COMPACT CONVOLUTIONAL TRANSFORMER FOR CLASSIFICATION OF RETINAL DISEASES FROM OPTICAL COHERENCE TOMOGRAPHY IMAGES

BY

Jarin Tias Meraj
ID: 241-25-013

This Report Presented in Partial Fulfillment of the Requirements for
The Degree of Master of Science in Computer Science and Engineering

Supervised By

Dr. Abdus Sattar
Associate Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

MAY, 2025

APPROVAL

This Project/Thesis titled **"Compact Convolutional Transformer for Classification of Retinal Diseases from Optical Coherence Tomography Images"**, submitted by **Jarin Tias Meraj**, ID No: **241-25-013** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of **MSc. in Computer Science and Engineering** and approved as to its style and contents. The presentation has been held on **24-05-2025**.

BOARD OF EXAMINERS

Chairman

Prof. Dr. Sheak Rashed Haider Noori
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dr. Md. Zahid Hasan
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dr. Naznin Sultana
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Dr. Md. Zulfiker Mahmud
Professor
Department of Computer Science and Engineering
Jagannath University

DECLARATION

I hereby declare that this research has been done by me under the supervision of **Abdus Sattar, Associate Professor, Department of CSE,** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Dr. Abdus Sattar
Associate Professor
Department of CSE
Daffodil International University

Submitted by:

MERAJ

Jarin Tias Meraj
ID: 241-25-013
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, I express my heartiest thanks and gratefulness to Almighty Allah for His divine blessing which makes it possible to complete the final year project/internship successfully.

I am really grateful and wish my profound indebtedness to **Abdus Sattar, Associate Professor**, Department of CSE, Daffodil International University, Dhaka, deep knowledge & keen interest of my supervisor in the field of Machine Learning to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stages have made it possible to complete this project.

I would like to express my heartiest gratitude to **Dr. Sheak Rashed Haider Noori**, **Head,** Department of CSE, for his kind help to finish our project and also to other faculty members and the staff of CSE department of Daffodil International University.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

ABSTRACT

In this study, we focus on how to use advanced deep learning models to identify retinal disorders from OCT images automatically. To increase the accuracy and speed of clinical diagnostics, the study introduces the Compact Convolutional Transformer (CCT) which is a transformer-based model that is both lightweight and performs well. A large OCT dataset made up of more than 130,000 images is used in the study to detect CNV, DME, Drusen and Normal conditions. A key part of our method is a data preparation process that uses image rotation, flipping and zooming to ensure classes are balanced. A 32×32 pixel image size was used for classification to lessen the computing burden without affecting how accurately the diagnosis was made. After extensive testing, the suggested model was matched against a standard Vision Transformer (ViT) model and a group of eight widely used transfer learning models, including DenseNet121, ResNet50 and EfficientNetB1. The model's effectiveness was tested by performing evaluation with several measures, including precision, recall, F1 score and training time. The study showed that the CCT model is more effective and resilient than the baseline and transfer learning models in medical picture classification tasks. This study highlights how transformer-based models, in particular CCT, can be included into AI-assisted retinal diagnostic systems in real time.

Keywords: CNV, DME, Drusen, Compact Convolutional Transformer, Vision Transformer, Convolutional Neural Network, Optical Coherence Tomography

.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Introduction

Globally, retinal disorders include Drusen, Diabetic Macular Oedema, and Choroidal Neovascularization (CNV) are among the main causes of blindness and visual impairment. Effective therapy and vision preservation depend on early and precise diagnosis of these disorders. High-resolution cross-sectional images of the retina are provided by optical coherence tomography (OCT), a potent non-invasive imaging method that makes it possible to identify even the smallest pathological alterations. However, manual OCT scan interpretation takes a lot of time, is prone to human error, and requires specific knowledge, which makes it challenging to scale in clinical settings with limited resources or high demand.

Deep learning has been particularly impactful in automating medical image processing as artificial intelligence advances quickly. Convolutional Neural Networks (CNNs) have proven to be effective in classification tasks, and more recently, transformer-based architectures have enhanced models' capacity to extract contextual information and long-range dependencies from images. This study suggests classifying OCT pictures into four groups using a Compact Convolutional Transformer (CCT): Normal, Drusen, DME, and CNV. Even if the images are not very clear, the CCT model is able to solve the problem by merging transformer and convolutional approaches. Its effectiveness and efficiency in clinical applications are shown by comparing its performance to that of several well-known transfer learning architectures.

1.2 Motivation

Because retinal disorders are becoming more common worldwide and ophthalmic services are under strain, we clearly need fast, reliable and automated ways to diagnose these conditions. Looking at OCT images by hand often takes doctors a long time and involves a lot of uncertainty. Automated tools for classifying retinal disorders help doctors diagnose patients faster, more accurately and from a distance. The main target of this

project is to build an AI tool for doctors to guide them in choosing treatments and eases the procedure on public health services.

1.3 Rationale of the Study

The purpose of this study is to respond to how deep learning is growing in medical imaging and diagnosis. Although CNNs were traditionally the main choice for image classification, the arrival of transformer-based models brings a new way to handle visual data. This CCT model performs well with small images and continues to obtain high classification results. The study aims to assess CCT's ability to distinguish retinal illnesses and to compare its performance with present transfer learning models, helping to decide if it can be used in real-world diagnosis.

1.4 Research Questions

- RQ1: To what extent does the Compact Convolutional Transformer (CCT) model proficiently classify retinal disorders from OCT pictures in comparison to conventional deep learning models?
- RQ2: What effect does picture resolution exert on the performance and efficiency of transformer-based classification models in medical imaging?
- RQ3: Do data augmentation strategies enhance classification performance amid class imbalance in OCT datasets?

1.5 Expected Output

- Creation of a resilient CCT-based model for precise and efficient categorization of retinal disorders using OCT scans. A comparative performance evaluation of the CCT model relative to leading transfer learning architectures.
- An exhibition of diminished computational duration and intricacy via the utilization of low-resolution images.
- A feasible AI-driven solution to aid ophthalmologists in real-time diagnosis and decision-making.

- A fundamental framework for subsequent research into transformer-based models in ophthalmic image analysis.

1.6 Handling projects and monetary aspects

This research was conducted independently from any outside financial sources or sponsorships – whether from individuals or organizations. The entire project was developed and carried out using only the researcher's vision, personal funding, and commitment. Lack of funding did not present challenges to the research process, on the contrary, it created a level of independence that presented an unadulterated study that was not impacted by any outside forces. This lack of financial backing highlighted the researcher's commitment to contributing to the academic community without letting financial stipulations contribute to any outcomes.

1.7 Report Layout

Chapter 1 delineates the context, aims, and research enquiries of the study. Chapter 2 investigates significant obstacles in the classification of retinal illnesses and analyzes pertinent studies that have advanced deep learning applications. Chapter 3 delineates the proposed methodology, encompassing dataset preparation, model architecture, and training methodologies. Chapter 4 delineates the experimental findings and analyses performance metrics. Chapter 5 examines sustainability, societal and environmental effects, and ethical implications. Chapter 6 finishes the findings and suggests avenues for further investigation in this domain.

CHAPTER 2
BACKGROUND

2.1 Preliminaries

Artificial intelligence is making imaging technology more important for finding retinal diseases. OCT helps to find and classify retinal problems in the early stages. OCT provides detailed images of the retina which can detect issues related to Choroidal Neovascularization (CNV), Diabetic Macular Oedema (DME) and Drusen [1]. Thanks to deep learning, particularly CNNs and transformer models, analyzing medical pictures has become much better. They are able to choose the most important aspects of a photo to ensure correct labeling. Even so, these methods have some issues, including working with data that is not balanced, making sure the model works in various scenarios and decreasing the amount of computation needed. In this paper, a novel model called Compact Convolutional Transformer (CCT) is presented to address these issues. The CCT model does this by combining convolutional and Transformer layers which allows it to find features better and keep the model efficient. You can make your photos 32×32 without losing accuracy and it will process faster. Also, rotating, flipping and zooming data can be used to address the problem of class imbalance.

