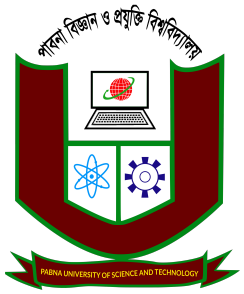
1in2in* 1in**

Performance Evaluation of Deep Learning Models for Detecting Visually Significant Cataracts Using Transfer Learning



Department of Computer Science and Engineering
Pabna University of Science and Technology, Pabna-6600

Course Title: Thesis
Course Code: CSE 4100 and CSE 4200

A thesis has been submitted to the Department of Computer Science and Engineering for the partial fulfillment of the requirement of B.Sc in Engineering in Computer Science and Engineering

Submitted By:
The Examinee of B. Sc Engineering Final Examination-2024
MD . Moshiur Rahman
Roll Number: 190117
Registration Number: 101833
Session: 2018-19

Supervised By:
Md. Shafiul Azam
Associate Professor
Department of Computer Science and Engineering
Pabna University of Science and Technology

Laboratory
Advanced Computer Lab, Department of Computer Science and Engineering
Pabna University of Science and Technology

January, 2025

DECLARATION

I declare that I have personally prepared this assignment. The work is my own, carried out personally by me unless otherwise stated and has not been generated using Artificial Intelligence tools unless specified as a clearly stated approved component of the assessment brief. All sources of information, including quotations, are acknowledged by means of the appropriate citations and references. I declare that this work has not gained credit previously for another module at this or another University.

I understand that plagiarism, collusion, copying another student and commissioning (which for the avoidance of doubt includes the use of essay mills and other paid for assessment writing services, as well as unattributed use of work generated by Artificial Intelligence tools) are regarded as offences against the University‚Äôs Assessment Regulations and may result in formal disciplinary proceedings.

I understand that by submitting this assessment, I declare myself fit to be able to undertake the assessment and accept the outcome of the assessment as valid.

Signature of the Examinee

CERTIFICATION

I am happy to certify that Md.Moshiur Rahman, Roll Number: 190117,Registration number: 101833, Session: 2018-2019, has completed a thesis work that enabled **"Performance Evaluation of Deep Learning Models for Detecting Visually Significant Cataracts Using Transfer Learning"**under my supervision to fulfill the requirements of the thesis course. This thesis was finished over the course of a year at the Department of Computer Science and Engineering, Pabna University of Science and Technology, Pabna-6600, Bangladesh.

According to my knowledge, this thesis paper has not been submitted elsewhere or replicated by another thesis paper before being submitted to the department.

Md. Shafiul Azam

Associate Professor,

Department of Computer Science and Engineering

Pabna University of Science and Technology, Pabna, Bangladesh

ACKNOWLEDGEMENT

First and foremost, I would like to express my gratitude to Allah (SWT) for granting me the strength and guidance to successfully complete my thesis. I would also like to extend my heartfelt appreciation to my respected supervisor, Md. Shafiul Azam, Assistant Professor, Department of Computer Science and Engineering (CSE), Pabna University of Science and Technology (PUST), for his intellectual guidance, invaluable support, encouragement, and insightful discussions throughout the completion of this thesis. I am deeply grateful for the opportunity to conduct this study under his supervision.

I am also thankful to the honorable Chairman, Dr. Md. Abdur Rahim, and all the esteemed faculty members of the Department of Computer Science and Engineering, Pabna University of Science and Technology, Pabna, Bangladesh, for their guidance and support during my thesis work.

Finally, I wish to express my sincere gratitude to my family, especially my parents, as well as to all my friends and well-wishers, for their unwavering support and inspiration.

January, 2025
Author

ABSTRACT

According to the World Health Organization report, one of the world's leading causes of blindness is cataracts. Even though cataracts majorly affect the elderly population, they can now be seen among minors too. Among the various types, three prominent types of cataracts affect large numbers of people: nuclear, cortical, and posterior subcapsular cataracts. Conventional methods of cataract diagnosis include slit-lamp image tests by doctors, which do not prove effective in classifying cataracts in the early stages and can also have inaccuracies in identifying the correct type of cataract. Existing work to automate the process has focused on classification based on binary detection only or has considered only one type of cataract among the mentioned types for further expanding the system. Furthermore, limited research has been done in the field of cataract type classification.

Our system works on the detection of cataracts and classification based on severity, namely mild, normal, and severe, in an attempt to reduce errors in manual detection of cataracts at early stages. The ODIR-5K dataset was utilized for training and validating our models, providing a robust foundation for our analysis. Our proposed system has successfully classified images as cataract-affected or normal with an accuracy of 97.35% using one of the CNN models, VGG19. Additionally, ResNet50 and Vision Transformer provided accuracies of 96.66% and 85.55%, respectively. Explainable AI (XAI) techniques were implemented, including LIME, SHAP, and Grad-CAM.

These methods were strategically used to ensure that the system's predictive accuracy is trustworthy and transparent, making it suitable for clinical applications.

Keywords: World Health Organization (WHO), Blindness Cataracts, ODIR-5K dataset, CNN models, VGG19, ResNet50, Vision Transformer, Accuracy, Explainable AI (XAI), LIME, SHAP, Grad-CAM, Clinical applications

TABLE OF CONTENTS

List of Tables

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

In this chapter, a review of our thesis has been clarified. For discussion convenience, there are a total of 4 sub-chapters under chapter 1, which we introduced. In section 1.1, we discussed the thesis Motivation; in section 1.2, we discussed the Research Questions and objectives; in subsection 1.2.1, we discussed Research question; In subsection 1.2.2, we discussed Objectives; in section 1.3, we discussed Organisation, in section 1.4, we include the discussion part about our thesis.

1.1 Research Justification

A cataract is a lenticular opacity that obscures the human eye's clear lens [1]. Light is usually converged onto the retina by the lens. Poor visual acuity results when this light is blocked and cannot reach the retina due to the cataract. It is the most common eye condition in the world, yet it does not impair vision at an early age. However, in people over 40, it can eventually impair vision and even lead to vision loss. Depending on its severity, early cataract detection can prevent blindness and avoid invasive, painful procedures. Approximately 285 million people worldwide suffer from visual impairment, according to the World Health Organization (WHO). The remaining individuals have poor vision, while 39 million have limited vision. Cataracts cause 51% of blindness and 33% of visual impairment. Early cataract detection is essential for treatment and can significantly lower the chance of blindness.

That is why a Computer-Aided System for Ocular Disease Recognition: Cataract Detection provides an opportunity for early cataract detection by using deep learning models. Three factors make using deep learning challenging: (i) the wide range of cataract lesions and human eye tones; (ii) the size, shape, and position of cataracts; and (iii) the dependency on age, gender, and eye type. Automatic cataract detection using various imaging modalities has been studied in recent years. Automatic cataract diagnosis and classification systems typically use one of four image types: fundus, ultrasonography, retro-illumination, or slit-lamp images. Given how simple it is for technologists or even patients to use a fundus camera, fundus images have garnered significant attention in this field among various imaging modalities. Slit-lamp cameras, on the other hand, must only be used by skilled ophthalmologists. As a result, timely treatments are delayed due to a shortage of qualified ophthalmologists, particularly in developing nations. Therefore, an automatic cataract diagnosis method based on fundus images is crucial to streamline the early cataract screening process[2].

Artificial intelligence-based cataract detection methods primarily rely on deep features (like deep CNN, which has attained better accuracy), local features (like local standard deviation), and global features (like discrete cosine transformation (DCT)). Despite the fact that several deep learning-based automatic cataract detection methods are available in the literature, they still have drawbacks, such as poor detection accuracy, a large number of model parameters, and significant computing costs.

The main contributions of this article are as follows: A cataract dataset is collected, reorganized, and preprocessed from the ODIR-5K standard dataset of Ocular Disease Recognition (EDA) published in the last two decades. It is then extended to a considerable number of images through the data augmentation process. To detect cataracts, a new 16-layer deep learning neural network, i.e., CataractNet, is proposed. The number of layers, activation functions, and loss functions are tuned to significantly improve detection accuracy. Five CNN models, i.e., VGG-19, Vision Transformer, and ResNet-50, have been implemented to compare and demonstrate the capability and accuracy of our proposed CataractNet.

1.2 Research Scope and Goals

1.2.1 Research Scope

Primary Question: " What methods can be used to improve the understandability of machine learning models in computer-based medical diagnosis systems to promote transparency and confidence?"
Secondary Question: "Can a comparative analysis of various convolutional neural networks (CNNs), in conjunction with explainable AI techniques like LIME, Grad-CAM and others, effectively address challenges 8 identified in prior studies, leading to improved accuracies and the establishment of a more reliable computer-aided diagnostic system for ocular diseases for potential clinical use?" [3]

1.2.2 Goals

Before writing the thesis, the following question comes up:

1. Construct distinct image classification models for each identified deep learning algorithm, encompassing various CNN architectures.
2. Conduct a comprehensive evaluation to assess the results and accuracies of these models.
3. Undertake an in-depth comparative analysis, contrasting the performance of each deep learning algorithm to identify respective strengths and weaknesses and obtain the best performance model.
4. Integrate LIME, GradCAM, and SHAP into the best-performing model to enhance interpretability.
5. Validate the outputs generated by XAI techniques for correctness and reliability.
6. Conduct a subsequent comparative analysis of the XAI techniques to determine superior accuracy and reliable validation [4].

The insights derived from these analyses will be pivotal in understanding how the implemented strategies contribute to transparency, trust, and the overall reliability of computer-aided diagnostic systems, significantly contributing to the refinement and advancement of diagnostic systems, particularly in the context of ocular eye diseases.

1.3 Thesis Organisation

Six sctions will make up the report, beginning with an introduction that provides a summary of the project's details, objectives, and research questions[1]. The background research section's literature review will be comprehensive, emphasizing the study's uniqueness and suggesting methods to apply the results to the dataset for improved outcomes. The study will next go into data collection and analysis methods, deep learning model implementation techniques, and a thorough description of each model's architecture. There will be descriptions of the XAI techniques

used. The implementation process, model accuracy, assessment criteria for the CNN architectures, and a discussion of the findings will all be included in the methodology section. The XAI research's visuals will be shown and examined. The implementation of the code will be evaluated before constraints are discussed. There will be a section on project management after this. The last section will acknowledge any limitations and offer a thorough assessment of the findings, viewpoints, and recommendations for further research. The code for the system, a project journal, and an ethical declaration form will all be included in the appendices.

1.4 Discussion

One of the main causes of blindness and visual impairment in the world, especially for people over 40, is cataracts. The disorder results from clouding of the lens, which makes it difficult for light to properly reach the retina [5] . Since early detection can avoid serious vision loss and lessen the need for intrusive procedures, it is essential for reducing the impact of cataracts.The use of artificial intelligence (AI) in medical diagnostics has opened up new approaches to addressing this problem in recent years. In particular, deep learning's ability to expedite early cataract identification has made it popular for application in ocular illness recognition. AI models can detect the existence and severity of cataracts with little assistance from humans by evaluating medical imaging, like eye fundus photos. This method allows for prompt diagnosis and treatment, which is especially advantageous in places with limited access to qualified ophthalmologists.Even while AI-based solutions seem promising, creating precise and effective models is extremely difficult. The detection method is made more difficult by variations in cataract presentation, variations in ocular anatomy, and demographic variables like age and gender. Furthermore, a lot of current models have drawbacks such excessive reliance on big datasets, poor accuracy, and high processing requirements.

Improvements in neural network topologies and picture preparation methods are being investigated to overcome these problems. Researchers hope to improve the accuracy and dependability of cataract detection systems by adjusting model parameters, refining data augmentation techniques, and utilizing reliable datasets. These developments could completely change the way cataracts are identified, increasing the effectiveness and accessibility of eye care globally.

