Enhancing Cataract Diagnosis: Towards Fine-grained Cataract Classification using Deep Learning and Progressive Image Processing Strategies

by

Sadman Safiur Rahman
21301411
Md. Ratul Mushfique
21301418
Mohammed Adib Ahnaf Hamim
21101054
Md. Mezbha Ul Haq Fahim
21301243

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

Department of Computer Science and Engineering
Brac University
June 2025

Declaration

It is hereby declared that

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

Student's Full Name & Signature:

Sadman Safiur Rahman
21301411

Md. Ratul Mushfique
21301418

Mohammed Adib Ahnaf Hamim
21101054

Md. Mezbha Ul Haq Fahim
21301243

Approval

The thesis titled "Enhancing Cataract Diagnosis: Towards Fine-grained Cataract Classification using Deep Learning and Progressive Image Processing Strategies" submitted by

1. Sadman Safiur Rahman(21301411)
2. Md. Ratul Mushfique(21301418)
3. Mohammed Adib Ahnaf Hamim(21101054)
4. Md. Mezbha Ul Haq Fahim(21301243)

Of Spring, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on June 20, 2025.

Examining Committee:

Supervisor:
(Member)

Mr. Md. Tawhid Anwar
Senior Lecturer
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)

Mr. Md. Sabbir Ahmed
Lecturer
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

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

Head of Department:
(Chair)

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

Abstract

The sense of seeing the world is one the most precious gifts of human life. However, cataract remains a major public health threat because of its high prevalence and frequent causing severe vision impairment or even blindness. Early diagnosis continues to be critical but reliance on ophthalmologists restricts availability, particularly in low-resourced areas. To solve this problem, in this paper, we present an AI-assisted fine-grained multiclass cataract classification framework, based on effective deep learning and progressive image preprocessing techniques. Unlike other prior works that only used public datasets, we assembled a real-world anterior eye image dataset acquired from hospitals in the city of Dhaka (Bangladesh), and annotated by experienced ophthalmologists. Our model is based on a variety of deep learning architectures that include VGG19, ResNet101, InceptionV3, Xception, EfficientNetB2, DenseNet121, MobileNetV2, FastViT, MobileViT, and ViT_B16, which were fine-tuned, using transfer learning, for the task of multiclass classification with four different cataract types as well as normal cases. A full preprocessing pipeline that includes resizing, normalization, augmentation and contrast enhancement was applied to enhance the generalization capability among the models. For further boost in performance, we utilized two ensemble methods, bagging and stacking, where bagging combines the predictions of those top-performing models via averaging the probabilities for final decision, whereas a stacking meta-learner (logistic regression) combines the base models output for improved estimation. Both systematic ensembler approaches yielded striking improvements in accuracy and performance across varied test conditions. Furthermore, Grad-CAM visualizations validated the clinical interpretability of the system, as heatmaps consistently indicated anatomically correct areas related to cataract pathology. This work introduces a scalable, clinically viable, and resource-efficient AI framework for cataract screening, which provides the groundwork for wider dissemination in ophthalmic diagnostics.

Keywords: Cataract, Progressive image processing, Convolutional neural networks, Attention mechanisms, Ophthalmology, Medical imaging, Transformer model, Ensemble Learning, Multi-class Classification

Dedication

- This work is dedicated to all cataract patients, whose determination and bravery motivated our search for improved methods of diagnosis.

- To the doctors and ophthalmologists who work tirelessly to restore hope and vision to all those who are affected.

- And to the spirit of medical science innovation, which brings us one step closer to innovation that may finally end the sufferings of mankind.

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption..

Secondly, we would like to thank our supervisor Mr. Md. Tawhid Anwar sir and our co-supervisor Mr. Md. Sabbir Ahmed sir for their advice and motivation throughout this work. Their frequent guidance and continuous encouragement played a major role from the start to the end of the whole project.

Thirdly, heartfelt appreciation goes to Prof. M. A. Bashar Sheikh and his team for helping us obtain and annotate the eye image dataset. His expert opinions have also helped us gain new knowledge and guide us to find the best solution for our work.

We also want to thank the panel members for their expert insights and valuable suggestions on future deployment methods. The feedbacks that we received from them greatly improved our work and their ideas have helped us understand how we can bring more improvement to the work as well as find efficient methods to deploy in real-life settings.

We would also like to appreciate all the people who permitted us to use their eye samples for our dataset. We could not have done this without your cooperation and you would be glad to know that your contributions have played a role in not only our success but also in helping the whole of mankind.