2.2 Similar works

Deep learning approaches to retinal disease detection have utilized a range of architectures. VGG16, ResNet and DenseNet CNN models have been widely used for deployment in classification of images from optical coherence tomography (OCT). While these architectures perform well, they often require both high-resolution input and significant long training times. For instance, the DenseNet121 and VGG19 model were trained for over two hours, and both at best performed comparably to newer networks. Vision Transformer and other transformer models have shown the ability to model the global context of visuals by leveraging self-attention, but vision transformer (ViT) architectures

generally require both extra data to model properly, as well as extra-long training times. The Compact Convolutional Transformer (CCT) architecture can overcome this additional requirement of computing resources and training time, as it utilizes tokenization and CNN methods to utilize lightweight transformer layers (2). As a result, the CCT can both achieve greater accuracy while using less data and computing resource as had been utilized previously by ViT architectures, CCT architecture achieved an accuracy score of 97.09%- compared ViT architectures only achieved 85.57% while multiple transfer learning models did not perform as well. In addition, the model also achieved a high F1-score of 96.87%. Furthermore, it also performed well in terms of precision (96.52%), recall (97.27%) and specificity (99.02%), confirming it is able to reliably and efficiently classify retinal OCT images.

2.3 The Problem's Scope

Identifying retinal conditions, for example CNV, DME and Drusen, with Optical Coherence Tomography (OCT) is important, but it is a lengthy process that depends on experts. Even though CNNs and transfer learning methods are commonly used in automating this process, they run into problems such as high computing costs, poor performance on datasets with unequal numbers of samples and limited ability to capture global information [1]. Usual CNNs excel in finding local features but often do not notice long-distance patterns, while transformer-based models which are powerful, generally need a lot of resources. Many current approaches disregard important evaluation criteria besides accuracy which are necessary for medical use. To address these issues, this study introduces a better version of the Compact Convolutional Transformer (CCT) that is more efficient, deals with class imbalance and performs well on all key metrics, making it suitable for real medical uses.

2.4 Challenges

It is challenging to classify retinal diseases in OCT images since the imbalance of images for each disease. The previous dataset had significantly fewer images of DME and Drusen than CNV or Normal images. We managed the problem with augmentations such as rotation, flipping and zooming. The use of the model can vary based on the environment, camera and patient features so it has to be robust. It has to generalize well with previous unseen data. So many traditional models had a problem over-learning the training dataset. The CCT model's hybrid design has helped it generalize much better than a lot of the evaluation tests. All these architectures require many resources. The CCT model's compact design allows it to make highly accurate evaluations after a fairly short training schedule.

CHAPTER 3
RESEARCH METHOLOGY

3.1 Description of the Dataset

The dataset that was utilized in this investigation was obtained from a Mendeley Data source that is accessible to the general public. The following four categories are used to classify the high-resolution greyscale OCT images that are included in this dataset: There are 37,206 images of choroidal neovascularization (CNV). 11,349 pictures of diabetic macular oedema (DME) come to mind. There are 8,617 pictures in Drusen. Image count of 51,140 (normal) These pictures are arranged in distinct folders according to the labels that are assigned to the classes, and every single picture is provided in JPEG format. This dataset was developed specifically for supervised classification tasks, which makes it an ideal candidate for deep learning experiments in the field of medical image diagnosis.

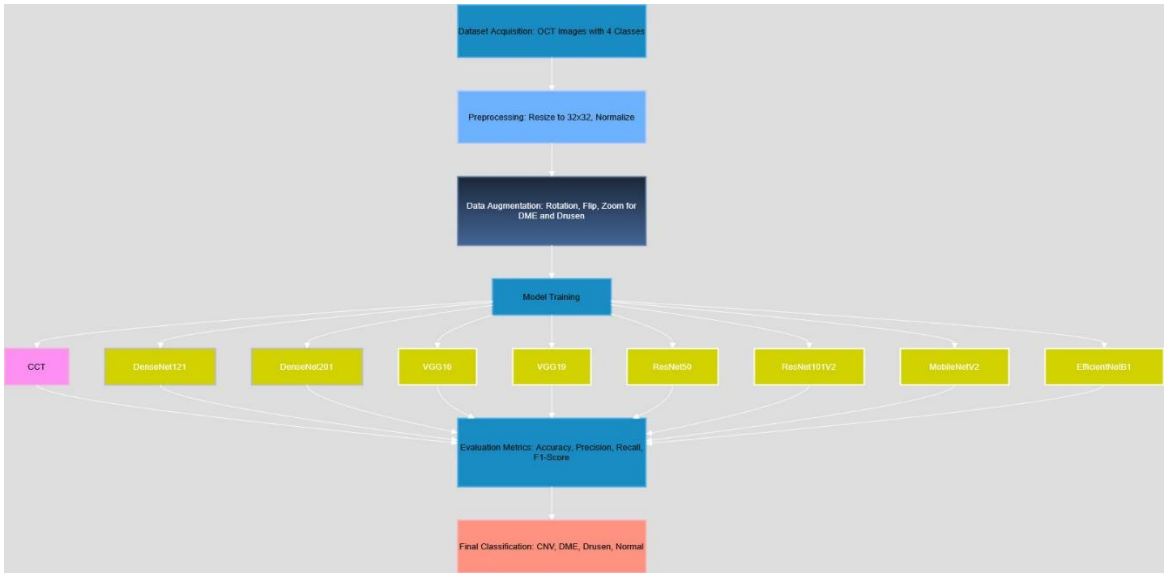


Fig. 3.1: Steps Diagram

3.2 Data Collection Procedure

To maintain uniformity in training and diminish computational complexity, various preprocessing measures were implemented: Resizing: All OCT images were scaled to 32×32 pixels utilizing bilinear interpolation. This down sampling maintains critical spatial

attributes while markedly decreasing memory usage and training duration. Normalization: Pixel intensity values were standardized to the range [0, 1] by dividing each pixel value by 255. This step guarantees that the neural network processes inputs on a uniform scale, thereby improving convergence during training. Class Imbalance Management: Due to the intrinsic imbalance within the dataset across various classes, augmentation techniques were judiciously employed for the minority classes (DME and Drusen) to establish a more equitable dataset. This aids in mitigating model bias and enhancing generalization across all categories.

3.3 Data Augmentation

The process of data augmentation is extremely important because it helps to improve the robustness of the model and increases the diversity of the training data. On the images of DME and Drusen, the following augmentation techniques were applied: Rotation (at random between -15 degrees and +15 degrees) Flipping in a horizontal direction the process of zooming (within a range of 0.9x to 1.1x) Using the augmentation strategy, the DME class was able to increase from 11,349 to 20,088 images, and the Drusen class was able to increase from 8,617 to 22,215 images.

Table 3.1: Impact of Data Augmentation on Class Distribution in OCT Dataset

Class	Original Count	Used for Augmentation	Augmented Count	Final Total
CNV	37,206	–	–	37,206
DME	11,349	2,000	8,739	20,088
Drusen	8,617	2,000	13,598	22,215
Normal	51,140	–	–	51,140
Total	**108,312**			**130,649**

This brought the DME and Drusen categories closer to aligning with the CNV and Normal categories.