CHAPTER 2

LITERATURE REVIEW

Deep learning architectures and a variety of classification techniques for computer-aided diagnostic systems have been extensively highlighted in earlier research related to the ODIR5K (Peking University, 2019) dataset[6]. Nevertheless, the examination of various XAI approaches and a comparative evaluation of interpretability in these investigations represent a clear void in the literature. Every study addresses the problem of class imbalance and offers solutions and recommendations.

2.1 Introduction

Khan et al. (2022) applied transfer learning, a machine-learning technique that uses information from one problem to tackle a related problem, and binary classification on the ODIR5K dataset [7].

(Khoshgoftaar, Wang, and Weiss, 2016). The model can generalize far better than if it was trained from start on the original tiny dataset by applying information from a related task with copious data. This makes it useful in situations when there is a limited amount of training data for the new task. Although a pre-trained VGG19 model is used to attain a 98.10% accuracy rate, the study failed to use explainable AI approaches, which made the model more difficult to interpret. The research proposed using generative adversarial networks (GANs) to create synthetic fundus images in order to further resolve the class imbalance and investigate image segmentation in order to further improve the accuracy in image classification of the study dataset. The imbalanced dataset was corrected using a combination of transfer learning and data augmentation. The biological characteristics of the patients in the study dataset were examined in another work (Hassan et al., 2023), which focused on age and gender and suggested FAG-Net for age and gender estimate using fundus images. Saliency maps were used in the study to establish causative regions and describe biological features with an accuracy of 91.87%. The study suggested applying more complex deep-learning models with attention processes and offered future recommendations for the use of saliency maps in medical image classification. A different study used both multiclass and multilabel classification algorithms to address the class imbalance in the study dataset. It generated discriminative feature attention maps with an accuracy of 96.08% by using an InceptionResNet architecture enhanced with a DKCNet block. Grad-CAM was utilized to improve the model's interpretability. To successfully lower the percentage of false positives, the images were subjected to a variety of artifacts, including "low-quality image," "optically invisible disk," "lens dust," and "image offset." A number of random sampling strategies were used to address the class disparity. The study recommended continued use of explainable AI methods for lesion diagnosis and localization, as well as the use of synthetic fundus image generators to address the class imbalance (Bhati et al., 2023). Using a pre-trained model that was initialized via transfer learning, EfficientNet was used for feature extraction through multi-label classification in the Wang et al. (2020) study. This model was combined with a specially designed multi-label classifier after its top layer was removed for improved feature extraction. Despite achieving a 73% accuracy rate, the study ignored interpretability. In the ODIR5K dataset, the 'O' label (signaling other diseases/abnormalities) includes a range of rare disorders, with some labels experiencing inadequate data availability, the researchers observed. This lack of data made it difficult to enhance the performance of the neural network model. The study's limitations were emphasized in the publication, which also recommended increasing the study's dataset size, especially for uncommon disorders, and adding demographic variables like age, gender, and family history to improve fundus disease identification even more. In order to explain the picture classifications for accountability, transparency, and debugging in the healthcare arena, Kinger and Kulkarni (2022) created a classification model for ocular illnesses that incorporates LIME. With a 92.81% accuracy rate, the study emphasized the value of XAI in healthcare and called for more research into XAI methods and other deep neural network algorithms for thorough comparison and optimization.

Dipu, Shohan, and Salam's (2021) multi-class classification implementation used an ImageNet pre-trained VGG-16 model and obtained 97.93% accuracy. Nevertheless, no explainability strategies were used. The study makes no recommendations for the future.

2.2 Discussion

Inspired by these studies, the current study intends to develop a dependable computer-aided diagnostic system that can be applied in the healthcare sector by utilizing multiple XAI techniques and incorporating multi-class classification in place of mult-label and similar deep-learning architectures. This will help them overcome their ten limitations in achieving accuracy. While addressing class imbalances with the study dataset, the research will employ appropriate but distinct deep-learning methods from those previously employed in an effort to improve multi-class categorization. To achieve a fair representation of classes, methods including data augmentation, under-sampling, and over-sampling will be used. To guarantee that the photos are consistent in size, resolution, and format, the study dataset will undergo preprocessing. The pre-trained ImageNet dataset models will be provided using Keras, an artificial neural network interface supported by Tensorflow, because transfer learning will be used. Researchers developed ImageNet, a natural picture dataset with over 15 million human-annotated photos in 1000 classes, to aid in the creation of computer vision algorithms (Brownlee, 2019). Although the photos in this dataset might not be identical to those in the research set, we found that using a variety of transfer learning models produced good accuracy. The goal will be to refine the model on the smaller dataset after replacing the final few layers of the pre-trained network with new ones that are appropriate for the task. we are also used feature extraction intstead of fine tuning.

Aspect	Feature Extraction	Fine-Tuning
Definition	Extracts pre-trained features from a model's intermediate layers and uses them as input to a new classifier or model.	Adapts a pre-trained model to a specific task by retraining some or all layers of the model.
Use Case	Suitable when computational resources are limited or the pre-trained model is sufficient for capturing relevant features.	Suitable when task-specific features need to be learned, especially if the target domain differs significantly.
Training Effort	Requires training only the final layers or classifiers, making it computationally efficient.	Requires training part or all of the model, which can be computationally intensive.
Data Requirement	Suitable for small datasets because it relies on pre-trained features.	Requires a larger dataset to avoid overfitting when retraining parts of the model.
Flexibility	Limited to the features learned by the pre-trained model; may not capture task-specific nuances.	Offers greater flexibility as the model can adapt its weights to the target domain.
Speed	Faster to implement and train since only the final classifier or a shallow model is trained.	Slower, as it involves retraining parts of the deep model.

2.2.1 CNN Model Selection

The goal of investigating the VGG-19 and ResNet50 CNN architectures is to add to the model set's variety. A strategic approach is suggested, even if VGG-19's increased complexity may cause possible overfitting problems. The class imbalance will be especially addressed through the use of transfer learning.Because ResNet50 performed well in earlier implementations but was criticized for lacking transparency, this study will try to clarify its decision-making procedures in order to increase confidence in its high accuracy. Vision Transformer will also be fulfilling. Since the combination of ResNet and Vision transformer has produced encouraging results in the past (Bhati et al., 2023), this work introduces Vision transformer to offer fresh perspectives on the dataset. Because of the findings of earlier research (Wang et al., 2020), which showed less than ideal outcomes, Efficient Net will be purposefully left out. This choice emphasizes how crucial it is to make well-informed architecture decisions, which is consistent with the main objective of creating efficient and successful deep learning models for the specified multiclass classification problem[8].

2.2.2 Selection of Explainable AI Methods

As noted by Kinger and Kulkarni (2022), for a thorough investigation of XAI approaches, this study will Use explainability techniques that are both model-specific and model-agnostic. Prior research has mostly centered on model-specific techniques that offer customized explanations, like Saliency Maps and Attention Maps to particular kinds or applications of machine learning models. Despite their insights, these techniques are fundamentally model-based.Bound and might be difficult to apply to different models. Model-agnostic approaches, on the other hand, provide explanations. They are not influenced by the structure of the neural network, instead concentrating on the inputs and outputs of the model to produce justifications. The study will include model-specific techniques such as GradCAM, which is specifically made for neural networks with convolutions (CNNs). GradCAM generates heat maps as visual explanations.Identifying key regions in a picture that affect the model's predictions. LIME (Local Interpretable Model-agnostic Explanations) and other model agnostic techniques We'll make use of SHAP (SHapley Additive Explanations). LIME makes it easier to comprehend individual forecasts by producing local estimates of the decision boundary of the model. In the meanwhile, SHAP values measure each feature's contribution to a model's prediction, providing a unified framework for feature evaluation significance[9].

Classification will be developed and evaluated using LIME, GradCAM, and SHAP, with comparisons drawn based on the following criteria:

1. **Consistency**: Evaluates the reproducibility of results under identical experimental conditions. For example, does running LIME multiple times yield similar explanations?

2. **Fidelity**: Measures the extent to which the explanation method accurately reflects the model's actual decision-making process. The quality of an XAI method's fidelity depends on both the truthfulness of the method to the model and the model's accuracy.

3. **Sensitivity**: Assesses the XAI method's responsiveness to changes in input data or its predictive class. This involves using augmented data to test the adaptability of the model,

as explanations should vary to reflect any modifications in the data.

4. **Clinical Relevance**: Determines the alignment of XAI method explanations with clinical considerations and expert knowledge, particularly critical for medical applications.

This research aims to encompass these criteria in the comparative analysis of explainability methods. Figure 1 below presents a visual comparison of the XAI techniques employed in this study[10].

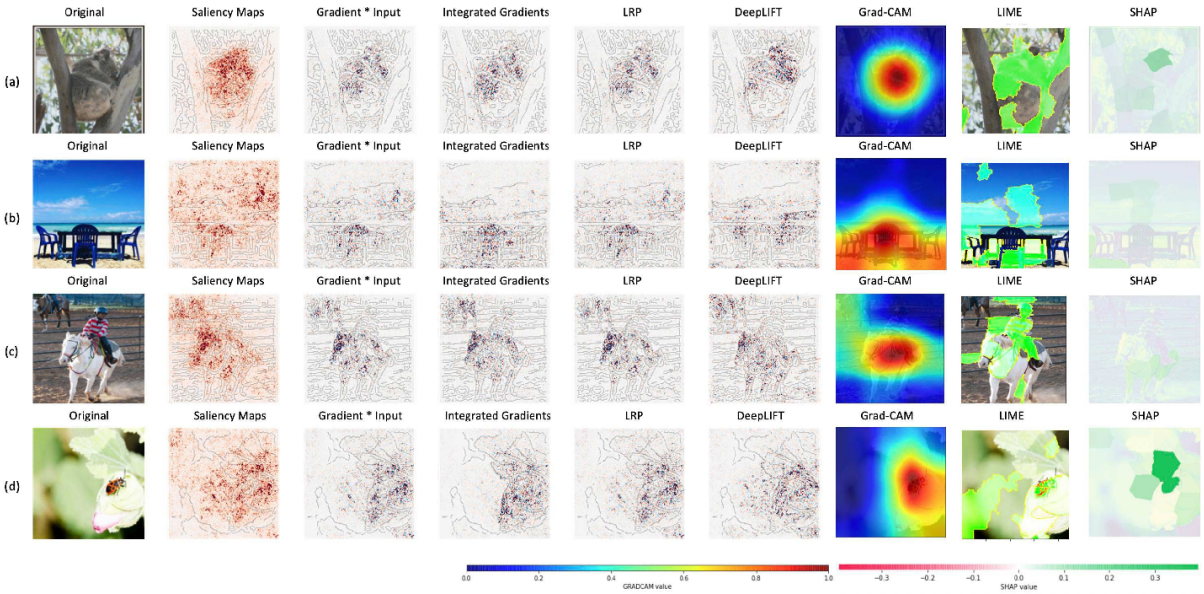


Figure 2.1: Figure 1 Comparative results of model-specific (e.g., Grad-CAM) and model-agnostic methods (Das and Rad, 2020)

CHAPTER 3

METHODS OF DETECTING OCULAR EYE DISEASES

3.1 Use of Deep Learning for Ocular Detection

To create a multi-class image classification model, transfer learning will be used on five CNN-based models that have already been trained. Each model's performance will be compared, and the model that performs the best will be further examined using three explainability techniques[11].

3.1.1 Convolution Neural Network for image classification

Because CNNs can efficiently handle picture data, they perform exceptionally well in image classification and object recognition (Brownlee, 2021). They apply filters to input images using convolutional layers, preserving important characteristics like edges, textures, and forms. These characteristics play a crucial role in the classification procedure.

