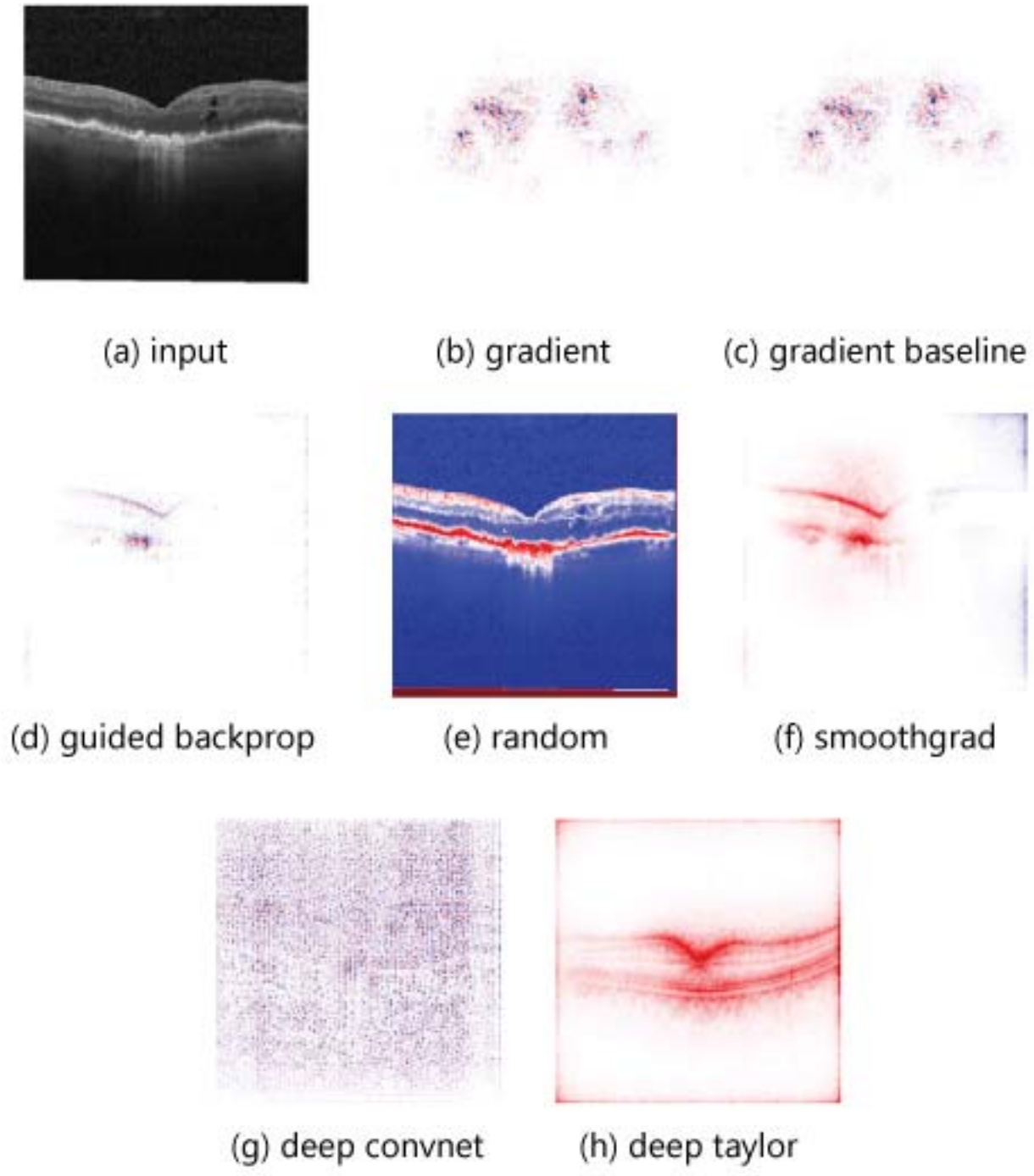

Figure 5.21: Comparison between Analysis methods of XAI

From Figure 5.21, we can clearly understand that the Deep Taylor method performs better than other methods in this particular data set for explanation. Which is why we have used the Deep Taylor method for analyzing the neural networks throughout our entire research.