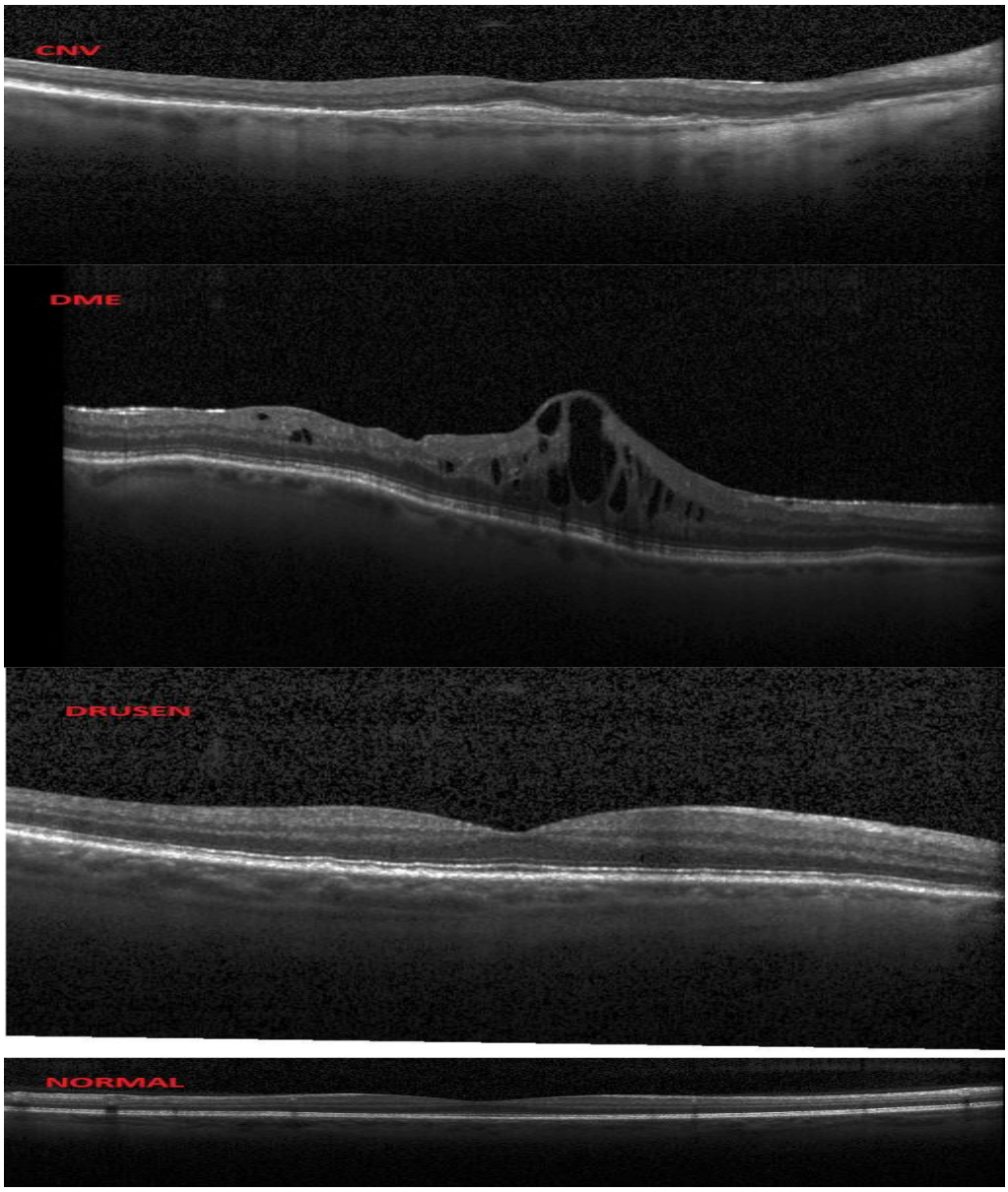


Fig. 3.2: Images of CNV, DME, DRUSEN and Normal

This augmentation not only helps the model generalize well on data that it has not previously seen, but it also helps the model balance the dataset by simulating variations that occur in the real world.

3.4 Deep Learning Models

This research developed model named **Compact Convolutional Transformer (CCT)** for classification.

3.4.1. CCT Architectural Reflection Model

The Compact Convolutional Transformer (CCT) incorporates aspects of convolutional neural networks and transformers to better classifiy medical imagery. The model resolves the issues of traditional ViT by implementing convolutional tokenization first and preserving local structure through the use of less data and positional information. Initially, local features and tokens from the image are generated with convolutional layers, then tokens are transformed in transformer blocks that use self-attention to understand global relationships. To primarily enhance generalization and robustness in lower-resolutions models, CCT altered positional encodings for sequence pooling. This is an ideal architecture, with a single class label, for classifying 32×32 OCT images can decrease the complexity of the model and improve the precision of image classification. CCT architecture is ideal for medical imaging such as retinal disease classification, because it is lightweight and can be effective at seeing miniscule visual details.
In addition, the hybrid CCT form keeps the main spatial details and also takes into account the global context, making it very flexible and important for clinical use in decision-making. Because it performs well with the least number of computations, it is very suitable for cases where computations are limited. Since CCT has helped classify retinal diseases with good results, it is even more useful for ophthalmologists and can improve the way diagnoses are made.

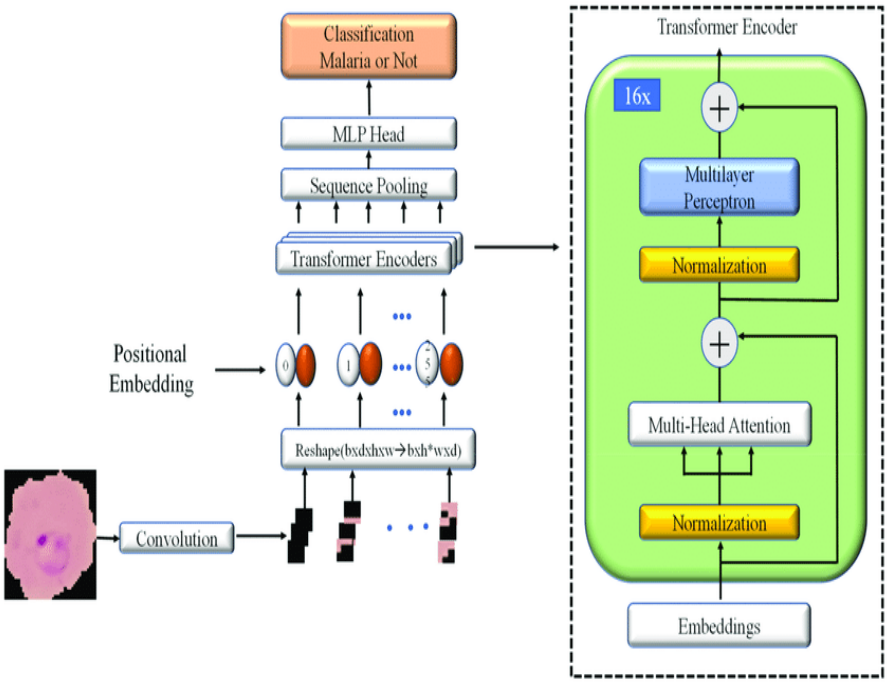


Fig. 3.3: Proposed CCT architecture

3.4.2 DenseNet121

DenseNet121 is an advanced convolutional neural network layout that comprises dense connections among the layers such that every layer can has inputs from all the previous layers while sending its feature maps to all the following layers. This complex connectivity structure promotes better information transfer and gradient flow within the network and reduces the vanishing gradient problem often observed in very deep networks. Dense connections also allow for feature reuse leading to better learning rates and less computing cost. The architecture has a number of dense blocks each of which is joined together by transition layers that perform convolution and pooling operations to control the dimensionality of feature maps. DenseNet121 is particularly good at medical image processing owing to the feature extraction component which is critical for picking up not only subtle but also complicated patterns of medical imaging. [3] Its ability to learn short

and unique representations makes it very suited for applications involving disease classification and anomaly detection in medical records.

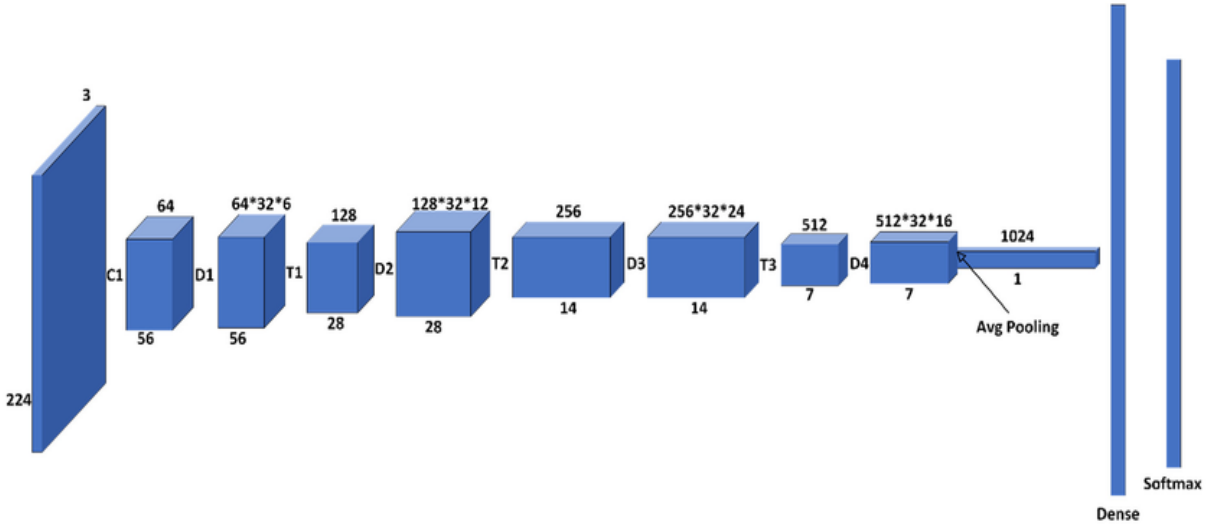


Fig. 3.4: Architecture of DenseNet121

3.4.3 VGG 19

Many folks use the VGG19 convolutional network architecture for it's ease of use and high accuracy for classifying images. The architectures from the Visual Geometry Group computing department at Oxford University have the same architecture and the same set of convolutional, max-pooling, and fully connected layers. While VGG19 has the same set of layers as VGG16, it has three additional convolutional layers so now has 19 weight parameters. Because it is deeper it thus learns representations from precented images much more efficiently, the VGG19 architecture is useful in more difficult detection tasks. VGG19 architecture occupies much more memory, but demonstrates the highest feature learning abilities and is typically used in the hardest deep computer vision tasks. Their conception is recognized as simple, as many adopt VGG16 and VGG19 architectures for image recognition and image transfer learning. VGG19 is an effective architecture that methodologically utilizes useful geometry-based thoughts, by an increased number of parametric weights. Convolutional architecture, layers, pooling architecture, layers,

activation functions and fully linked layers are all included.