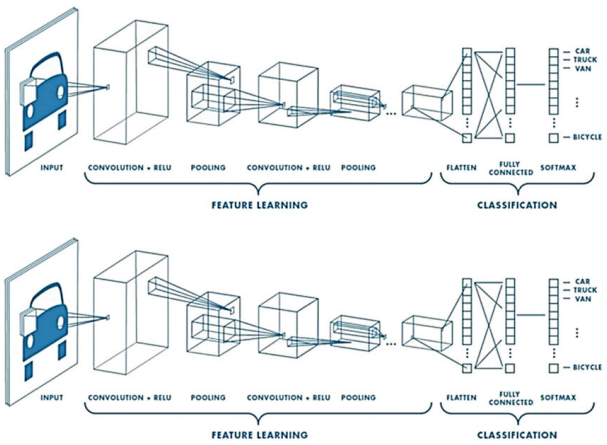


Figure 3.1: Architecture-of-a-network-of-convolutional-neural-networks-as-an-illustration.png

Figure 3.1 illustrates a diagram example of image classification using the CNN algorithm. The process starts with the initial image input, which then goes through a series of convolutional layers that utilise filters to identify features. Following each convolutional layer, it is common to include a Rectified Linear Unit (ReLU) activation function to introduce non-linearity, allowing the neural network to understand complex patterns. Following each convolutional layer, there is a pooling layer that decreases the representation's size, resulting in a reduction of parameters and computations in the network, ultimately aiding in preventing overfitting. This procedure also

guarantees that the recognition of features stays reliable even if changes occur in dimensions and orientation, enhancing the overall accuracy of the image categorisation system. The process of feature learning concludes with a flattening step, which transforms the 3D result from the last pooling layer into a 1D feature vector. The vector is inputted into a sequence of fully connected layers where the "learning" occurs through adjustments of weights. In the end, a SoftMax layer categorises the image into different classes like car, truck, van, and bicycle by giving probabilities to each class, and the class with the highest probability is the model's output. This structure allows the CNN to analyse unprocessed image data, understand characteristics, and forecast the image's content in a step-by step approach.

3.1.2 Transfer Learning

Transfer Learning is applied using the following steps below:

1. Obtain layers from[1] a previously trained model on the ImageNet dataset.
2. Freeze them, to avoid destroying any existing knowledge they contain during future training rounds.
3. Add some new trainable layers on top of the frozen layers. These new layers will learn to turn the old features into predictions on the new dataset.
4. Then train the new layers on the study dataset.
5. Apply fine-tuning, which consists of unfreezing a few layers of the entire model obtained and retraining it on the study dataset at a low learning rate to achieve more improvements[12].

3.2 CNN Models

The image classification system is implemented in this study using the TF-Keras Application models for the specified architectures.

3.2.1 VGG19

With three more layers ,VGG19 is an extension of VGG16 that might be able to understand more complex features and marginally improve performance in some tasks. However, the task and dataset may have a significant impact on the performance gains, which may be modest. Furthermore, more processing power is needed to achieve the additional depth,[2] .

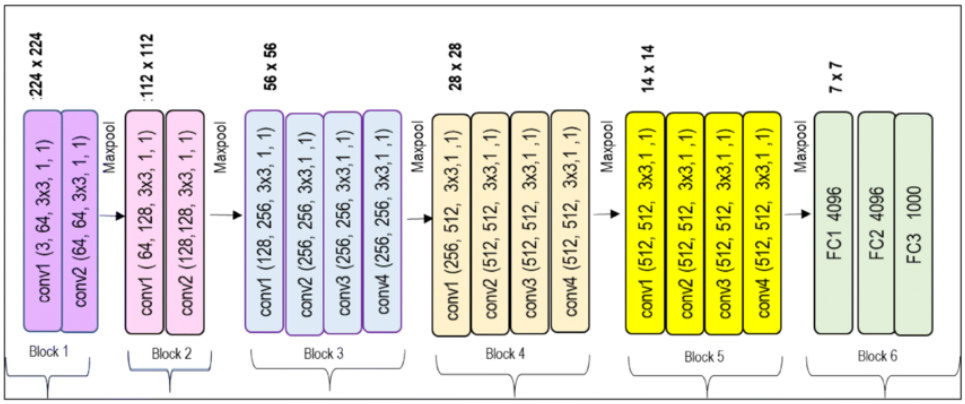


Figure 3.2: VGG19 Model Architecture

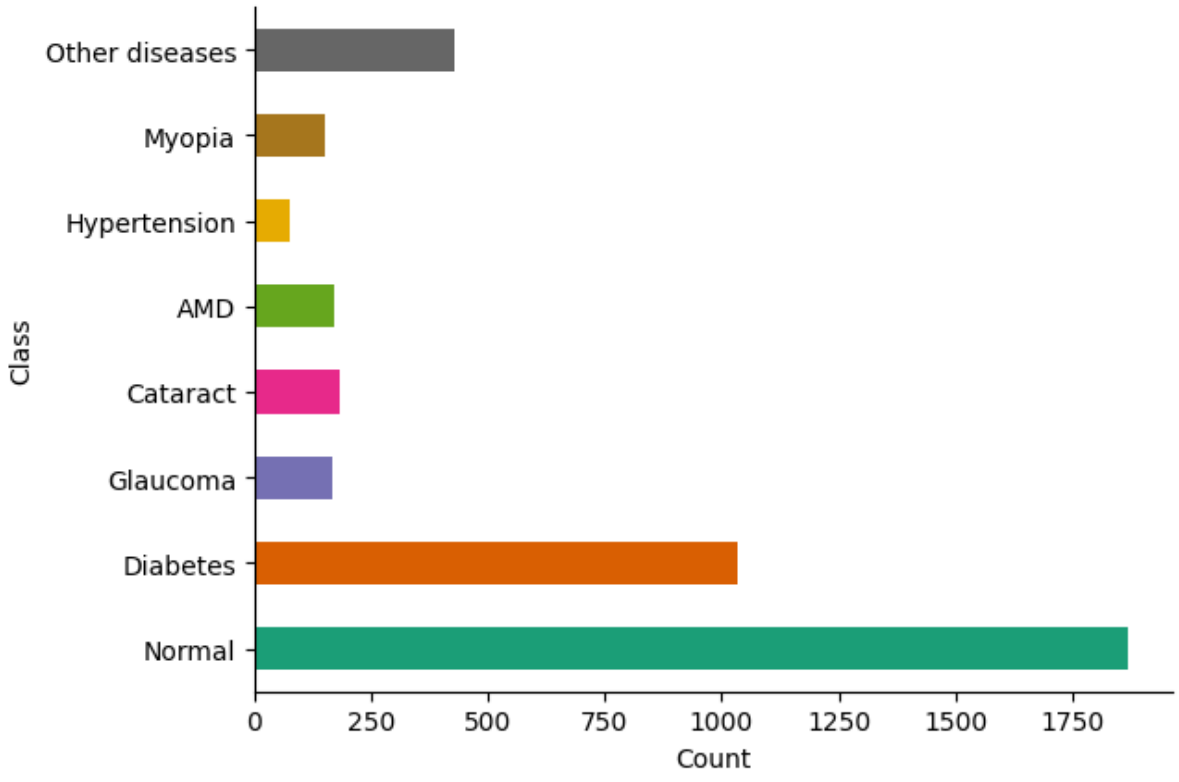


Figure 3.3: Dataset classification image

.

3.2.2 ResNet50

ResNet-50 is a pre-trained model that won the 2015 ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) competition. It was trained on a portion of the ImageNet database. The model can classify photos into 1000 item categories and is trained on more than a million photographs. It contains 177 layers in total, which corresponds to a 50-layer residual network (224, 224, 3) [18]. The ResNet architecture (figure 3.4) is considered to be among the most popular Convolutional Neural Network architectures around. Residual Networks (abbreviated ResNet) were first described by Xiangyu Zhang, Kaiming He, Jian Sun, and Shaoqing Ren in their 2015 computer vision research paper titled "Deep Residual Learning for Image Recognition" [18]. ResNet was later introduced by Microsoft Research in 2015 and set numerous records[13] .

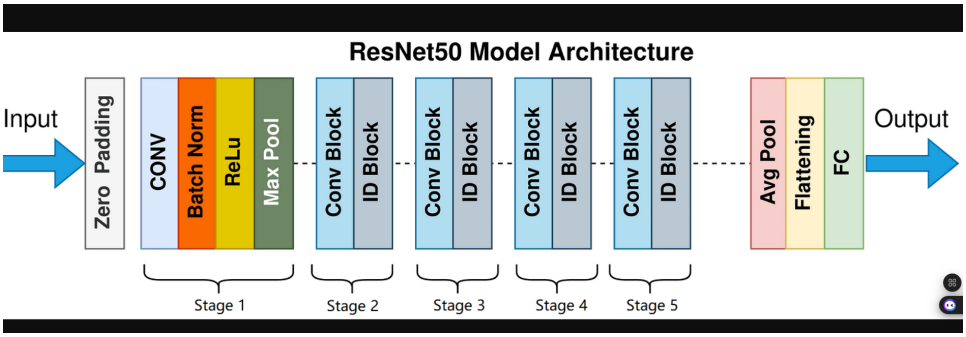


Figure 3.4: ResNet50 Architecture

Mathematical Explanation of How Residual Blocks Work in ResNet50

Assume you have an input x. In a traditional neural network layer, you might pass x through a set of operations (like convolution, batch normalisation, and activation) to get an output, which we can call $H(x)$.

In a residual block, rather than learning $H(x)$ directly, the layers learn a residual function $F(x)$ with the idea that learning this residual is easier:

(1) $$F(x) = H(x) - x$$

The output of the residual block then becomes:

(2) $$H'(x) = F(x) + x$$

3.2.3 Vision Transformer

Vision Transformers (ViT) has recently emerged as a competitive alternative to Convolutional Neural Networks (CNNs) that are currently state-of-the-art (SOTA) in different image recognition computer vision tasks. ViT models outperform the current SOTA CNNs by almost x4 in terms of computational efficiency and accuracy.

Transformer models have become the de facto status quo in Natural Language Processing (NLP). For example, the popular ChatGPT AI chatbot is a transformer-based language model. Specifically, it is based on the GPT (Generative Pre-trained Transformer) architecture. This uses self-attention mechanisms to model the dependencies between words in a text.

We can find several proposals for vision transformer models in the literature. The overall structure of the vision transformer architecture consists of the following steps:

1. Split an image into patches (fixed sizes)
2. Flatten the image patches
3. Create lower-dimensional linear embeddings from these flattened image patches
4. Include positional embeddings
5. Feed the sequence as an input to a SOTA transformer encoder
6. Pre-train the ViT model with image labels, then fully supervised on a big dataset
7. Fine-tune the downstream dataset for image classification

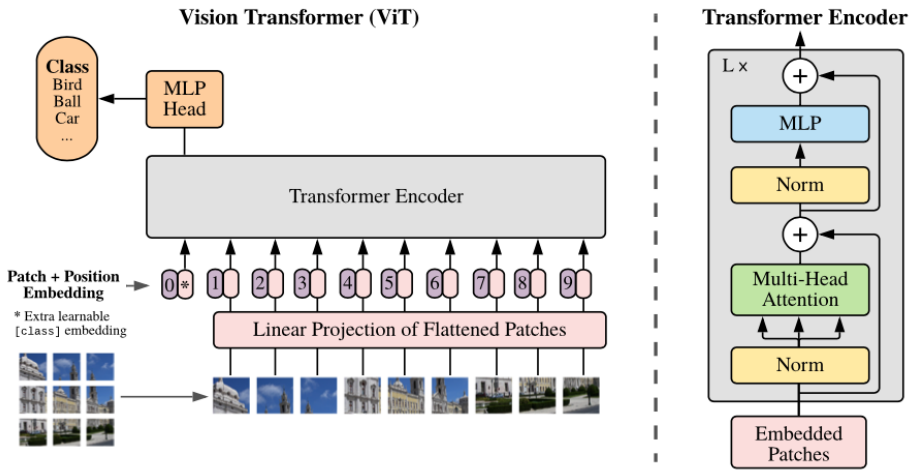


Figure 3.5: Vision Transformer Architecture

CHAPTER 4

DATA COLLECTION AND ANALYSIS

4.1 Dataset

The Ocular Disease Intelligent Recognition (ODIR) dataset, available on Kaggle's "Ocular Disease Recognition" (2020) challenge, is a comprehensive collection of patient data compiled by Shanggong Medical Technology Co., Ltd. in collaboration with various hospitals and medical facilities across China. This dataset contains color fundus photographs captured using different camera models, leading to variations in image quality and resolution[14].