Hence, our proposed model got lower accuracy than the other research group from [45] but we explained our model through Explainable AI's methodology $LRP - \gamma$. There is a global observation that more accuracy comes with less explanation and less accuracy comes with more explanation, so our research is not an exception of that.

Chapter 6

Conclusion

The main purpose of this work is to identify retinal disease with more precision of accuracy and explainability and this goal reaches towards the acceptance of machines while detecting retinal disease over human interaction. It can be said that with the developing era of machine domination, machine learning has the potential to be more accurate and make the life easy of humans. With the motive in mind we tried to overcome the problems of the works that were done previously on retinal disease identification using image classification. Most of the work that has been done with deep learning methodologies but there was not much explanation about the black boxes of those models that underlines the drawbacks of detecting the diseases. To cut down that drawback of lack of explainability of those deep neural methodology. We merged our work by using deep neural models with the implementation of XAI. Here the deep neural networks that we used Xception, VGG16, RestNet50, Inception v3 validates the accuracy of our model and LRP(XAI) explains the unknown predictors that have been passed through as an output by DNNS.

Though we have been successful implementing DNNS with XAI and getting almost the desired outcome of retinal disease identification, we underline drawbacks here as well as it failed to classify some of the DRUSEN images properly.

In the future, we want to implement federated learning through our proposed model, as patients' medical images are a matter of privacy to themselves, so we want to maintain their privacy and implement a federated learning model in the longer term.

Bibliography

[1] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Transactions on systems, man, and cybernetics*, no. 6, pp. 610–621, 1973.

[2] P. Szolovits, R. S. Patil, and W. B. Schwartz, "Artificial intelligence in medical diagnosis," *Annals of internal medicine*, vol. 108, no. 1, pp. 80–87, 1988.

[3] J. Carreira, H. Madeira, and J. G. Silva, "Xception: A technique for the experimental evaluation of dependability in modern computers," *IEEE Transactions on Software Engineering*, vol. 24, no. 2, pp. 125–136, 1998.

[4] A. Abdelsalam, L. Del Priore, and M. A. Zarbin, "Drusen in age-related macular degeneration: Pathogenesis, natural course, and laser photocoagulation–induced regression," *Survey of ophthalmology*, vol. 44, no. 1, pp. 1–29, 1999.

[5] A. G. Podoleanu, J. A. Rogers, D. A. Jackson, and S. Dunne, "Three dimensional oct images from retina and skin," *Optics Express*, vol. 7, no. 9, pp. 292–298, 2000.

[6] A. F. Fercher, W. Drexler, C. K. Hitzenberger, and T. Lasser, "Optical coherence tomography-principles and applications," *Reports on progress in physics*, vol. 66, no. 2, p. 239, 2003.

[7] ——, "Optical coherence tomography-principles and applications," *Reports on progress in physics*, vol. 66, no. 2, p. 239, 2003.

[8] G. J. Jaffe and J. Caprioli, "Optical coherence tomography to detect and manage retinal disease and glaucoma," *American journal of ophthalmology*, vol. 137, no. 1, pp. 156–169, 2004.

[9] G. Arden, R. Sidman, W. Arap, and R. Schlingemann, "Spare the rod and spoil the eye," *British journal of ophthalmology*, vol. 89, no. 6, pp. 764–769, 2005.

[10] L. M. Chalupa and R. W. Williams, *Eye, retina, and visual system of the mouse.* Mit Press, 2008.

[11] F. E. Hirai, M. D. Knudtson, B. E. Klein, and R. Klein, "Clinically significant macular edema and survival in type 1 and type 2 diabetes," *American journal of ophthalmology*, vol. 145, no. 4, pp. 700–706, 2008.