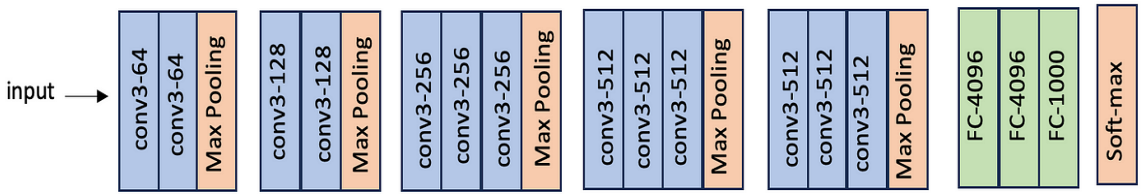


Fig. 3.5: Architecture of VGG19

3.4.4. DenseNet201

DenseNet201 is part of the Dense Convolutional Network family, working to ensure data and gradients can travel all over the network efficiently. In comparison to regular networks, DenseNet connects every layer to all the other layers in a simple feedforward structure. All layers draw feature maps from all previous layers which helps features be reused more often and avoids the problem of disappearing gradients. The deep structure and dense connections in DenseNet201[3] allow it to gather useful, distinctive features in less space. Such architecture helps the whole system operate more efficiently and successfully, mainly when the dataset is not very large, as is common in the medical imaging field. The structure of this architecture is well-suited to retinal disease categorization which requires high-level detail from images. The network's ability to share gradients better and more fully between layers helps detect fine patterns in OCT images, proving that DenseNet201 can handle medical image classification well in deep learning systems.

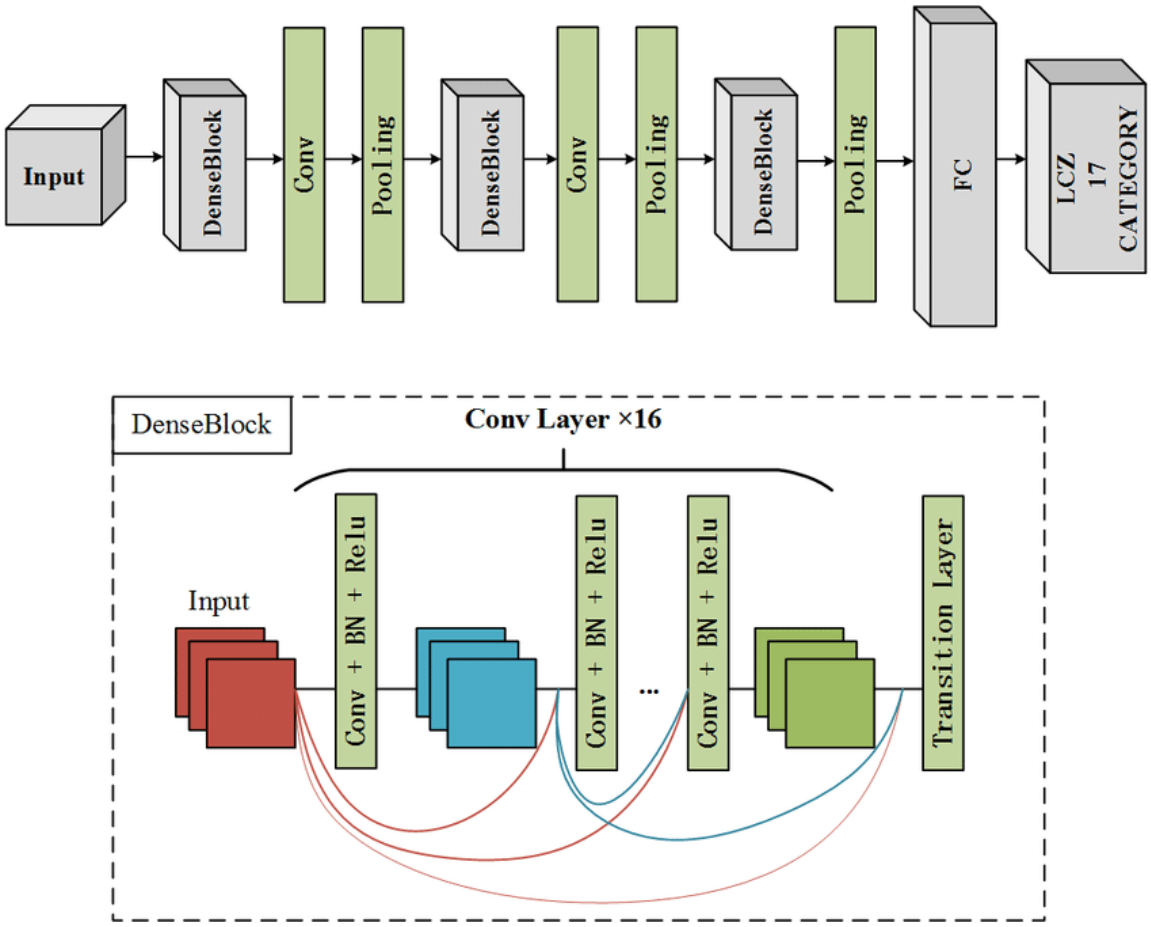


Fig. 3.6: Architecture of DenseNet201

3.4.5. ResNet50

It is a convolutional neural network within the group known as ResNet. Compared to other versions, this one has only 50 layers. Because residual connections are used, data can easily travel from one layer to another in the network, instead of facing the "vanishing gradient" issue seen in wider networks. A lot of image recognition tools such as those for object detection and picture classification, commonly select ResNet50. Even today, the model can be used effectively for multiple computer vision tasks. Deep learning experts and researchers admire this type of architecture thanks to residual connections that ensure it remains accurate as the network deepens.

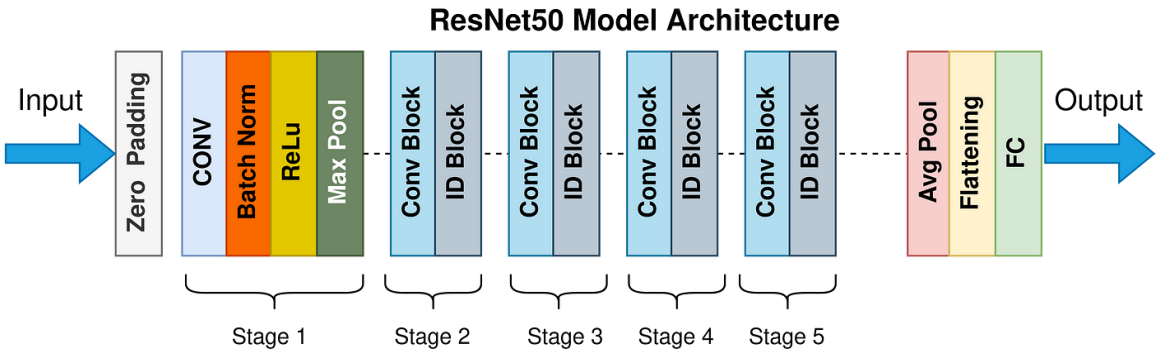


Fig. 3.7: Architecture of ResNet50

3.4.6. MobileNetV2

MobileNetV2 is a network architecture built to provide efficiency and accuracy in general computer vision on mobile devices and other embedded systems. MobileNet is improved by adding two key features: inverted residuals and linear bottlenecks. As a result, the number of parameters and computations drops and the model still works just as well. MobileNetV2 works well on limited-resource and fast-inference tasks in the medical picture classification field. Despite being small, it manages to record significant visual features which qualifies it for use in retinal OCT analysis. Its quick architecture lets it work well for real-time uses and edge devices which could mean faster and more accessible diagnostic tools for use in clinics.