The dataset is organized into multiple components, including a detailed data frame file, `full_df.csv`, available in Excel format. Additionally, pre-processed versions of the training images are provided. The dataset is carefully structured, with separate folders for training and testing, which are pre-split to facilitate model evaluation. Figure ?? illustrates the dataset structure, presenting the first five rows of the `full_df.csv` file.

	ID	Patient Age	Patient Sex	Left-Fundus	Right-Fundus	Left-Diagnostic Keywords	Right-Diagnostic Keywords	N	D	G	C	A	H	M	O	filepath	labels	target	filename
0	0	69	Female	0_left.jpg	0_right.jpg	cataract	normal fundus	0	0	0	1	0	0	0	0	../input/ocular-disease-recognition-odir5k/ODI...	['N']	[1, 0, 0, 0, 0, 0, 0, 0]	0_right.jpg
1	1	57	Male	1_left.jpg	1_right.jpg	normal fundus	normal fundus	1	0	0	0	0	0	0	0	../input/ocular-disease-recognition-odir5k/ODI...	['N']	[1, 0, 0, 0, 0, 0, 0, 0]	1_right.jpg
2	2	42	Male	2_left.jpg	2_right.jpg	laser spot，moderate non proliferative retinopathy	moderate non proliferative retinopathy	0	1	0	0	0	0	0	1	../input/ocular-disease-recognition-odir5k/ODI...	['D']	[0, 1, 0, 0, 0, 0, 0, 0]	2_right.jpg
3	4	53	Male	4_left.jpg	4_right.jpg	macular epiretinal membrane	mild nonproliferative retinopathy	0	1	0	0	0	0	0	1	../input/ocular-disease-recognition-odir5k/ODI...	['D']	[0, 1, 0, 0, 0, 0, 0, 0]	4_right.jpg
4	5	50	Female	5_left.jpg	5_right.jpg	moderate non proliferative retinopathy	moderate non proliferative retinopathy	0	1	0	0	0	0	0	0	../input/ocular-disease-recognition-odir5k/ODI...	['D']	[0, 1, 0, 0, 0, 0, 0, 0]	5_right.jpg

Figure 4.1: Showing the first five rows of the full df csv file.

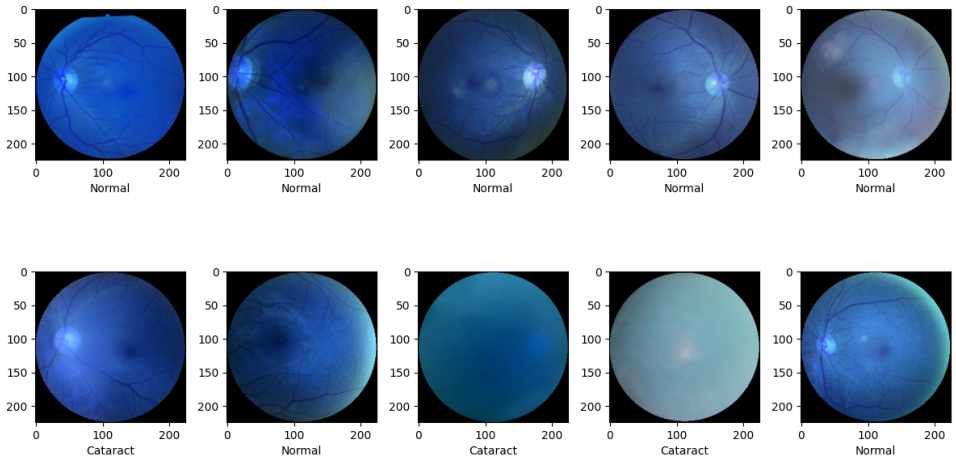


Figure 4.2: Cataract And Normal Image

Each patient is assigned classification labels for both eyes, along with informative diagnostic keywords. All eight classes are represented in the `target` column, which is one-hot encoded and structured for multi-class classification. A binary value of 1 or 0 indicates the presence or absence of a particular ocular disease[15].

A sample of fundus Cataeact photographs from the training set is shown in Figure 4.2. These images demonstrate the dataset's value for this investigation and the variability in image quality, which underscores the need for robust preprocessing techniques[16].

4.2 Challenges in the Dataset

The Ocular Disease Intelligent Recognition (ODIR-5K) dataset presents several challenges that are crucial for the development of deep learning models for ocular disease classification. These challenges include:

1. **Class Imbalance:** The dataset exhibits significant class imbalance, with certain diseases such as glaucoma and myopia having fewer instances compared to others like normal and diabetes. This imbalance can lead to biased model performance, where the model may favor the majority class, resulting in reduced sensitivity for the minority classes.

2. **Variability in Image Quality:** Fundus images in the dataset are captured using various devices from different manufacturers, leading to variations in image quality, resolution, and color representation. This variability can complicate the feature extraction process and affect the model's ability to generalize across different image qualities.

3. **Multilabel Classification:** Each patient in the dataset can have multiple diagnoses, making the classification task multilabel in nature. This requires models to predict multiple labels simultaneously, which can be more complex than single-label classification tasks.

4. **Data Privacy and Ethical Considerations:** The dataset contains sensitive patient information, including age and medical conditions. Ensuring data privacy and adhering to ethical standards are crucial when handling and processing such data[17].

Addressing these challenges is essential for developing robust and accurate models capable of effectively diagnosing ocular diseases from fundus images.

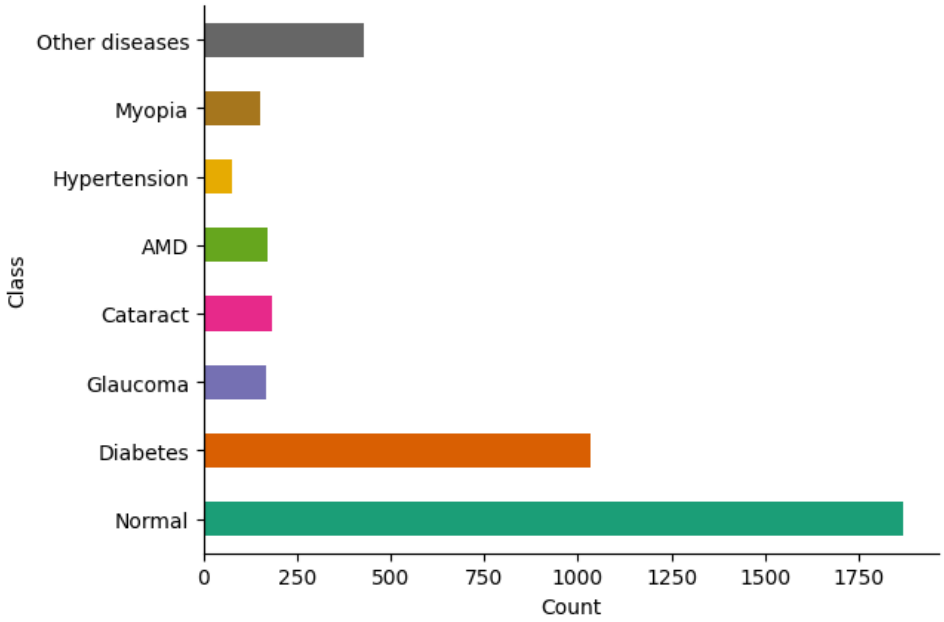


Figure 4.3: Example of a fundus image from the dataset.

CHAPTER 5

HUMAN-CENTRIC AI EXPLANATIONS

5.1 AI Decision Transparency

Human-centric AI explanations approach are described in below .

5.1.1 Local interpretable model-agnostic explanations

Ribeiro, Singh, and Guestrin (2016) introduced LIME, a technique designed to explain individual predictions made by classifier models[9]. For image data, LIME utilizes a superpixel approach, where the algorithm segments an image into superpixels,Äîgroups of pixels sharing similar properties. These superpixels serve as features, allowing LIME to analyze the model's behavior. By modifying superpixels and observing how the model's predictions change, LIME identifies the most significant features of an image. This method simplifies the explanation of complex concepts, especially in scenarios where visual information plays a key role[18].

5.1.2 Shapley additive explanations

In 2017, Lundberg and Lee introduced SHAP (Shapley Additive Explanations), a method based on Shapley values from game theory to interpret individual predictions of machine learning models[19]. In this framework, each feature in the dataset is treated as a "player" in a cooperative game, with the model's prediction serving as the "reward." SHAP calculates the contribution of each feature (its Shapley value) by comparing the model's predictions with and without the feature across all possible subsets of features. This process identifies the marginal contribution of each feature, providing a detailed assessment of its influence on the prediction. By summing these contributions, SHAP delivers comprehensive and intuitive explanations, making it especially valuable for complex models where straightforward interpretation is challenging[3].

5.1.3 Gradient-weighted class activation mapping

Grad-CAM (Gradient-weighted Class Activation Mapping) is a technique developed to explain the decisions of convolutional neural networks (CNNs), particularly for visual tasks such as image classification[20]. It works by computing the gradients of a target output, such as a specific class label, with respect to the feature maps of a selected convolutional layer. These gradients are then aggregated to create weights that indicate the importance of each feature map for predicting the target class. By combining these weighted feature maps, Grad-CAM produces a

"heatmap" that highlights the areas in the input image most relevant to the model's prediction. This heatmap, when overlaid on the original image, provides a visual representation of the regions that influenced the classification decision. Grad-CAM is especially useful for understanding how a model perceives and prioritizes different parts of an image during its decision-making process.

CHAPTER 6

METHODOLOGY

6.1 Experiment Process

Using the ODIR-5K dataset, The main focus is on creating a balanced dataset of cataract and normal images, combining left and right eye images, and preparing them for deep learning model training.[21]. But For the work at hand, these resources were not adequate. As a result, We switched to Google Collab, which offered the processing capacity required for more complex computer vision-based activities[9].

Cataract Detection Experiment

The experiment involves the detection of cataracts using images from a dataset. Here's a breakdown of the process:

Cataract and Normal Dataset Creation

The code first separates the dataset into two categories: images of cataracts and normal images (without cataracts). The data is divided based on diagnostic keywords present in the dataset (for example, "normal fundus" for normal images). The dataset is further filtered for images that are either related to the left eye or the right eye, depending on the diagnostic keywords in the *Left-Diagnostic Keywords* and *Right-Diagnostic Keywords* columns. After the initial separation, 250 random samples of normal images are selected from both left and right eyes for balance[22].

Combining the Datasets

The cataract dataset for both left and right eyes is combined into a single array, and similarly, the normal dataset for both eyes is combined. This forms the basis of the two primary categories: cataract and normal. The cataract and normal datasets are then ready for model training, where the images will be fed into a deep learning model for classification[23].

Dataset Size Information

The code prints the number of images in both the cataract and normal categories to confirm the dataset's size before training the model[24].

Image Dataset Creation

The `create_dataset` function is responsible for creating the final dataset, where each image is read, resized, and paired with its corresponding label. The label for cataract images is `1`, and for normal images, it is `0`. Images are loaded from a specified directory (*dataset_dir*), resized to the target size of 224x224 pixels, and added to the dataset with their respective labels. The dataset is then shuffled to ensure randomness before feeding it into the model for training. We made sure to use reliable and well-established models as the basis for our classifier by sourcing all of the pre-trained models used in this study from the Keras 2 API documentation.