Finally, we would like to extend a special thanks to our parents, family members, friends and fellow peers for their boundless love, prayers, and support. Your belief in us made this dream come true.

Table of Contents

List of Figures

Chapter 1

Introduction

Cataract is a disease that is identified by the clouding of the eye's lens. This happens when too much protein builds up in the lens and changes its normal shape. This opacity problem could cause the patient's vision to be unclear. Among the several factors that could lead to cataracts include aging, smoking, radiation, diabetes, and so on. Also, many people are losing their vision because of cataracts; as a result, it is crucial to diagnose these disorders precisely and promptly so that patients can receive the appropriate therapy. Cataract can account for more than half of cases of blindness. In 2010, cataracts were associated with around 33.3% of all cases of blindness and 18.4% of all cases of visual impairment globally. [31][32]

Cataracts are also well recognized as an age-related condition; according to research, 15% of westerners over the age of 50 and 40% of those over the age of 50 in developing nations experience this eye condition. In next 25 years, 40 million people will suffer with cataract as the elder population rises. [1]

At present, cataract diagnosis is largely dependent on ophthalmologists, which presents accessibility challenges and requires more efficient screening procedures.In order to overcome these challenges, our research developed a highly accurate cataract classification system by combining deep learning with advanced image processing methods. Nevertheless, this also has a few drawbacks, such as a scarcity of benchmark datasets, suboptimal model performance, and challenges with generalizability and real-world clinical application. Additionally, cataracts in the early stages are not detectable by the current models. Our objectives are to enhance interpretability, optimize model performance, ensure clinical validity, reduce computational complexity and inference time, and increase dataset diversity and quality. To overcome these drawbacks, we present a deep learning-based classification system for cataract in this study, which uses convolutional neural networks (CNNs) and an advanced image preprocessing methodology. To this end, our approach comprises transfer learning, data augmentation, and model interpretability tools, in order to create a trustable, accurate and explainable AI-assisted diagnostic tool. One of the key distinction of this work is the creation of a high-quality real-world dataset of images acquired in a clinical setting in hospitals, yet from the context of Dhaka that better captures the variabilities seen in practical settings. The models were validated by hard metrics and visualization skills for robustness and trust in decision making. Our study is not only focused on enhancing the classification performance, but also to guaran-

tee the robustness of these models in various clinical scenarios. Through thorough testing and evaluation of the model, our objective is to offer an affordable, efficient, and scalable AI-driven diagnostic tool that can greatly improve the early cataract detection, especially in low-resource scenarios. By completing the mentioned goals, we intend to incorporate our system into clinical processes, giving ophthalmologists more efficient diagnostic models to improve cataract identification and management.

1.1 Research Problem

Despite a number of models and methods that have emerged in the field of cataract classification, none have been identified without having a major obstacle. To start with, much of the research that is conducted to create these models often use a very small dataset. For example, in [29] the ACCV model based on DL, the dataset consists of the video data of only 38 people and in another research [16] based on ML, the dataset consists of only 31 patients from a single hospital situated in southern Nigeria. Small datasets are very helpful to train models and detect its usability, but the actual performance and accuracy of the model cannot be concluded without a vast amount of data since generalisation is a major issue for a model's acceptance.

Size is not the only factor in the case of problems seen in the dataset generalisation. The lack of quality and diversity among the data is another concerning issue in finding datasets. For instance, in [46] the study used 1600 fundus retinal images sourced from various open-access databases, but there was a scarcity of benchmark datasets for cataract detection, impacting the generalizability and robustness of the developed model. Similarly, in [47], authors used a softmax classifier and a CNN model to achieve 92.7% accuracy but had a major drawback in finding a large labelled dataset with the image quality to extract retinal features such as tiny blood vessels accurately. Due to this same situation a lot of the existing models do not perform well in real life clinical scenarios where image quality and lighting conditions might significantly differ from the fundus images in the datasets. For example, in [49] the AI model used alongside CNN does not have algorithms' training sets that can adequately reflect real-world clinical settings, leading to potential biases and decreased efficiency in patient outcomes.

Feature extraction limitation is another problem that can be seen in various papers. In [24] the model uses ResNet18 and GLCM to extract high-level features and texture features but is unable to capture relevant information of fundus images that may lead to errors in grading. Furthermore, in [47] the methods used rely heavily on hand-crafted feature extraction, leading to inconsistencies and inefficiencies. In addition, the model failed to properly grade and identify cataracts because it failed to take into account micro blood vessels details on retinal pictures. But, it is also a matter of fact that even advanced DL models struggle with extracting minute retinal features accurately.