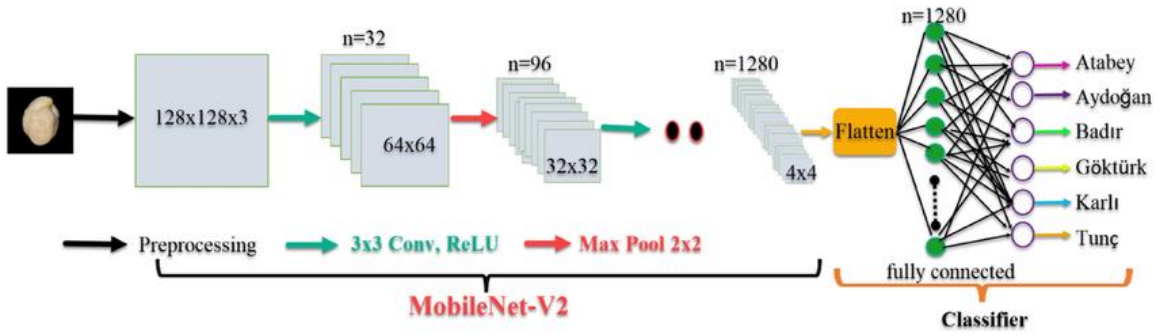


Fig. 3.8: Framework of MobileNetV2

3.4.7. ResNet101V2

ResNet101V2 achieves this after making the ResNet backbone more efficient and preventing the problems associated with vanishing gradients. Since the output of each layer is an input to the next, the model's performance improves due to dealing with the problem of degradation. It surpasses the original in detecting problems in retinal scans and performs better under training conditions. The model improves optimization on the residual connections and adds batch normalization, thus allowing the updates on ResNet101 referred to as ResNet101V2. The combination of propogating through each layer's details and image identification gives this model the ability to identify deviations such as in retinal scans. This variant, because of its detailed feature extraction, can perform complex tasks like image classification and is used in the detection of senile changes in medical photographs of the retina. Among other sophisticated models, ResNet101V2 still retains a smaller set of parameters which when compared grants it remarkable efficiency for tasks that requer precision without many computational resources.

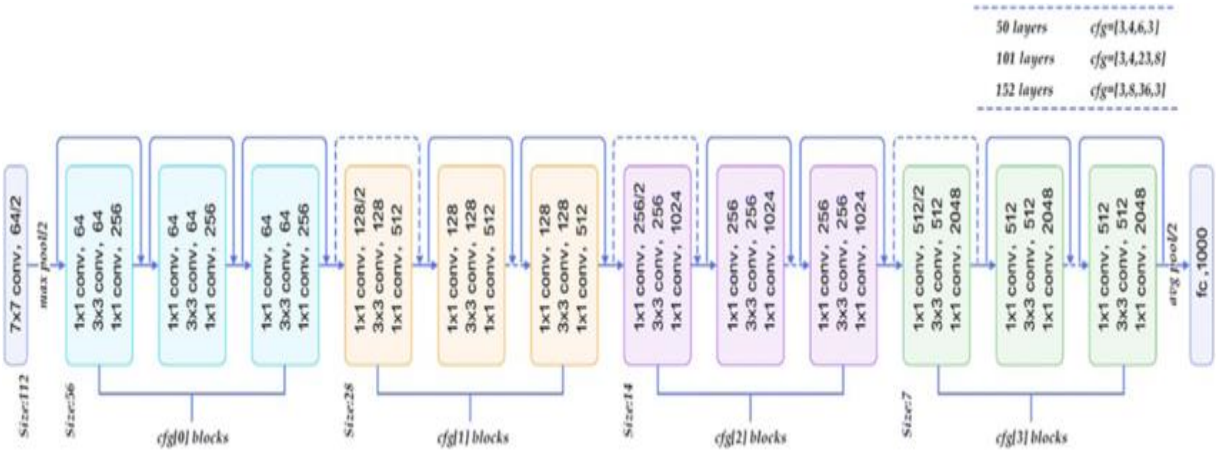


Fig. 3.9: Framework of ResNet101V2

3.4.8. VGG16

VGG16 is one of the convolutional neural networks architectures widely known for its strcuture depth and simpicity. Developed by the Visual Geometry Group of the University of Oxford, it consists of 16 weighted layers: 13 convolutional layers and 3 fully linked layers. Its distinctive mark is the usage of 3x3 small convolution filters which are organized in a way that helps to capture sophisticated features while avoiding high computational costs. Because of its simple and consistent architecture, VGG16 has become popular in medical imaging and numerous image classification problems. The model is quite efficient in recovering hierarchical characteristics ranging from edges and textures to intricate patterns in later stages. It is also more robust in parallel fashion than older models like AlexNet [6] but at the cost of requiring more intense computations and memory, which could be problematic in resource restrictive environments. Regardless, VGG16 stands out as a reliable choice for feature extraction and transfer learning tasks because it is consistent in its structure and performance.

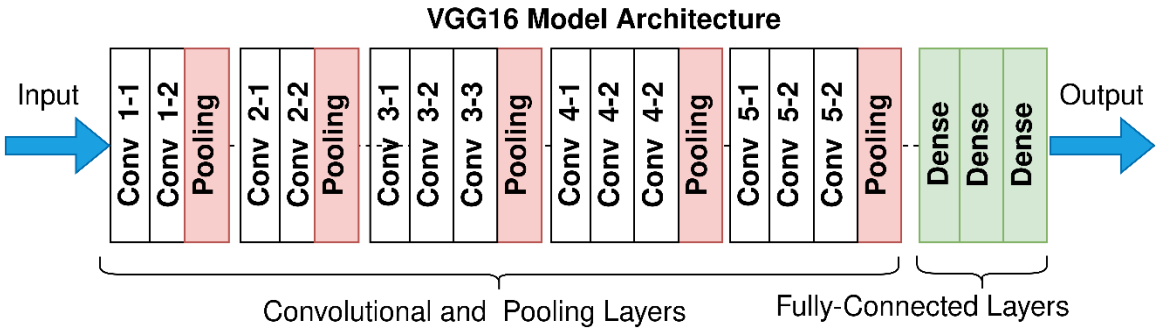


Fig. 3.10: Framework of VGG16

3.4.9. EfficientNetB1

EfficientNetB1 is a member of the EfficientNet family, a collection of convolutional neural networks created by Google that emphasize performance optimization with reduced parameters and diminished processing expenses. EfficientNetB1 enhances the foundational EfficientNetB0 model by proportionately increasing its depth, width, and resolution through a compound scaling approach. It utilizes MBConv blocks (mobile inverted

bottleneck convolutions) and squeeze-and-excitation optimization to improve speed while maintaining a lightweight design. EfficientNetB1 provides an optimal balance between speed and accuracy, rendering it appropriate for use in real-time or resource-constrained settings. Notwithstanding its tiny construction, it has exhibited robust performance in many picture classification applications, including medical image analysis.

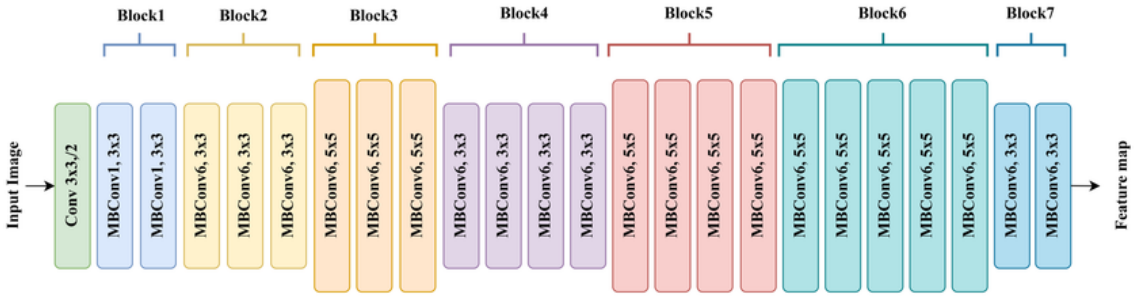


Fig. 3.11: Framework of EfficientNetB1

CHAPTER 4
EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Image Preprocessing

This study manually examined the impact of picture preprocessing techniques that, as the author has suggested, might enhance the data quality for training and evaluation of classification models. Optical Coherence Tomography, (OCT) images, as one of the major inputs in the study, underwent numerous preprocessing procedures for accuracy enhancement as focus feature extraction was performed to maximize accuracy. The exact steps of the treatment pipeline were: (1) Raw Image: The dataset contains original OCT images, or raw pictures, which are untouched. They contain complete structural information, but raw pictures almost always represent a low-quality view of the object that contains a large background noise and some artifacts that interfere hindering accurate classification. (a) Border Elimination: In order to concentrate the model's attention to the relevant retinal area of interest, the superfluous border surrounding the retinal OCT scan were also removed. That led to a lowering in input space complexity and retains the more characteristic center region of the scan. [7] (c) Erosion: Erosion morphological technique was applied to remove some of the decorating pixels in the tissues as well as some insignificant pixels located in the foreground where some noise imagine was superimposed. It also highlighted the details of the structures that forms a pixel like the tissues of the retina by enabling pixels in the background that do not represent the objects of interest to be gotten rid of. (d) Median Filtering: A median filtering was done after salt and pepper speckle noise has been added to the noise free image.