6.1.1 Data Preprocessing

The `train_test_split` function from the `sklearn.model_selection` library was used to separate the dataset into training and test sets, allocating 20% of the data to the test set and the remaining 80% to training. This division preserved a strong evaluation set while guaranteeing enough data for training[15]. This involves loading, resizing, labeling, and shuffling images, followed by combining the data into a usable format for training.
Data processing involves several important steps to prepare the images for use in training deep learning models. Below are the steps followed in this process:

Loading and Resizing Images

Images are loaded using the `cv2.imread()` function, which reads the image in color (using the `cv2.IMREAD_COLOR` flag). After reading the image, it is resized to a standard size of 224x224 pixels using the `cv2.resize()` function. This ensures that all images have a uniform size, which is crucial for feeding them into deep learning models that expect fixed input sizes.

Labeling Images

Each image is associated with a label, where:

- **1** represents a cataract image.
- **0** represents a normal image.

The labels are assigned based on the diagnostic keywords in the dataset (for example, "normal fundus" for normal images and "cataract" for cataract images).

Shuffling the Dataset

After the images are processed and labeled, the dataset is shuffled using the `random.shuffle()` function. Shuffling ensures that the model is not biased by the order of the images, promoting more robust learning[25].

Creating a Dataset for Model Training

The `create_dataset` function compiles the processed images and labels into a dataset that is ready for model training. This dataset is returned as a list of pairs, where each pair consists of an image (as a numpy array) and its corresponding label (as a numpy array). This dataset will then be used for training a deep learning model to classify images as either having cataracts or being normal[5].

Data Augmentation

TensorFlow's `Sequential` API was used to construct a pipeline for data augmentation but only for Vission Transformer. In order to provide diversity to the training data, this pipeline consisted of layers that performed zoom operations, rotations, and random flips both vertically and horizontally. Only the training set was subjected to these changes, guaranteeing the integrity of the test set for objective assessment. Before application, the augmentation pipeline was modified to fit the training data, maximizing its impact on the data set[14].But in our code for first two models does not include data augmentation, it is common to apply techniques such as rotations, flipping, or brightness adjustments to artificially expand the dataset and make the model more robust to different variations of the input images. This can be implemented using Keras' `ImageDataGenerator` or custom transformations[26].

Handling Class Imbalance

To address the issue of class imbalance in the dataset, the following techniques were employed:

- **Oversampling the Minority Class:** This technique involves increasing the number of samples in the minority class by duplicating existing samples. It helps to balance the dataset by ensuring that the model has enough examples from the minority class to learn from[27].
- **Undersampling the Majority Class:** In this approach, the number of samples from the majority class is reduced by randomly removing some of its instances. This ensures that both classes are represented equally, preventing the model from being biased toward the majority class.
- **Using Class Weights in Model Training:** Another effective method for dealing with class imbalance is to assign different weights to each class during the training process. Many deep learning frameworks, including Keras, provide the ability to assign higher weights to the minority class, allowing the model to pay more attention to these underrepresented instances during training. This technique avoids altering the dataset size while improving the model's sensitivity to the minority class[28].

Conclusion

Although consistent and in line with the demands of transfer learning architectures, this method offered a solid basis for training deep learning models.

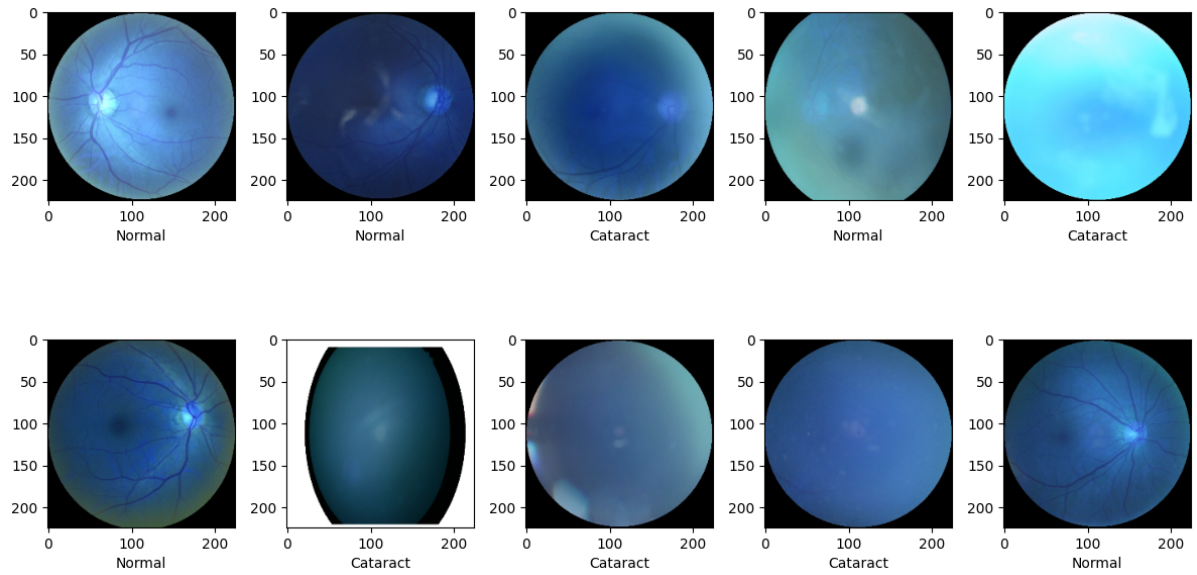


Figure 6.1: Train imgae

```
dataset = create_dataset(cataract,1)
len(dataset)
```

Python

```
100%|██████████| 594/594 [01:23<00:00,  7.09it/s]

519
```

```
dataset = create_dataset(normal,0)
len(dataset)
```

Python

```
100%|██████████| 500/500 [01:02<00:00,  8.00it/s]

751
```

```
plt.figure(figsize=(12,7))
```

Figure 6.2: How to train

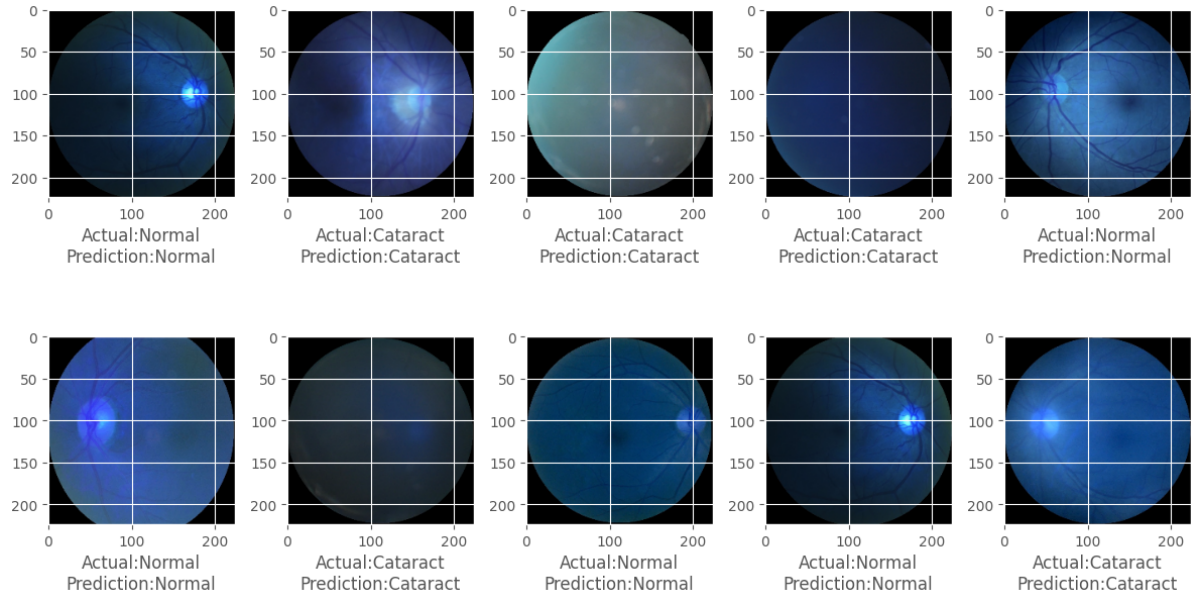


Figure 6.3: Testing image

6.1.2 Hyperparameter configuration

Since they control the learning process and have a big impact on the models' performance, hyperparameters are crucial to machine learning models. The hyperparameter configurations I used for the base model and the model with random sampling, two iterations of the image classification models I created, are shown in Tables 2 and 3. I changed each parameter separately during the hyperparameter tuning process to see how it affected the model's accuracy and loss charts. This method made sure the model was overfitting-free and generalizing well. Furthermore, because of Google Colab's computing resource constraints, I gave priority to characteristics that allowed for quicker training timeframes. For each of the three models, the following hyperparameters were maintained[29].

Configuration	**Value**
Optimisation Function	Adam
Epoch	10 (Complete training), 80 (Finetuning)
BatchSize	32
Learning Rate	0.0001
Dropout	0.2
ReduceLROnPlateau	monitor='val_loss', factor='0.5', patience='3', min_lr='1.00E-07'
EarlyStopping	monitor='val_loss', patience='5'
ModelCheckpoint	save_best_only=True

Table 6.1: Hyperparameters for resampled training data model

Configuration	**Value**
Optimisation Function	Adam
Epoch	80 (Complete training), 80 (Finetuning)
BatchSize	32
Learning Rate	0.0001
Dropout	0.2
L2 Regularisation	kernel_regularizer=l2(0.001)
ReduceLROnPlateau	monitor='val_loss', factor='0.5', patience='3', min_lr='1.00E-07'
EarlyStopping	monitor='val_loss', patience='5'
ModelCheckpoint	save_best_only=True

Table 6.2: Hyperparameters for resampled training data model (with L2 regularization)

Categorical crossentropy was chosen as the best loss function because this project deals with a multi-class classification problem. This decision is based on how well it handles many classes, with each instance being solely assigned to one class out of many. The difference between the actual distribution of the classes and the anticipated probabilities is measured by categorical crossentropy, commonly referred to as SoftMax loss. In order to minimize the loss by modifying the model weights to forecast a probability distribution that is as near to the actual distribution as feasible, it calculates the likelihood that a given input belongs to each of the classes[30].

6.1.3 Model Architectures

I used pre-trained Keras models that were initially trained on the ImageNet dataset to create the picture categorization model. With the top layer removed and average pooling applied to account for the specific classification problem, these models formed the basis of the architecture. All pre-trained layers were frozen to non-trainable to ensure the preservation of the learned features from ImageNet. This step was essential to maintain the integrity of the previously learned features, which are often relevant to a variety of visual recognition tasks.Later, new layers were added to the model to tailor it to the image categorization problem of the study dataset[18]. These modifications included:

- **Dense Layer**: A dense layer with 512 units and ReLU (Rectified Linear Unit) activation was incorporated to introduce non-linearity into the model, enabling it to recognize increasingly intricate patterns.
- **Dropout Layer**: A regularization dropout layer was included to prevent overfitting. Dropout enhances the model's resilience by randomly setting a fraction of the input units to 0 during each update in training, reducing the likelihood of memorizing the training data.
- **SoftMax Layer**: A final dense layer was added to generate predictions for each of the eight classes using a SoftMax activation function. The SoftMax layer is critical for multi-class classification tasks as it converts the model's raw predictions into probabilities by normalizing the exponential of each output[29].