Just like feature extraction, limitations can also be detected in getting suboptimal model performance from constructed models. This lack of model performance is caused by various factors. In [41], the model performance is hampered by the misclassification of a notable portion of eye images in model prediction phase. In [37],

There is a lack of standardised metrics for evaluating the performance of AI diagnostic tools. In [24] model performance is predicted to decrease when generalising to new data due to the model complexity and overfitting.

While model performance can be gained by various methods, without safety and efficacy in the process, a model cannot be deemed perfect for use in real-life scenarios. Such as in [33] the model is based on AI algorithms based on DL techniques that offered comprehensive solutions for improved patient outcomes but had critical concerns over clinical acceptance and data security. Data privacy and security issues are also major factors that have raised eyebrows in numerous models, especially the ones that associate mobile applications with their models such as [29] where a mobile app is used to collect and transmit patient eye videos for remote screening and diagnosis and in [45], an AI smartphone app is used find and gauge the severity of cataracts compared with ophthalmologists guided by the slit-lamp exam.

1.2 Research Objective

The goal of this work is to construct a scalable, interpretable, and accurate deep learning solution for a multiclass cataract classification system which performs well in clinical scenarios, and is robust to real world imaging setups. In contrast to past works, which used idealized datasets or binary classification, this methods tackles the realistic challenges needed for real-world screening by creating a custom dataset clinically-verified and comprising both slit-lamp as well as predominantly mobile-captured anterior eye images, to ensure the model's ability to deal with the variability and noise found in field-level data collection, specifically in rural and marginalized areas.

A variety of deep learning models, VGG19, ResNet101, Xception, InceptionV3, EfficientNetB2, DenseNet121, MobileNetV2, ViT-B16, FastViT, and MobileViTV2 were trained and tested. Each architecture was additionally finetuned from transfer learning and modified for the multiclass classification that the model should also recognize: Normal, Nuclear Cataract, Cortical Cataract, Posterior Subcapsular Cataract, and Hypermature Cataract. The top CNN and ViT based models according to performance scores were integrated in ensemble bagging and stacking method with the meta-classifier based on logistic regression for more reliable, less overfit, and better generalized predictions for images across sources.

Additional objectives include:

- Using integrated image preprocessing pipeline, including resizing, normalization, contrast enhancement and data augmentation, to make the model robust to the variation in resting state scanning.
- Adding Grad-CAM visualization to interpret and validate the attentional region of the model focus areas in relation to clinically meaningful features.
- Exploring performance comparison across architectures to uncover accuracy trade-offs in the presence of different cataract presentations.

The objective of this study is to offer a robust and medically pertinent AI tool for the screening of cataracts by deep learning, which can support clinical decision-making in various healthcare settings which may not always adhere to standard imaging requirements.

1.3 Workflow

To establish a clinically usable, fine-grained cataract classification system, the process in our workflow was grounded within clinical demands and accounted for both technical accuracy and ethical data acquisition. As opposed to the majority of prior works which are based solely on public data, we began our data acquisition by teaming up with ophthalmologists and volunteers from local eye hospitals in Dhaka. Patients and volunteers readily supplied high-quality anterior eye images from slit-lamp and smartphone-based acquisition under clinical supervision. All samples were confirmed and classified by certified eye practitioner for clinical reliability. This people-centric methodology enabled us to build a real-world dataset with a diversity of light, pupil dilation, and cataract stages conditions that are frequently not covered in the benchmark datasets. The images collected were preprocessed by resizing to 224×224 dimensions to meet the input demand of deep learning models. A large augmentation pipeline introduced variability that was essential for generalization that included gaussian blur, horizontal flip, random zoom, brightness change, and salt-and-pepper noise, the latter in line with the dust, grain, and digital noise that may be encountered in medical images captured on mobile devices. These reinforcement techniques allowed the models to learn from imperfections and variability as is often encountered in the real world. The model evolution extended from the classical architectures like VGG19, ResNet50 and InceptionV3 to modern architectures such as ResNet101, Xception, EfficientNetB2, MobileNetV2, DenseNet121 and transformer-based architecture like ViT-B16 and FastViT and MobileViTV2. All models were also pretrained with ImageNet weights and fine-tuned with a shared classification head consisting of GAP, dense layers with ReLU activation, dropout (rate 0.5), and a five-unit softmax output for multiclass prediction. The five classes of targets were: Normal, Nuclear Cataract, Cortical Cataract, Posterior Subcapsular Cataract and Hypermature Cataract. After both personal training and analysis, we bagging ensemble model as well as applied a stacked ensemble learning model. The best CNN-based and ViT-based models were chosen according to performance metrics and ensembled together with all the three ViT-based models through logistic regression as the meta-classifier. The consistency and robustness of the predictions were enhanced by this strategy for diverse test images. Learning rate was held at 0.0001 for all models, with 40 epoch training and early stopping and model saving. Clinically, Grad-CAM was utilized to interpret model attention and verified that the decisions were made due to pathologically related regions (e.g., lens opacity and central focus). Lastly, the outputs of the system were examined by an experienced ophthalmologist and verified that the predictions were consistent with clinical expectation. The expertise review confirmed that the model can support in a real application for screening and decision in low resource settings.