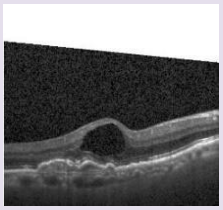
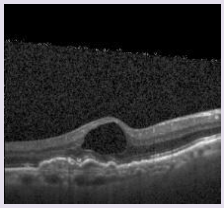
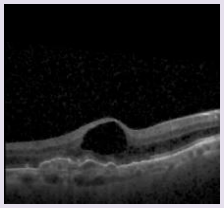
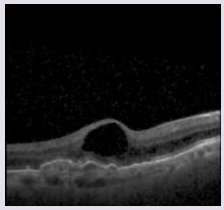
(a) Raw image	(b) After removing border	(c) Erosion	(d) Median
			

Figure 4.1: Image Preprocessing

4.2 Hyperparameter Impact Analysis

Ablation studies were done to optimize the model and enhance its efficacy in diagnosing retinal disorders from OCT pictures. This research examined the influence of several architectural and training elements on the model's accuracy and training efficiency. The findings are encapsulated below:

Study 1: Effect of Transformer Layers

Table 4.1: Effect of Transformer Layers

Config	Transformer Encoder Blocks	Parameters	Training Time	Accuracy	Findings
1	1	0.24M	19s	91.24%	Close to highest
2	2	0.41M	38s	91.41%	Moderate
3	3	0.57M	60s	91.53%	Highest Accuracy

Study 2: Effect of Kernel Size

Table 4.2: Effect of Kernel Size

Config	Kernel Size	Parameters	Training Time	Accuracy	Findings
1	1	0.17M	23s	87.30%	Close to high accuracy
2	2	0.20M	24s	92.12%	Close to highest accuracy
3	3	0.24M	19s	93.57%	Highest accuracy, lowest complexity

Study 3: Impact of Activation Functions

Table 4.3: Impact of Activation Functions

Config	Activation Function	Training Time	Accuracy	Findings
1	ReLU	19s	93.57%	Close to highest accuracy
2	ELU	19s	93.82%	Highest accuracy
3	Tanh	19s	93.37%	

Study 4: Pooling Layers

Table 4.4: Pooling Layers

Config	Pooling Type	Training Time	Accuracy	Findings
1	Max	19s	93.82%	Highest accuracy
2	Average	19s	92.55%	

Study 5: Comparison of Optimizers

Table 4.5: Comparison of Optimizers

Config	Optimizer	Training Time	Accuracy	Findings
1	Adam	19s	95.48%	Highest accuracy
2	Adamax	19s	94.17%	Accuracy improved
3	Nadam	19s	93.82%	Lower accuracy

Study 6: Learning Rate Analysis

Table 4.6: Learning Rate Analysis

Config	Learning Rate	Training Time	Accuracy	Findings
1	0.01	19s	84.06%	Lowest accuracy
2	0.001	19s	96.82%	Highest accuracy
3	0.006	19s	89.30%	

Study 7: Loss Function Comparison

Table 4.7: Loss Function Comparison

Config	Loss Function	Training Time	Accuracy	Findings
1	Categorical Cross entropy	19s	97.09%	Highest accuracy
2	Binary Crossentropy	19s	95.23%	Close to highest
3	Mean Squared Error	19s	95.39%	Close to highest

4.3 Evolution Methods

After segmentation, the segmented images were then forecasted using a Transfer Learning model. The evaluation uses the confusion matrix, including metrics such as accuracy, precision, recall, and F1 score. True positive (TP) values accurately reflect reality. False positives (FP) arise when erroneous results are inaccurately classified. The third variety, false negative (FN), transpires when an accurate number is erroneously classified as negative. TN and FN are the fourth and fifth choices. A true negative (TN) is a positive instance erroneously classified as negative. The fourth is true negative (TN)

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \text{.....(i)}$$

$$Precision = \frac{TP}{TP+FP} \text{.....(ii)}$$

$$Recall = \frac{TP}{TP+FN} \text{.....(iii)}$$

$$F1\ Score = 2 \times \frac{precision \times recall}{precision+recall} \text{.....(iv)}$$

4.4 Experimental Results & Analysis

This section provides a comparative investigation of the classification performance of the proposed CCT model in relation to the traditional transformer (ViT) and several transfer learning models on the segmented OCT retinal pictures. Table 4.1 delineates the ideal results regarding accuracy, precision, recall, and F1 score.

Table 4.8: Pinpointing the highest quality result produced by the transfer learning technique

Table: Performance Comparison of Classification Models				
Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**
CCT	**97.09**	96.52	97.27	96.87
ViT	85.57	82.19	86.24	83.64
DenseNet121	72.21	-	-	-

DenseNet201	66.48	-	-	-
ResNet50	55.95	-	-	-
MobileNetV2	53.79	-	-	-
ResNet101V2	53.74	-	-	-
VGG16	50.51	-	-	-
VGG19	41.11	-	-	-
EfficientNetB1	39.22	-	-	-

The CCT model achieved the highest classification accuracy of 97.09%, along with strong precision (96.52%) and recall (97.27%), resulting in a robust F1 Score of 96.87%. These metrics demonstrate the model's high reliability in identifying disease categories from OCT images, with balanced performance across all metrics. In contrast, ViT, although a transformer-based model, fell behind with an accuracy of 85.57%, precision of 82.19%, and recall of 86.24%. While still effective, it lacks the compact and efficient architectural advantages of CCT, as evident in both performance and training time.

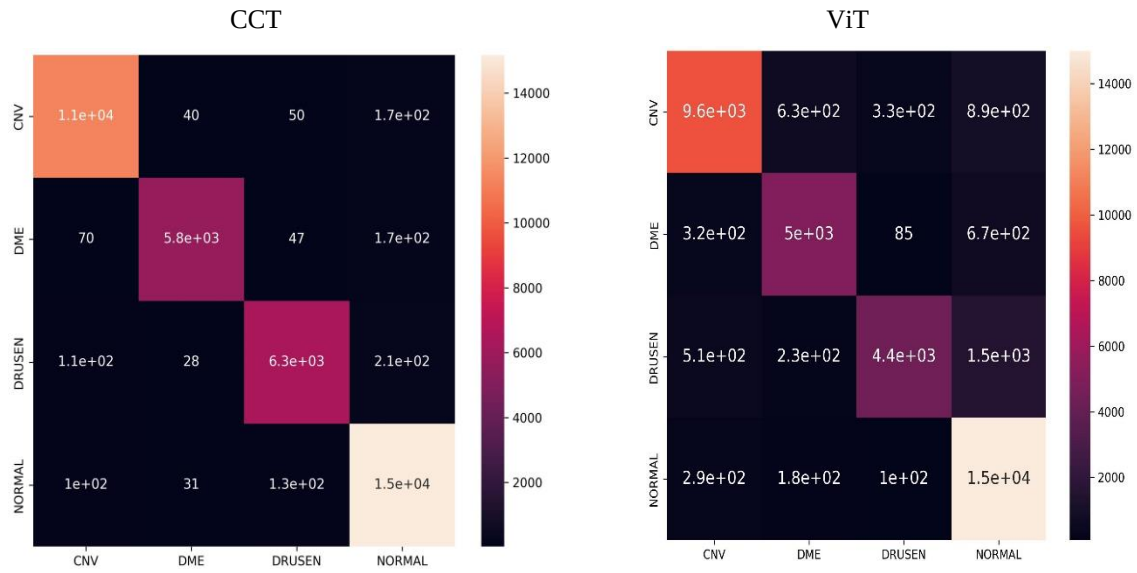


Fig 4.2. Confusion Matrix of the models' prediction

Among transfer learning models, DenseNet121 performed relatively better with an accuracy of 72.21%, while models like ResNet50, MobileNetV2, and EfficientNetB1 achieved significantly lower accuracy scores, indicating limitations in their generalization to the OCT dataset. These results highlight the superiority of the CCT architecture in handling complex visual features and ensuring effective classification. Furthermore, it showcases the importance of designing lightweight yet powerful architectures tailored to medical imaging datasets.