Figures below illustrate the basic structure used for all base model implementations, along with the architectural configuration for each model.To enable fine-tuning, the final four layers of the model were unfrozen during the later training phases. This fine-tuning process was crucial for better tailoring the model to the specific needs of the new task. By making these layers trainable, they could adjust their weights in response to fresh information, capturing task-specific subtleties.The model was recompiled with a significantly lower learning rate during this phase to minimize disruption to the previously learned features. This conservative approach ensured that adjustments were subtle, enhancing the model's performance on the new task by refining rather than overwriting the previously learned details[26].

resnet50_input	input:	[(None, 224, 224, 3)]
InputLayer	output:	(None, 224, 224, 3)
resnet50	input:	(None, 224, 224, 3)
Functional	output:	(None, 2048)
dense	input:	(None, 2048)
Dense	output:	(None, 512)
dropout	input:	(None, 512)
Dropout	output:	(None, 512)
dense_1	input:	(None, 512)
Dense	output:	(None, 8)

Table 6.3: ResNet50 Model Architecture

vgg19_input	input:	[(None, 224, 224, 3)]
InputLayer	output:	(None, 224, 224, 3)
vgg19	input:	(None, 224, 224, 3)
Functional	output:	(None, 512)
dense	input:	(None, 512)
Dense	output:	(None, 512)
dropout	input:	(None, 512)
Dropout	output:	(None, 512)
dense_1	input:	(None, 512)
Dense	output:	(None, 8)

Table 6.4: VGG19 Model Architecture

vit_input	input:	[(None, 224, 224, 3)]
InputLayer	output:	(None, 224, 224, 3)
vit	input:	(None, 224, 224, 3)
Embedding	output:	(None, 224, 224, 768)
TransformerBlock	input:	(None, 224, 224, 768)
TransformerBlock	output:	(None, 224, 224, 768)
Dense	input:	(None, 768)
Dense	output:	(None, 512)
dropout	input:	(None, 512)
Dropout	output:	(None, 512)
dense_1	input:	(None, 512)
Dense	output:	(None, 8)

Table 6.5: Vision Transformer Model Architecture

6.1.4 Training of the models

This section outlines the training process of the models under study. The training was conducted in two distinct phases: the base implementation phase and the sampling implementation phase. Both phases utilised specific hyperparameter configurations as detailed in Table 2 and Table 3, respectively[31].

VGG19

The VGG19 shows a convergence of training and validation accuracies similar to the VGG16 for the basic models, indicating a similar capacity for learning from the training data. Interestingly, compared to the VGG16 model, the model loss for VGG19, as shown in Figure 6.6, shows less variability and a greater decline in validation loss. The VGG19 architecture's extra layers may have contributed to this steeper and more gradual drop in loss by giving it a better capacity to identify patterns in the training data more quickly. The VGG19 (Figure 6.6) no needs additional

epochs to show comparable patterns in the accuracy and loss graphs to the VGG16 when comparing the sampling models[24].

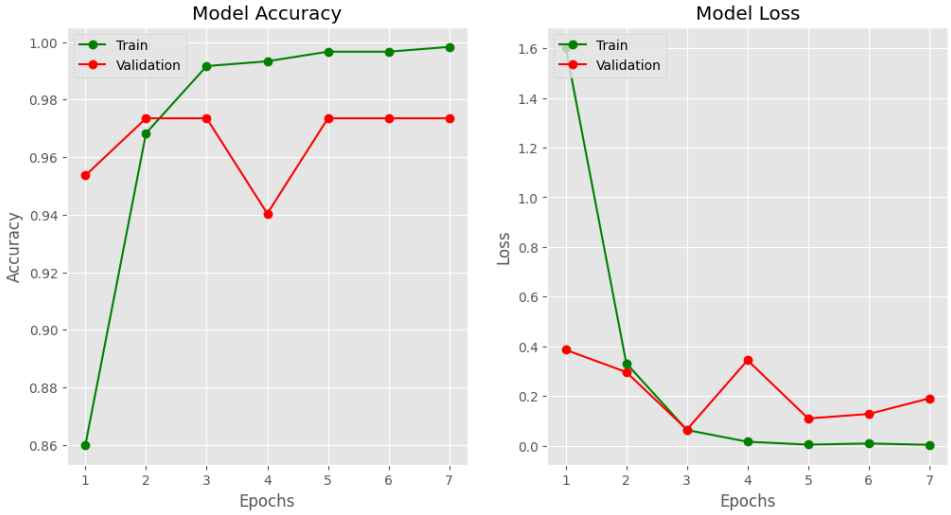


Figure 6.4: Vgg19 Model Accuracy and Model Loss

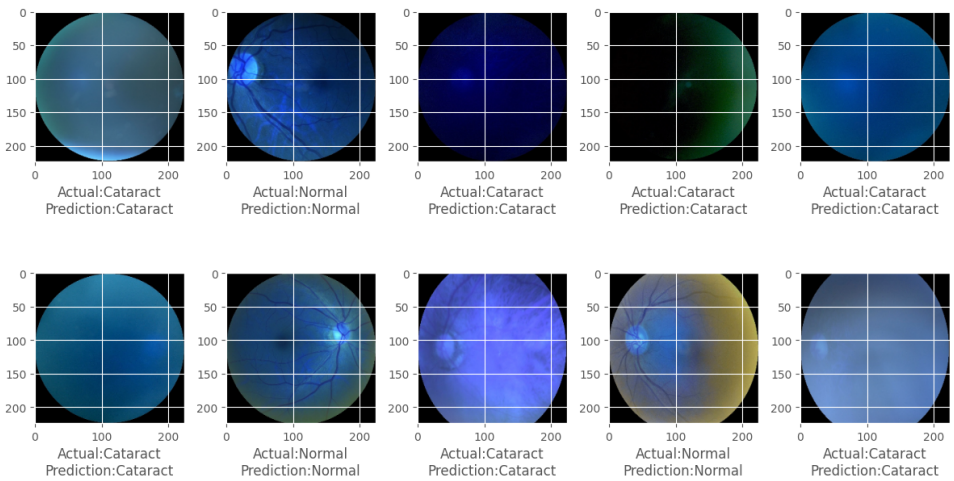


Figure 6.5: Vgg19 model Prediction Table

According to this finding, the VGG19 may gain from its deeper architecture in terms of learning efficiency, but when class balancing strategies are used, it also requires a longer training time to fully realize this benefit[32].

ResNet50

While the validation accuracy for the basic model plateaus early and stays relatively low, the accuracy graph of ResNet50 (Figure 6.7) shows a steady gain in training accuracy, indicating a higher level of overfitting than that shown in the VGG models. This is also supported by the loss graph, which shows significant overfitting with a more noticeable and quick decline in training loss and a noticeably slower decline in validation loss. One possible explanation for this difference in learning dynamics between the VGG and ResNet50 models is the employment of a different preprocessing function designed for the pre-trained ResNet50 version. Because these preprocessing changes have a direct impact on the representation of input data that the model learns from, they can have a substantial impact on the model's capacity to generalize. The use of oversampling approaches, as seen in Figure 23, does not significantly improve the alignment between training and validation performance indicators; instead, the ResNet50 model still shows severe overfitting[22].

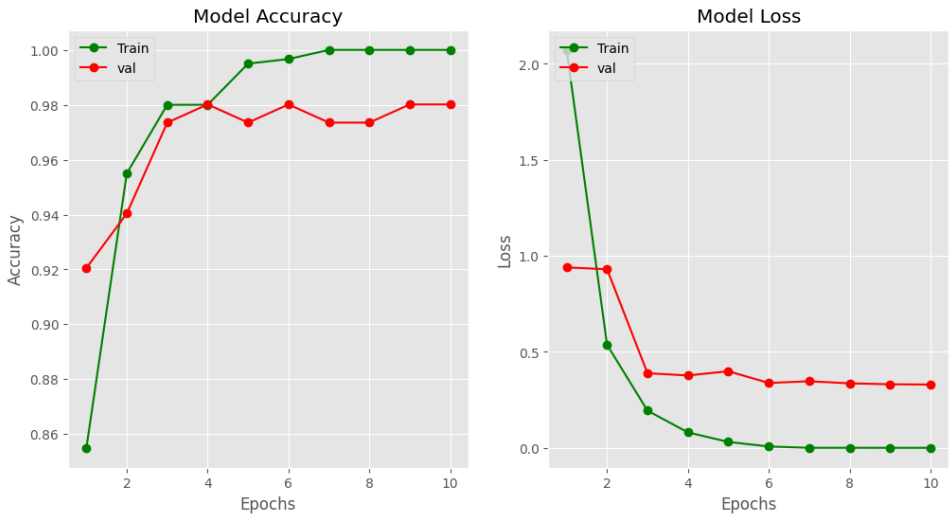


Figure 6.6: ResNet50 Model Accuracy and Model Loss

This benefit is offset by the larger difference between the training and validation accuracy and loss, which suggests worse generalization to unknown data, even though the model was trained for fewer epochs than the VGG19.

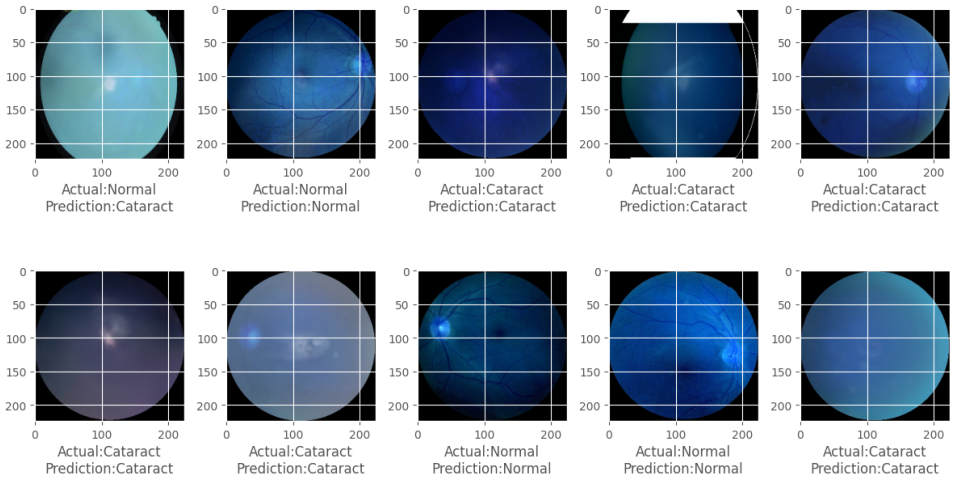


Figure 6.7: ResNet50 Prediction table

Vision Transformer

The Vision Transformer (ViT) base model demonstrated remarkable training performance, achieving an impressive accuracy of 98.12%, with a sharp decrease in training loss to 0.021. However, its validation accuracy was notably lower at 85.46%, accompanied by a relatively high validation loss of 0.349, indicating substantial overfitting. This performance disparity underscores the model's struggle to generalize effectively to unseen data despite its strong learning capacity on the training set. The persistent gap between training and validation performance metrics persisted even after applying oversampling and augmentation techniques designed to address class imbalance and enhance data diversity[24].
These results highlight the inherent challenges associated with training Vision Transformers, particularly in scenarios involving limited or imbalanced datasets. While ViT excels at capturing intricate patterns in the training data due to its self-attention mechanisms and ability to model global dependencies, the model's reliance on large-scale, high-quality data becomes apparent. The application of advanced preprocessing strategies, such as data augmentation pipelines tailored to the model's requirements, or integrating additional regularization techniques like dropout, weight decay, or stochastic depth, could potentially mitigate overfitting. Moreover, fine-tuning the pre-trained ViT on domain-specific data with a reduced learning rate and unfreezing selective layers might further improve its generalization capabilities[22].