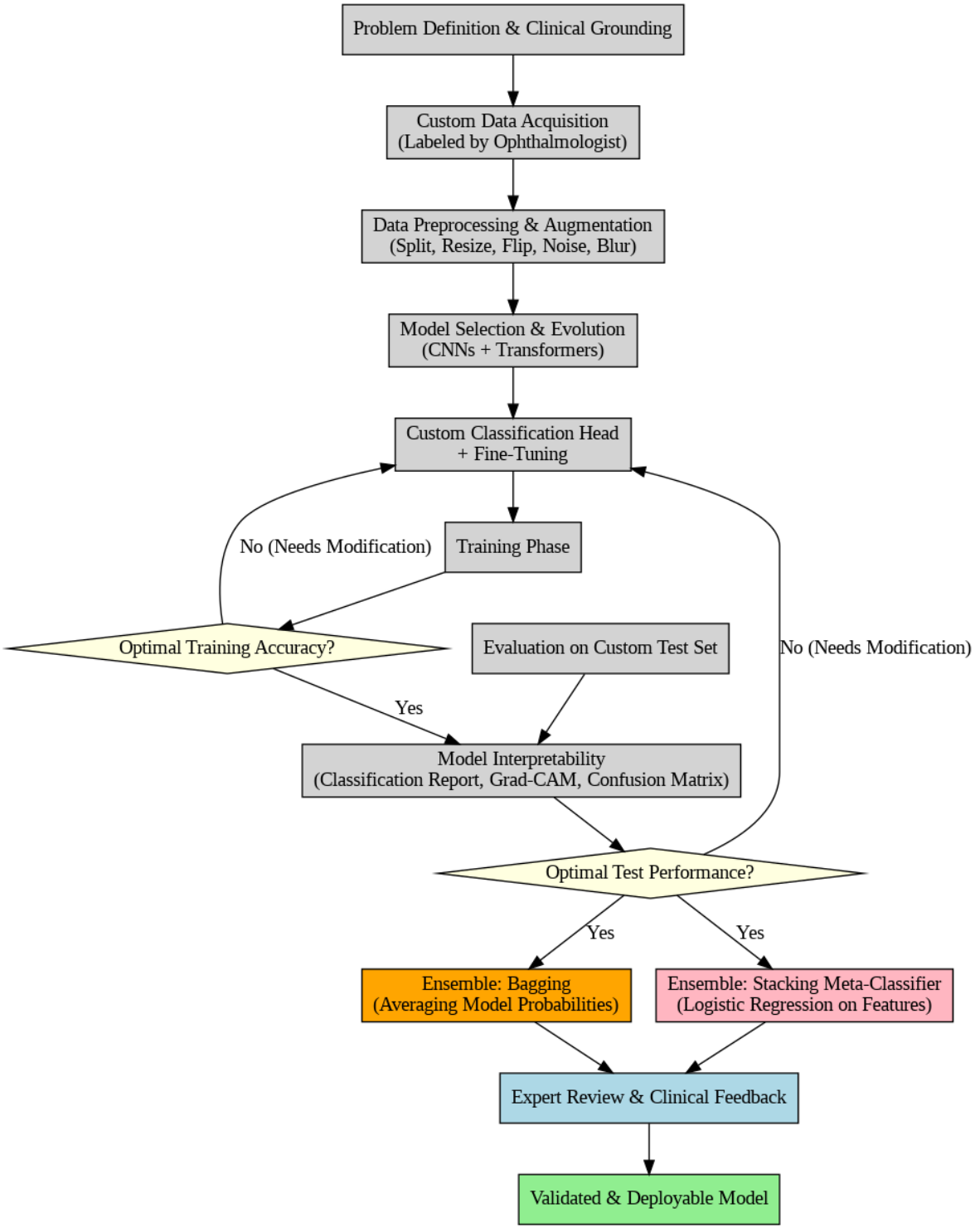

Figure 1.1: Workflow Diagram

Chapter 2

Literature Review

The field of classifying eye cataracts has shown promising results with recent advances in numerous methods such as deep learning, machine learning, artificial intelligence, image processing, etc. The models that have been proposed in this field also have variances in their architecture and design. The model structure can be new or a refinement of one from a prior model. Also It could easily adopt one or more methods or approaches to create one like incorporating deep learning with image analysis. Furthermore, a model's technique in categorizing cataracts can also be novel or an improvement and could consist of several algorithms that have been applied singly to build different models entirely. For instance, CNN gets much attention for extracting features from the images, and when compounded with RNN, the new combined model gets to extract features sequentially from the frames of a video.

Deep Learning approaches such as CNN have created revolutionary changes in the field of cataracts just like methods of Machine Learning. Machine learning has been quite famous for a long time in research for the medical sector based on IT. Traditional, ML algorithms such as support vector machines (SVM) and decision trees have been renowned on detecting relevant details from data. On the field of classifying cataracts, ML methods are also giving quite significant results. For example, Egejuru et al. (2017) identified risk factors of cataracts using Naïve Bayes, Decision Trees and the Multi-layer Perceptron (ANN) classifiers. The model was validated by mental health professional using WEKA software and the dataset was collected from a hospital in southern Nigeria which showed 9 demographics and 17 risk factor variables. While identifying these variables, ANN demonstrated 100% accuracy in identifying hidden patterns in the variables. These patterns were used to create the predictive model based on C4.5 Decision Trees and Naïve Bayes classifiers. The model was quite efficient as both methods had accuracy rates of 87% and 84%, respectively. [16]

In the case of DL, however, accuracy rates have been quite better than the ML approaches. Thus, inspired by Zhang et al. (2019), a six-level cataract grading system was developed that integrates ResNet18 for detecting high-level features and the GLCM matrix for detecting texture features. To grade these features a frame is used that contains two SVM classifiers and an FCNN consisting of two fully connected layers. The SVMs generate the probability outputs for each fundus image while the FCNN serves as a meta-learner for producing final classification results.