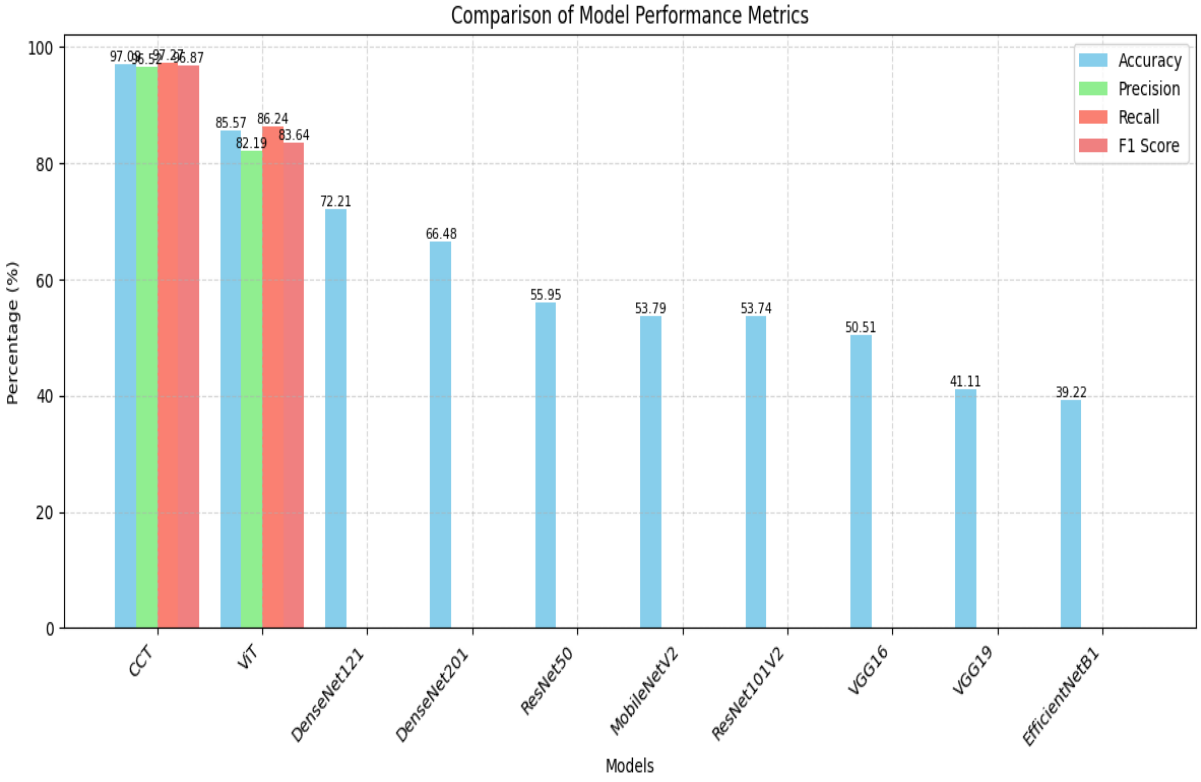


Fig 4.3. Comparison of the results between different models

The graph compares the performance of three convolutional neural network models: VGG16, ResNet152V2, and ResNet50, evaluating their Test Accuracy, Precision, Recall, and F1 Score, all expressed as percentages. Both VGG16 and ResNet152V2 demonstrate nearly identical performance, achieving close to 100% across all metrics. This indicates their greater proficiency in reliably to classify, identify relevant instances, and maintain a balance between precision and recall. In contrast, ResNet50, while still performing well,

shows slightly lower Test Accuracy and Precision, approximately 90% and 85% respectively. Moreover, it exhibits a more substantial decline in Recall and F1 Score, around 70% and 80%. This suggests that ResNet50, though robust, is not as proficient as VGG16 and ResNet152V2, particularly in correctly identifying all relevant instances, leading to a lower overall balance of precision and recall.

4.5 Discussion

The effectiveness and dependability of predictive models, especially in the classification of retinal diseases, depend on the accuracy of segmentation of the images—rotary imaging being one of the many modalities used in a clinical setting. Segmentation functions as an important prerequisite procedure which marks out a salient area ROI and hence guarantees minimal interference from less applicable areas. With effective segmentation, deep learning algorithms are able to extract clinically relevant features, such as shape, texture, and intensity more accurately. This is most important with OCT pictures where CNV, DME, and Drusen conditions have subtle differences. Accurate segmentation of the input image ensures maximum pattern recognition and hence optimum reduction of false positives and false negatives. In addition, correct segmentation increases feature extraction in the training phase and therefore strengthens the model's ability to adapt to different data. Training a classification model on distinctly segmented retinal layers builds the model's capability to classify similar-looking diseases distinctly, thus enhancing overall diagnostic accuracy. Segmentation also aids explainability for other visual representation frameworks like Grad-CAM. Ensuring effective segmentation means that the model’s activations

CHAPTER 5

Effects on Society, Ecosystems, and Sustainable Practices

5.1 Changes in societal dynamics

Employing deep learning for classifying retinal diseases through OCT (Optical Coherence Tomography) imaging has great promise in improving retinal healthcare. Early and accurate diagnosis of conditions like diabetic retinopathy, macular degeneration, and retinal edema can prevent vision loss and greatly enhance the quality of life for the patients. The application of automated AI diagnosis systems allows healthcare professionals to help reduce ophthalmologist's burden and provide prompt and accurate diagnostic services, especially in rural or isolated places with little access to specialists. These systems can effectively speed up the initial assessment of patients, optimize clinical activities, and improve tailored treatment strategies, thereby enhancing long-term results. To maintain fairness, correcting the inequities of technology access and guaranteeing equal advantage to all people regardless of group division is critical criterion.

5.2 Environmental Impact

The categorization of medical images, typically from the inspection and analysis of retinas based on optical coherence tomography (OCT), requires training through the use of 'deep learning' models which necessitates a lot of computation. This computational work usually requires the use of energy-hogging GPUs and a sizable number of images for analysis. These GPU-hunger workloads can supplement environmentally problematic global expansion such as the increase in energy consumption along with carbon dioxide emissions. However, solutions to the impact can be achieved through the utilization of sustainable practices such as establishing the energy efficiency of the model designs, using transfer learning will shorten the time to train with fewer iterations and eventually utilizing renewable energy to power the data centers.

5.3 Ethical Considerations

Using artificial intelligence (AI) for the assessment of retinal ailments has raised some important ethical questions. The patient imaging data must be protected to ensure confidentiality and privacy. Explainability in the model algorithms should be sufficiently recognized, particularly in medical situations where explainability affects treatment decisions and their reliance on clinician judgment. Reducing and avoiding bias in algorithms where accuracy differs between age groups, ethnicities, and strata is of equal importance. Building trust in automated systems requires validation, regulatory scrutiny, and audits of false predictions, but also adequate rationalization behind how and why decisions were made.

5.4 Sustainability Strategy

Sustainability in formally integrating deep learning into OCT-based retinal diagnosis includes the following factors: perpetual development and validation to enable continuous uptake of new clinical data, ongoing education of clinicians regarding how to read and apply AI tools in practice, energy efficient computing, and the collaboration between hospitals, AI technology developers, and policy makers that emphasize responsible and meaningful integration. Sustainability as a concept also means building AI systems that are clinically relevant, adaptable to the shifting composition of the patient population, and ethically grounded in the responsibilty to provide patients with appropriate level of care whilst maintaining an environmental responsibility.

CHAPTER 6

Closing Remarks and Future Investigations

6.1 Overview of the Research

The aim of this research was to implement deep learning methods such as convolutional neural networks (CNNs) and transformers to classify eye diseases from Optical Coherence Tomography (OCT) images. The models analyzed were several CNNs, including DenseNet, VGG, ResNet, MobileNet, EfficientNet, and also transformer models (ViT and CCT), all with pre-trained weights using transfer learning. The CCT performed the best in classifying OCT images, and outperformed the other models in all metrics, including accuracy, precision, recall, and f1 score. This study highlights the capability of transformer-based architectures to capture complex visual characteristics in order to produce accurate predictions in medical imaging.

6.2 Wrap-up

The findings of this study show the power of deep learning, namely the CCT and ViT models, for accurately classifying retinal diseases through OCT images. While standard CNN architectures such as ResNet and VGG showed moderate to poor performance, transformer-based methods showed significantly better performance as they discern more complex patterns and structures in OCT images. Nevertheless, these findings demonstrate the importance of advanced deep learning architectures in ophthalmology and can allow for early diagnosis, improve clinical decision making, and help specialists care for patients as automation can do the analysis.