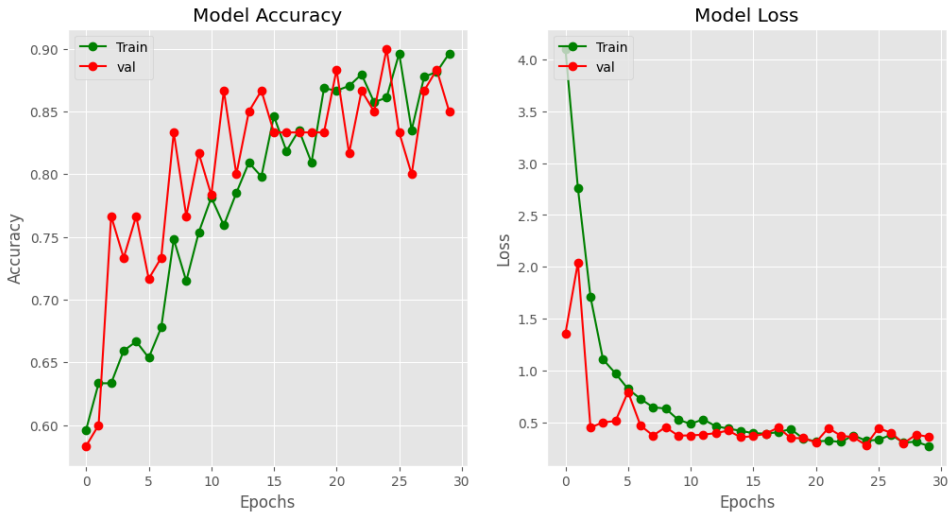


Figure 6.8: Vission Transformer Model Accuracy and Model Loss

This analysis underscores the importance of adopting a holistic approach when deploying Vision Transformers for complex classification tasks, balancing their robust learning capabilities with techniques to reduce overfitting and ensure consistent performance across both training and validation datasets.

6.2 Evaluation Metrics

To analyze the performance of the classification models, evaluation metrics such as accuracy, precision, recall, F1 score, balanced accuracy score, and ROC AUC (Receiver Operating Characteristic Area Under the Curve) were employed. The formulas for these metrics are provided below:

- **Accuracy:**
$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$
- **Precision:**
$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$
- **Recall:**
$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$
- **F1 Score:**
$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$
- **Balanced Accuracy Score:**
$$\text{Balanced Accuracy Score} = \frac{1}{N} \sum_{i=1}^{N} \frac{\text{TP}}{\text{TP} + \text{FN}}$$

where:

- **TP (True Positive)**: Number of correctly identified positive cases.
- **FP (False Positive)**: Number of cases incorrectly predicted as positive.
- **TN (True Negative)**: Number of correctly identified negative cases.
- **FN (False Negative)**: Number of cases incorrectly predicted as negative.
- **N**: Total number of classes.

The *Balanced Accuracy Score* evaluates the performance of each class individually by considering true positives and false negatives, then averages the accuracies across all classes. This approach ensures that all classes, regardless of size, contribute equally to the overall metric, making it particularly suitable for imbalanced datasets[26].
The *ROC AUC* is used to evaluate the model's ability to distinguish between classes in multi-class classification problems. It provides a comprehensive measure of model performance beyond accuracy, particularly in datasets with uneven class distributions. A higher AUC value indicates improved model performance in correctly predicting the positive and negative classes.

6.3 Results Analysis

Model	Accuracy on Test Dataset	Precision (macro average)	Recall (macro average)	F1-Score (macro average)	Balanced Accuracy Score	AUC ROC Score
VGG19	97.35%	0.96	0.98	0.97	96.31%	96.46%
ResNet50	96.68%	0.95	0.98	0.97	96.41%	96.44%
Vision Transformer	86.67%	0.83	0.82	0.81	82.39%	84.73%

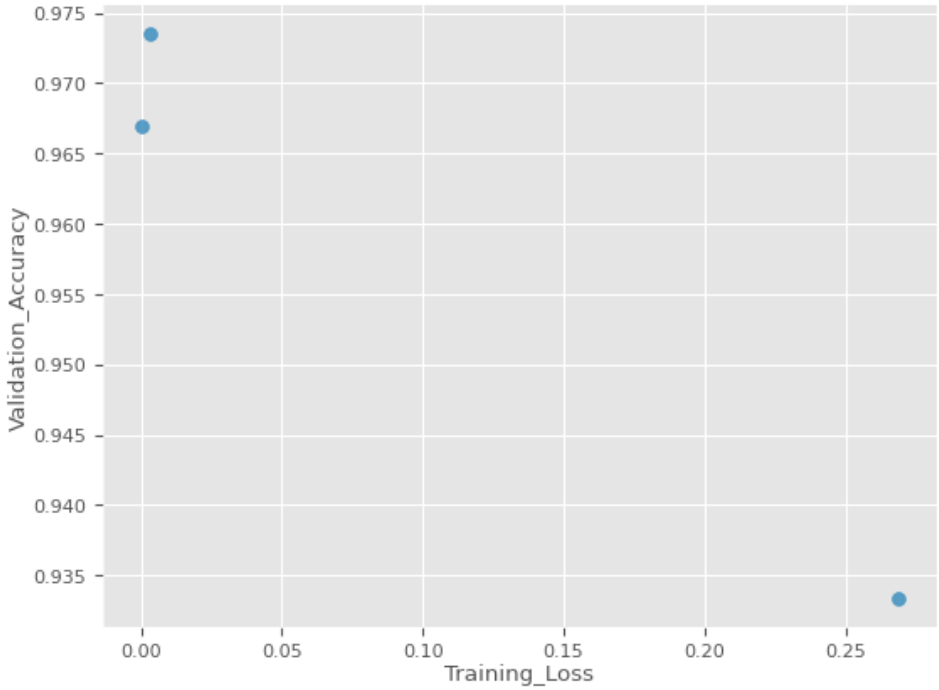


Figure 6.9: Validation Accuracy and Training Loss for 3 models

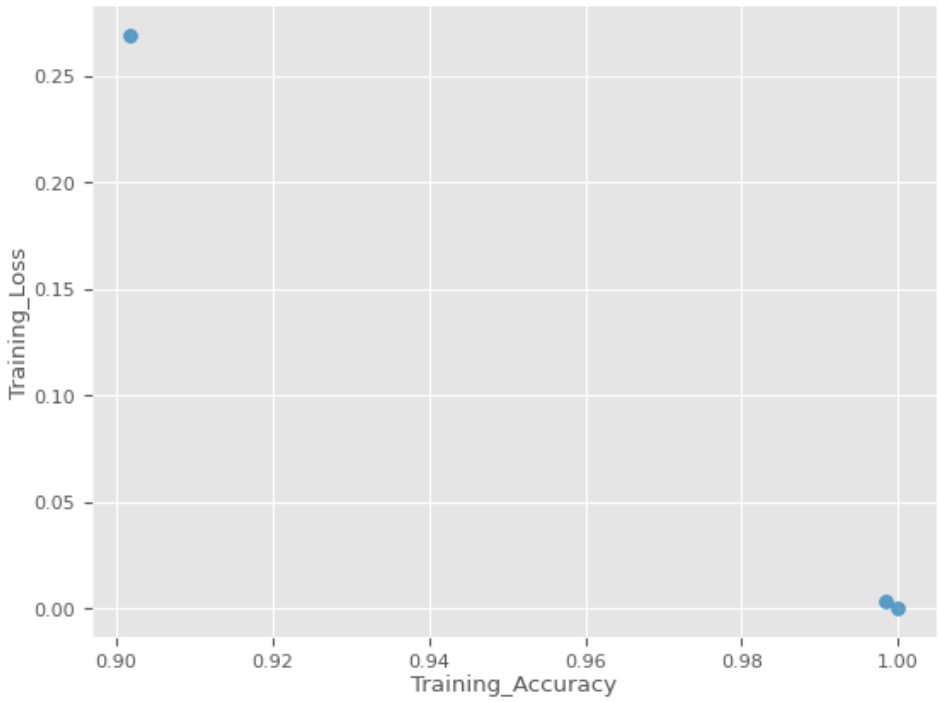


Figure 6.10: Tranning Loss And Trainning Accuracy for 3 models

The best-performing model is VGG19, as it demonstrates superior performance across all evaluated metrics, it achieves the highest Model Accuracy (97.65%), Balanced Accuracy Score (97.65%), and Area Under the Receiver Operating Characteristic Curve (AUC-ROC) score (96.46%), indicating not only its ability to correctly classify a higher percentage of the test dataset but also its robustness in handling imbalanced classes and in distinguishing between class labels. ResNet50 follows closely, showing similar strengths, particularly in its ability to differentiate between classes as evidenced by its AUC-ROC score (96.68%). Vission Transform present competitive, though slightly inferior, performances, with challenges in balanced accuracy, which suggests difficulties in inequitable performance across classes. metrics,Äîmost notably in balanced accuracy (86.67%) and AUC-ROC (84.73%), which hints at substantial difficulties both in dealing with class imbalance and in distinguishing between class labels effectively[6].