The authors have gone to great lengths in the collection of data by collecting 1352 images and categorising them according to the severity of the disease. All of these images were captured by Topcon and verified by ophthalmologists. Their hard work came to fruition as the proposed grading method achieved an average accuracy of 92.66%, with a peak accuracy of 93.33%. On the other hand, for four-level cataract grading, the method reaches 94.75% accuracy, surpassing existing methods by at least 1.75%, proving the method's higher grading performance in the field. [24]

The CatractNet model was developed by Junayed et al. (2021) using a specialized 16-layer deep-learning neural network. Its purpose is to automate cataract identification, overcome the drawbacks of human feature extraction, and lower the computing cost of convolutional neural networks (CNNs). The authors used HRF, FIRE, ACHIKO-I, IDRiD, DRIVE, and a color fundus image dataset, and employed all the data augmentation techniques in order to avoid overfitting and channel out the difference between testing and training data sets. It was found that CataractNet outperformed other pre-trained CNNs of VGG-16, VGG-19, MobileNet, Inception-v3 and ResNet-50 having an accuracy of nearly 99.13% and the time of running of the program was 1035 seconds. But, a major limitation was found which is that the model is unable to differentiate age-related cataracts and cannot precisely grade cataracts. [30]

An approach that was anticipated by Yu et al. (2019) to automate phase classification from manually pre-segmented phases in cataract surgery recordings was the use of deep learning and machine learning algorithms. The authors gathered data from one hundred cataract surgeries performed by residents and instructors in an ophthalmology residency program between 2011 and 2017. There were five distinct algorithms used: a support vector machine, an RNN, a CNN, an image-data input, an image-data input with a CNN-RNN, and instrument labels with a CNN-RNN. Results showed that the unweighted accuracy of the five algorithms ranged from 0.915 to 0.959, with minor differences between them. Surprisingly, the image-only CNN-RNN had a substantially greater AUC than the cross-sectional CNN for detecting instrument labels from images. The degree of specificity was also consistently high with the level of sensitivity being 0. 005 to 0. 974 and the whole range of precision depends on a certain coefficient and can be within the scale of 0. 283 to 0. 963, this result means that the deep learning approach can be applied to identify the phase of cataract. [23]