6.3 Consequences for Subsequent Research

This study offers a number of intriguing avenues for future research. One avenue involves the use of multiple modalities (combining OCT with particular imaging surgeries or with

other metadata) that may assist in bolstering diagnostic quality and assisting with model interpretability and understanding. An important consideration is to enhance model explainability and interpretability, to provide transparency over decision making or justification for diagnosis with said models. This may include using tools like Grad-CAM, or SHAP, that seek to provide some trust and better adoption within clinical settings. Future research into small, computationally efficient models may help extend the capacity for real-time diagnostics, particularly in assessments without high-functioning computerized diagnostics. The potential for continued use of models in clinical practice with models that maintain learning systems in place, could show the accuracy and generalizability of the models in other populations or imaging modalities and technologies.

This work is not without limitations. The single dataset usage could limit the extent that the model can generalize its predictions beyond the training data., or data that is made available in a more diverse manner. In addition, due to biases in data collection, under-represented populations in the study there may therefore be an even larger performance gap between the conduit of academic performance and actual clinical performance. Future studies should address these limitations and further support the developing trust the robustness and fairness of the models with large, diverse, datasets and with some concurrent external clinical validation

REFERENCES

[1] N. D. Koseoglu, A. Grzybowski, and T. Y. A. Liu, "Deep learning applications to classification and detection of age-related macular degeneration on optical coherence tomography imaging: A review," *Ophthalmol. Ther.*, vol. 12, no. 5, pp. 2347–2359, Oct. 2023, doi: 10.1007/s40123-023-00775-0.

[2] A. Wali, Z. Suhail, S. Naz, and I. Younas, "An ensemble deep learning model for OCT image detection and classification," *ResearchGate*, Aug. 2024. [Online]. Available: https://www.researchgate.net/publication/384008411

[3] A. Butola et al., "Deep learning architecture 'LightOCT' for diagnostic decision support using optical coherence tomography images of biological samples," *Biomed. Opt. Express*, vol. 11, pp. 5017–5031, 2020.

[4] A. Abdi and A. M. Abdulazeez, "OCT images diagnosis based on deep learning: A review," *Indones. J. Comput. Sci.*, vol. 13, no. 1, pp. 114–130, Mar. 2024. [Online]. Available: https://www.researchgate.net/publication/379153820

[5] S. P. K. Karri, D. Chakraborty, and J. Chatterjee, "Transfer learning based classification of optical coherence tomography images with diabetic macular edema and dry age-related macular degeneration," *Biomed. Opt. Express*, vol. 8, no. 2, pp. 579–592, Feb. 2017, doi: 10.1364/BOE.8.000579.

[6] G. Litjens et al., "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, 2017, doi: 10.1016/j.media.2017.07.005.

[7] J. De Fauw et al., "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nat. Med.*, vol. 24, no. 9, pp. 1342–1350, 2018, doi: 10.1038/s41591-018-0107-6.

[8] T. Schlegl et al., "Fully automated detection of diabetic macular edema using deep learning on OCT volumes," *Invest. Ophthalmol. Vis. Sci.*, vol. 58, no. 8, p. 316, 2017, doi: 10.1167/iovs.17-21873.

[9] University Eye Clinic of Trieste, "OCT-based deep-learning models for the identification of retinal key signs," *Sci. Rep.*, vol. 13, no. 14628, 2023, doi: 10.1038/s41598-023-41362-4.

[10] Gachon University et al., "OCTNet: A modified multi-scale attention feature fusion network with InceptionV3," *Mathematics*, vol. 12, no. 19, p. 3003, 2024, doi: 10.3390/math12193003.

[11] J. De Fauw et al., "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nat. Med.*, vol. 24, no. 9, pp. 1342–1350, 2018, doi: 10.1038/s41591-018-0107-6.

[12] A. Jaiswal, N. Gianchandani, D. Singh, V. Kumar, and M. Kaur, "Classification of the COVID-19 infected patients using DenseNet201 based deep transfer learning," *J. Biomol. Struct. Dyn.*, vol. 38, pp. 1–8, 2020, doi: 10.1080/07391102.2020.1788642.

[13] G. Cao, Y. Wu, Z. Peng, Z. Zhou, and C. Dai, "Self-attention CNN for retinal layer segmentation in OCT," *Biomed. Opt. Express*, vol. 15, pp. 1605–1617, 2024.

[14] N. Jinsakul, C.-F. Tsai, C.-E. Tsai, and P. Wu, "Enhancement of deep learning in image classification performance using Xception with the Swish activation function for colorectal polyp preliminary screening," *Mathematics*, vol. 7, no. 12, p. 1170, 2019, doi: 10.3390/math7121170.

[15] W. Al Ayoubi, W. M. Shalash, and M. F. Abulkhair, "Diabetic retinopathy detection through deep learning techniques: A review," *Informatics Med. Unlocked*, vol. 20, p. 100377, 2020, doi: 10.1016/j.imu.2020.100377.

[16] P. Selvakumar and S. Hariganesh, "The performance analysis of edge detection algorithms for image processing," in *Proc. Int. Conf. Comput. Technol. Intell. Data Eng. (ICCTIDE)*, 2016, pp. 1–5, doi: 10.1109/ICCTIDE.2016.7725371.

[17] C. I. Suciu, A. Marginean, V.-I. Suciu, G. A. Muntean, and S. D. Nicoară, "Diabetic macular edema optical coherence tomography biomarkers detected with EfficientNetV2B1 and ConvNeXt," *Diagnostics*, vol. 14, no. 1, p. 76, Jan. 2024, doi: 10.3390/diagnostics14010076.

[18] O. Tan et al., "A hybrid deep learning classification of perimetric glaucoma using peripapillary nerve fiber layer reflectance and other OCT parameters from three anatomy regions," *arXiv preprint*, arXiv:2406.03663, Jun. 2024.

[19] C. Lam et al., "Performance of artificial intelligence in detecting diabetic macular edema from fundus photography and optical coherence tomography images: A systematic review and meta-analysis," *Diabetes Care*, vol. 47, no. 2, pp. 304–319, Feb. 2024, doi: 10.2337/dc23-0993.

[20] Y. Guo et al., "Automated segmentation of retinal fluid volumes from structural and angiographic optical coherence tomography using deep learning," *arXiv preprint*, arXiv:2006.02569, Jun. 2020.

[21] K. Chen, X. Yang, J. Na, and W. Wang, "Denoising, segmentation and volumetric rendering of optical coherence tomography angiography (OCTA) image using deep learning techniques: A review," *arXiv preprint*, arXiv:2502.14935, Feb. 2025.

[22] S. Akça et al., "Automated classification of choroidal neovascularization, diabetic macular edema, and drusen from retinal OCT images using vision transformers: A comparative study," *Lasers Med. Sci.*, vol. 39, p. 140, May 2024, doi: 10.1007/s10103-024-04089-w.

[23] V. Thanikachalam, K. Kabilan, and S. K. Erramchetty, "Optimized deep CNN for detection and classification of diabetic retinopathy and diabetic macular edema," *BMC Med. Imaging*, vol. 24, p. 227, Aug. 2024, doi: 10.1186/s12880-024-01406-1.

[24] K. Pradeep et al., "Artificial intelligence and hemodynamic studies in optical coherence tomography angiography for diabetic retinopathy evaluation: A review," *Proc. Inst. Mech. Eng. H*, vol. 238, no. 2, pp. 1–12, Feb. 2024, doi: 10.1177/09544119231213443.

[25] S. Kim et al., "Deep learning model based on 3D optical coherence tomography images for the automated detection of pathologic myopia," *Diagnostics*, vol. 12, no. 3, p. 742, Mar. 2022, doi: 10.3390/diagnostics12030742.

241-25-013

ORIGINALITY REPORT

19%	15%	7%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	9%
2	Submitted to Daffodil International University Student Paper	2%
3	Sadia Sultana Chowa, Md. Rahad Islam Bhuiyan, Israt Jahan Payel, Asif Karim et al. "A Low Complexity Efficient Deep Learning Model for Automated Retinal Disease Diagnosis", Journal of Healthcare Informatics Research, 2025 Publication	2%
4	123dok.com Internet Source	1%
5	www.mdpi.com Internet Source	<1%
6	arxiv.org Internet Source	<1%
7	www.scaler.com Internet Source	<1%
8	Submitted to Gitam University Student Paper	<1%
9	C. Haritoglou. "Paracentral scotomata: a new finding after vitrectomy for idiopathic macular hole", British Journal of Ophthalmology, 2001 Publication	<1%