[12] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, IGI global, 2010, pp. 242–264.

[13] D. Ciregan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *2012 IEEE conference on computer vision and pattern recognition*, IEEE, 2012, pp. 3642–3649.

[14] S. G. Jarrett and M. E. Boulton, "Consequences of oxidative stress in age-related macular degeneration," *Molecular aspects of medicine*, vol. 33, no. 4, pp. 399–417, 2012.

[15] D. Pascolini and S. P. Mariotti, "Global estimates of visual impairment: 2010," *British Journal of Ophthalmology*, vol. 96, no. 5, pp. 614–618, 2012.

[16] T. Eisenberger, C. Neuhaus, A. O. Khan, C. Decker, M. N. Preising, C. Friedburg, A. Bieg, M. Gliem, P. C. Issa, F. G. Holz, *et al.*, "Increasing the yield in targeted next-generation sequencing by implicating cnv analysis, non-coding exons and the overall variant load: The example of retinal dystrophies," *PloS one*, vol. 8, no. 11, e78496, 2013.

[17] K. Da, "A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[18] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," *arXiv preprint arXiv:1404.2188*, 2014.

[19] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" In *Advances in neural information processing systems*, 2014, pp. 3320–3328.

[20] H.-P. Hammes, R. Welp, H.-P. Kempe, C. Wagner, E. Siegel, R. W. Holl, D. I.-G. B. C. Network, and D. Mellitus, "Risk factors for retinopathy and dme in type 2 diabetes—results from the german/austrian dpv database," *PloS one*, vol. 10, no. 7, e0132492, 2015.

[21] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458*, 2015.

[22] D. Castelvecchi, "Can we open the black box of ai?" *Nature News*, vol. 538, no. 7623, p. 20, 2016.

[23] S. Collins, "Bmj analysis calls medical errors third leading cause of death, shines new light on ongoing problem," *Pharmacy Today*, vol. 22, no. 7, pp. 36–37, 2016. DOI: 10.1016/j.ptdy.2016.06.022.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[25] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[26] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[27] C. S. Lee, D. M. Baughman, and A. Y. Lee, "Deep learning is effective for classifying normal versus age-related macular degeneration oct images," *Ophthalmology Retina*, vol. 1, no. 4, pp. 322–327, 2017.

[28] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep taylor decomposition," *Pattern Recognition*, vol. 65, pp. 211–222, 2017.

[29] D. C. Neely, K. J. Bray, C. E. Huisingh, M. E. Clark, G. McGwin, and C. Owsley, "Prevalence of undiagnosed age-related macular degeneration in primary eye care," *JAMA ophthalmology*, vol. 135, no. 6, pp. 570–575, 2017.

[30] E. Rezende, G. Ruppert, T. Carvalho, F. Ramos, and P. De Geus, "Malicious software classification using transfer learning of resnet-50 deep neural network," in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, 2017, pp. 1011–1014.

[31] I. Rocco, R. Arandjelovic, and J. Sivic, "Convolutional neural network architecture for geometric matching," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6148–6157.

[32] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, 2017.

[33] ——, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, 2017.

[34] N. Eladawi, M. M. Elmogy, M. Ghazal, O. Helmy, A. Aboelfetouh, A. Riad, S. Schaal, and A. El-Baz, "Classification of retinal diseases based on oct images," *Front Biosci*, vol. 23, pp. 247–264, 2018.

[35] N. Sharma, V. Jain, and A. Mishra, "An analysis of convolutional neural networks for image classification," *Procedia computer science*, vol. 132, pp. 377–384, 2018.

[36] M. Alber, S. Lapuschkin, P. Seegerer, M. Hägele, K. T. Schütt, G. Montavon, W. Samek, K.-R. Müller, S. Dähne, and P.-J. Kindermans, "Innvestigate neural networks!" *J. Mach. Learn. Res.*, vol. 20, no. 93, pp. 1–8, 2019.