Confusion Matrix

Vgg19

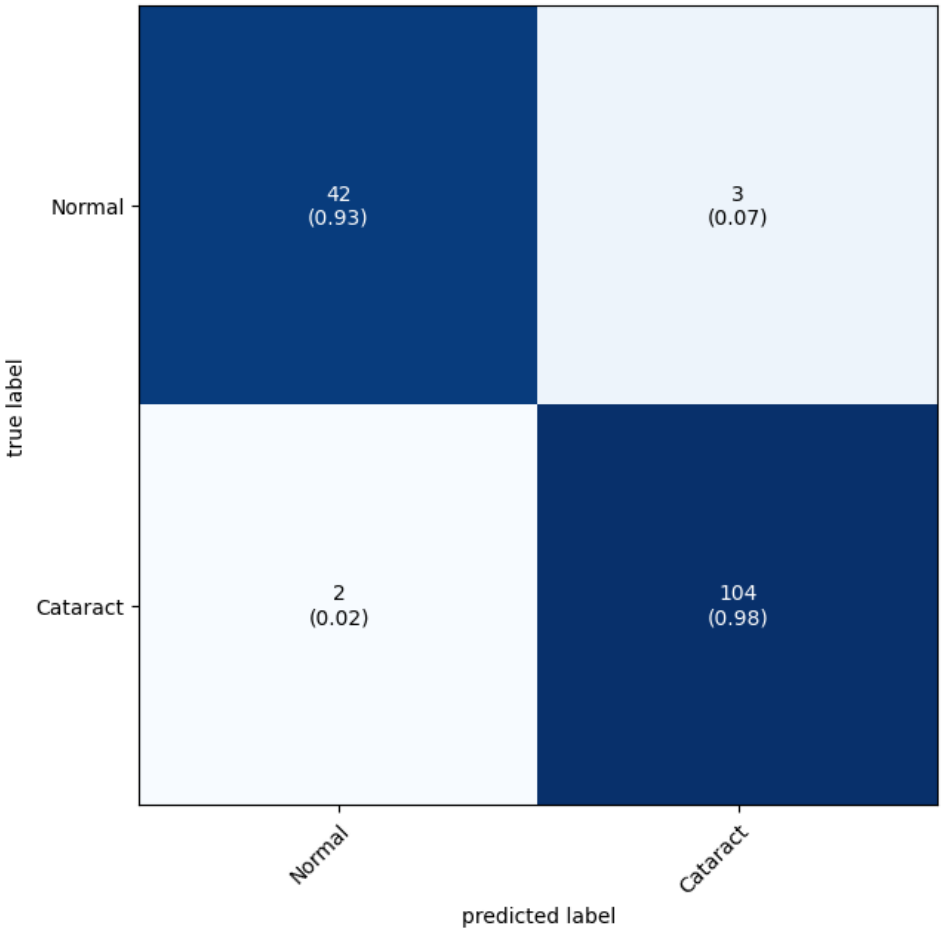


Figure 6.11: Confussion matrix for Vgg19

ResNet50

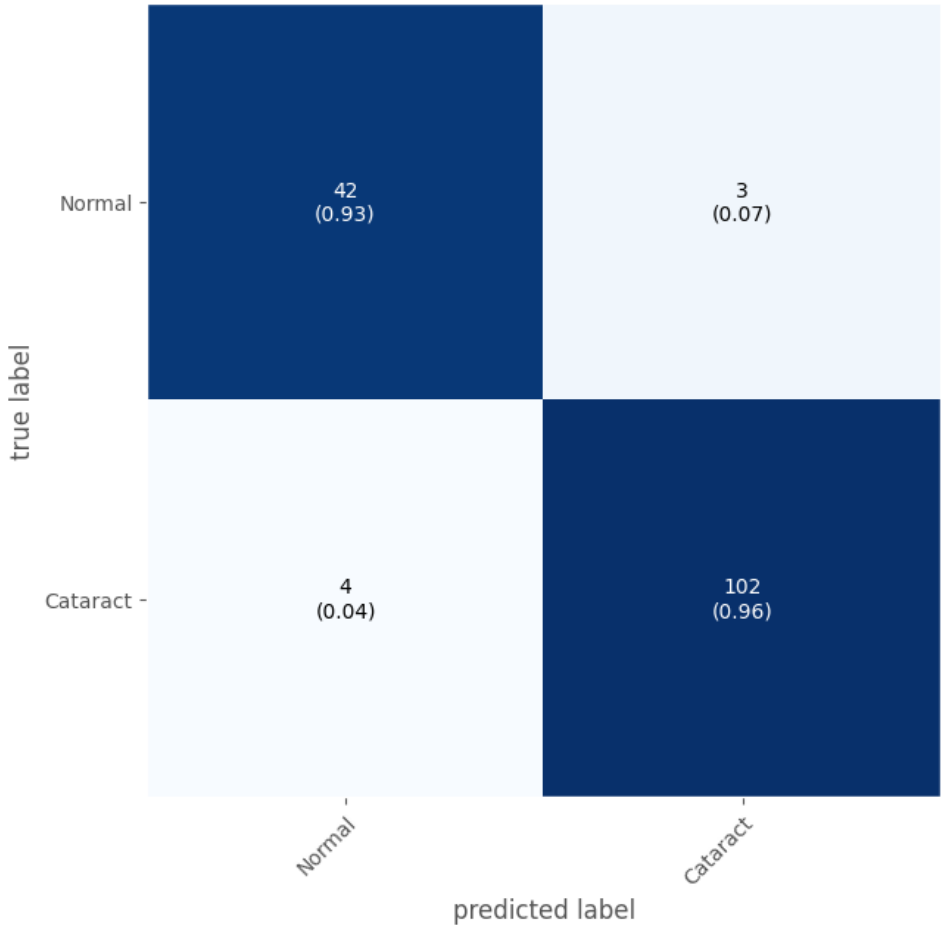


Figure 6.12: Confussion Matrix for ResNet50

6.3.1 Limitation

The primary limitation I faced was the low accuracy of the image classification model, which did not improve despite applying class balancing strategies. Another significant challenge was the lack of computational resources, which notably impacted the accuracy as extensive hyperparameter tuning could have potentially enhanced the model,Äôs performance. In this context, I considered using Keras Tuner, an automated hyperparameter tuning tool that optimises model configurations within specified limits to achieve the best possible performances. Keras Tuner systematically tests a set of hyperparameters, evaluating their effectiveness using a defined metric. However, when I attempted to use it for tuning ResNet50, the process demanded extensive computational power and ran for an extended period, straining the resources I was funding out of pocket. Time management also proved problematic as I spent a considerable amount of time coding until the very end, which prevented me from recognising and addressing mistakes early in the process. Additionally, I encountered issues with data leakage among the train, test, and validation datasets, which initially resulted in misleadingly high accuracies. Fortunately, I identified and rectified this error in time. Another oversight was my failure to run LIME on augmented images to assess the explainability method,Äôs responsiveness to changes in making predictions. This would have provided deeper insights into how well the explainability techniques adapted to new data scenarios, thereby enhancing the robustness and reliability of the diagnostic predictions[5].

6.3.2 Conclusion and Future Work

This study aimed to develop a transparent, computer-aided diagnostic system for ocular disease recognition.Despite the advanced methodologies applied, the system faced significant challenges in achieving high accuracy, which was further compounded by computational resource limitations. The highest achieved accuracy of the image classification models was modest, reflecting the difficulty of balancing model complexity with performance on a complex dataset like ODIR-5K[21].
Although the use of Explainable AI (XAI) techniques such as LIME, SHAP, and Grad-CAM enhanced the interpretability of the models, the overall effectiveness was constrained by the models performance.
Limitations included the low accuracy of predictions and the substantial computational resources required for optimal hyperparameter tuning and model training, which were not fully accessible due to budget
constraints. These factors hindered the potential of achieving better performance and more robust explanations of the model decisions.

Future Work

Future work should focus on:

- **Enhancing Model Accuracy:** Exploring more sophisticated data augmentation techniques and advanced neural network architectures could potentially improve model accuracy.
- **Sourcing External Funding for Computational Resources:** Utilizing cloud computing platforms with the help of external funding may provide the necessary power for more extensive hyperparameter tuning and training deeper models.
- **Extending Explainability:** Applying XAI techniques to more complex models or developing new methods tailored to specific medical imaging tasks to improve both transparency and user trust.
- **Clinical Validation:** Collaborating with medical professionals to validate the model's predictions against clinical outcomes, thus ensuring its utility in practical scenarios.

In conclusion, while the project faced several setbacks primarily due to resource constraints, it laid a foundation for further research into the application of deep learning and explainable AI in medical diagnostics, particularly in the detection and classification of ocular diseases.

REFERENCES

[1] T. Pratap and P. Kokil, "Computer-aided diagnosis of cataract using deep transfer learning," *Biomedical Signal Processing and Control*, vol. 53, p. 101533, 2019.

[2] "Using transfer learning technique as a feature extraction phase for diagnosis of cataract disease in the eye," no. 1.

[3] S. M. Saqib, M. Iqbal, M. Zubair Asghar, T. Mazhar, A. Almogren, A. Ur Rehman, and H. Hamam, "Cataract and glaucoma detection based on transfer learning using mobilenet," *Heliyon*, vol. 10, September 2024.

[4] X. Leng, R. Shi, Y. Wu, S. Zhu, X. Cai, X. Lu, and R. Liu, "Deep learning for detection of age-related macular degeneration: A systematic review and meta-analysis of diagnostic test accuracy studies," *PLOS ONE*, vol. 18, pp. 1–20, 04 2023.

[5] M. S. Khan, N. Tafshir, K. N. Alam, A. R. Dhruba, M. M. Khan, A. A. Albraikan, and F. A. Almalki, "[retracted] deep learning for ocular disease recognition: An inner-class balance," *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, p. 5007111, 2022.

[6] A. Bhati, N. Gour, P. Khanna, and A. Ojha, "Discriminative kernel convolution network for multi-label ophthalmic disease detection on imbalanced fundus image dataset," *Computers in Biology and Medicine*, vol. 153, p. 106519, 2023.

[7] K. Simonyan, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[8] B. Κατσάρα, "Prediction of the retention time of natural product metabolites using transfer learning strategies," 2024.

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

[10] J. Wang, L. Yang, Z. Huo, W. He, and J. Luo, "Multi-label classification of fundus images with efficientnet," *IEEE access*, vol. 8, pp. 212499–212508, 2020.

[11] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

[12] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, pp. 1–40, 2016.

[13] S. Mukherjee, “The annotated resnet-50,” *Towards Data Science*, vol. 18, 2022.

[14] A. Kumar, L. Nelson, and S. Gomathi, “Cataract prediction with vgg19 architecture using the ocular disease dataset,” in *2024 2nd World Conference on Communication & Computing (WCONF)*, pp. 1–7, IEEE, 2024.

[15] S. Umrani and H. M. Pande, “Clahe-enhanced transfer learning with vgg19 & densenet201 for ocular disease classification,” in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pp. 1–8, IEEE, 2024.

[16] D. Lin, J. Chen, Z. Lin, X. Li, K. Zhang, X. Wu, Z. Liu, J. Huang, J. Li, Y. Zhu, *et al.*, “A practical model for the identification of congenital cataracts using machine learning,” *EBioMedicine*, vol. 51, 2020.

[17] R. Sigit, M. Kom, M. B. Satmoko, D. K. Basuki, and S. Si, “Classification of cataract slit-lamp image based on machine learning,” in *2018 International Seminar on Application for Technology of Information and Communication*, pp. 597–602, IEEE, 2018.

[18] Y. Kumar and S. Gupta, “Deep transfer learning approaches to predict glaucoma, cataract, choroidal neovascularization, diabetic macular edema, drusen and healthy eyes: an experimental review,” *Archives of Computational Methods in Engineering*, vol. 30, no. 1, pp. 521–541, 2023.

[19] M. T. Ribeiro, S. Singh, and C. Guestrin, “" why should i trust you?" explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

[20] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: visual explanations from deep networks via gradient-based localization,” *International journal of computer vision*, vol. 128, pp. 336–359, 2020.

[21] A. Bhati, N. Gour, P. Khanna, and A. Ojha, “Discriminative kernel convolution network for multi-label ophthalmic disease detection on imbalanced fundus image dataset,” *Computers in Biology and Medicine*, vol. 153, p. 106519, 2023.

[22] A. I. Herrera-Chavez, E. A. Rodríguez-Martínez, W. Flores-Fuentes, J. C. Rodgíruez-Quiñonez, J. C. García-Gallegos, O. H. Montiel-Ross, F. F. Gonzàalez-Navarro, and O. Sergiyenko, “Multi-label image classification for ocular disease diagnosis using k-fold cross-validation on the odir-5k dataset,” in *2024 IEEE 33rd International Symposium on Industrial Electronics (ISIE)*, pp. 1–6, IEEE, 2024.

[23] A. Ivanescu, S. Popescu, A. Braha, B. Timar, T. Sorescu, S. Lazar, R. Timar, and L. Gaita, “Diabetes and cataracts development,Äîcharacteristics, subtypes and predictive modeling using machine learning in romanian patients: A cross-sectional study,” *Medicina*, vol. 61, no. 1, p. 29, 2024.

[24] D. Jameel and A. M. Abdulazeez, “Ocular disease recognition based on deep learning: A comprehensive review,” *The Indonesian Journal of Computer Science*, vol. 13, no. 3, 2024.

[25] S. E. Alexeeff, S. Uong, L. Liu, N. H. Shorstein, J. Carolan, L. B. Amsden, and L. J. Herrinton, "Development and validation of machine learning models: electronic health record data to predict visual acuity after cataract surgery," *The Permanente Journal*, vol. 25, 2021.

[26] G. D. Aranha, R. A. Fernandes, and P. H. Morales, "Deep transfer learning strategy to diagnose eye-related conditions and diseases: An approach based on low-quality fundus images," *IEEE Access*, vol. 11, pp. 37403–37411, 2023.

[27] T. Li, J. Stein, and N. Nallasamy, "Evaluation of the nallasamy formula: a stacking ensemble machine learning method for refraction prediction in cataract surgery," *British Journal of Ophthalmology*, vol. 107, no. 8, pp. 1066–1071, 2023.

[28] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, "Not just a black box: Learning important features through propagating activation differences," *arXiv preprint arXiv:1605.01713*, 2016.

[29] H. Bakır and Ş. Yılmaz, "Using transfer learning technique as a feature extraction phase for diagnosis of cataract disease in the eye," *Uluslararası Sivas Bilim ve Teknoloji Üniversitesi Dergisi*, vol. 1, no. 1, pp. 17–33, 2022.

[30] R. R. Maaliw, A. S. Alon, A. C. Lagman, M. B. Garcia, M. V. Abante, R. C. Belleza, J. B. Tan, and R. A. Maaño, "Cataract detection and grading using ensemble neural networks and transfer learning," in *2022 IEEE 13th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pp. 0074–0081, IEEE, 2022.

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

[32] K. Vijay, S. Vishnu, S. Sankar, and E. Manohar, "Innovative approaches in cataract detection: Exploring transfer learning and image segmentation techniques," in *2024 International Conference on Smart Systems for applications in Electrical Sciences (ICSSES)*, pp. 1–6, IEEE, 2024.