[37] B. I. Dodo, Y. Li, D. Kaba, and X. Liu, "Retinal layer segmentation in optical coherence tomography images," *IEEE Access*, vol. 7, pp. 152 388–152 398, 2019.

[38] L. Fang, C. Wang, S. Li, H. Rabbani, X. Chen, and Z. Liu, "Attention to lesion: Lesion-aware convolutional neural network for retinal optical coherence tomography image classification," *IEEE transactions on medical imaging*, vol. 38, no. 8, pp. 1959–1970, 2019.

[39] A. Garcıa-Floriano, Á. Ferreira-Santiago, O. Camacho-Nieto, and C. Yáñez-Márquez, "A machine learning approach to medical image classification: Detecting age-related macular degeneration in fundus images," *Computers & Electrical Engineering*, vol. 75, pp. 218–229, 2019.

[40] 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.

[41] M. Mesquida, F. Drawnel, and S. Fauser, "The role of inflammation in diabetic eye disease," in *Seminars in Immunopathology*, Springer, 2019, pp. 1–19.

[42] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, *Explainable AI: interpreting, explaining and visualizing deep learning*. Springer Nature, 2019, vol. 11700.

[43] C. Wang, D. Chen, L. Hao, X. Liu, Y. Zeng, J. Chen, and G. Zhang, "Pulmonary image classification based on inception-v3 transfer learning model," *IEEE Access*, vol. 7, pp. 146 533–146 541, 2019.

[44] ——, "Pulmonary image classification based on inception-v3 transfer learning model," *IEEE Access*, vol. 7, pp. 146 533–146 541, 2019.

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

[46] A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (xai): A survey," *arXiv preprint arXiv:2006.11371*, 2020.

[47] D. Mojahed, R. S. Ha, P. Chang, Y. Gan, X. Yao, B. Angelini, H. Hibshoosh, B. Taback, and C. P. Hendon, "Fully automated postlumpectomy breast margin assessment utilizing convolutional neural network based optical coherence tomography image classification method," *Academic radiology*, vol. 27, no. 5, e81–e86, 2020.

[48] A. A. Saraiva, D. Santos, P. Pimentel, J. V. M. Sousa, N. M. F. Ferreira, J. d. E. B. Neto, S. Soares, and A. Valente, "Classification of optical coherence tomography using convolutional neural networks.," in *BIOINFORMATICS*, 2020, pp. 168–175.

[49] T. Schnake, O. Eberle, J. Lederer, S. Nakajima, K. T. Schütt, K.-R. Müller, and G. Montavon, "Xai for graphs: Explaining graph neural network predictions by identifying relevant walks," *arXiv preprint arXiv:2006.03589*, 2020.

[50] T. Ridnik, H. Lawen, A. Noy, E. Ben Baruch, G. Sharir, and I. Friedman, "Tresnet: High performance gpu-dedicated architecture," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Jan. 2021, pp. 1400–1409.

[51] usanne Dandl Christoph Molnar, "Counterfactual explanations," [Online]. Available: https://christophm.github.io/interpretable-ml-book/counterfactual.html.

[52] V. Dibia, "Ml interpretability: Lime and shap in prose and code," [Online]. Available: https://blog.cloudera.com/ml-interpretability-lime-and-shap-in-prose-and-code/.

[53] "Figure 2f from: Irimia r, gottschling m (2016) taxonomic revision of rochefortia sw. (ehretiaceae, boraginales).
biodiversity data journal 4: E7720., doi=10.3897/bdj.4.e7720.figure2f,,"

[54] M. Gurucharan, "Basic cnn architecture: Explaining 5 layers of convolutional neural network," [Online]. Available: https://medium.com/techiepedia/binary-image-classifier-cnn-using-tensorflow-a3f5d6746697.