A novel approach to performing the complete classification of cataract videos through ACCV without the need for manual feature extraction is presented by Hu et al. (2021). The authors used a YOLOv3 pupil detection approach and it was used to perform frame selection while also estimating a classification, as well as a probability from the pixel data through a process of regression. The study utilised the observations from 76 eyeballs that belonged to 38 unique individuals and acquired the data with the help of an iSpector-mini mobile phone slit lamp. Here, ACCV utilized the iSpector Mini slit lamp on the phone to read the video ahead from the previous segment, then take each frame to YOLOv3. Finally, generated by UCF, the complete video files containing blanks are used by YOLOv3 to complete the automatic classification. Results showed that the ACCV algorithm outperforms

other methods such as VGG19, Inception-v3, Resnet50, MobileNet, and Xception, and produces more accurate classifications. Furthermore, the utilization of YCrCb space significantly improves the real-time performance and recognition speed of the method. This makes it possible for the method to precisely identify the pupil position in each frame and classify the light knife section of each video frame. [29]

Son et al. (2022) delivered an AI platform that was based on deep learning, technique which diagnoses and grade cataract using slit lamp and retroillumination lens photos consequently applying LOCS III system. This kind of research used a convolutional neural network trained with 918 slit lamp photographs to identify NO as well as NC cataracts, while its input data is expanded with 435 retro illumination photographs to cover CO or PSC cataracts diagnostic cases- The model demonstrated the combination of strong diagnostic value and grading prediction for a wide range of types of cataract. The authors conducted the training and validation of AI models through several techniques that include local detection network, data augmentation, generalized cross-entropy loss and so on. They were also reinforced by other algorithms like ResNet18, WideResNet50-2 and ResNext50. The diagnostic accuracy of the CAD system for cataracts was nearly perfect with AUC above 0.99 and around 98% specificity for NO and NC using slit-lamp photos, and AUC up to 0.97 and accuracy over 92% detection of CO and PSC with retroillumination images . The grading performance was great too; it had AUCs around 0.90 joined with accuracies above 87%. [36]

To identify and categorize cataracts in their early stages, Yadav et al. (2023) presented a new approach that integrates deep learning (DL) with the 2D-discrete Fourier transform (DFT) spectrum of fundus pictures. By transforming fundus images into the frequency domain using 2D-DFT and employing a DL model for feature extraction, followed by classification with a softmax classifier, their system achieves a four-class accuracy of 93.10%, surpassing previous state-of-the-art methods. The study also incorporates a module to filter out poor-quality images, ensuring robust performance. This approach highlights the potential for developing cost-effective, accurate diagnostic tools that can improve accessibility to early cataract detection and treatment.[47]

In another research, Yadav et al. (2023) enhanced their deep learning-based computer-assisted diagnostic method for cataract screening on fundus retinal images. They used fundus retinal images from different open access databases amounting to 1600 images. A proficient ophthalmologist classified these photos into four categories (mild, moderate, normal, and severe). The research achieved a 97.34% accuracy in classifying and grading cataracts by using an ensemble learning method that incorporated a deep convolutional neural network with three classification techniques: Support Vector Machine, Naive Bayes, and Decision Tree classifiers. This development could lead to an inexpensive test that can help catch the earlier stages of cancer, which would be a great stride towards global health. [46]

Similarly, Ganokratanaa et al. (2023) devised a new CNN-based model in which they performed pre-processing using machine learning techniques and predicted the presence of cataracts in eye images. They use a Convolutional Neural Network

(CNN) architecture, including convolutional layers for feature extraction and max-pooling layers for dimensionality reduction. Training the model involves many steps such as loading data, preprocessing and compiling the model. Small but important hyper-parameters we use such as: learning rate, batch size, epochs to get the optimal model for your data. Training and testing dataset of 3500 eye images with 80 percent training and 20 percent testing used for evaluation of CNN model. The evaluation metrics show the model's performance using various metrics such as accuracy, sensitivity, specificity, and so on, demonstrating that the model is effective in distinguishing between cataract-affected and normal eyes. In the end, authors believe that this study will it to a level of clinical settings to help in early diagnosis and treatment of cataract from other related diseases, among other things. [41]

According to Goh et al. (2020) the condition that is characterised by the formation of a cloud on the lens of the eye is called cataracts, which is the main reason for visual impairment,and with the growth of the number of elderly people. They usually depend on adept ophthalmologists for detection constituting a noteworthy quandary for developing countries. Traditional IOL power calculations are not perfect either and can vary with each individual patient at times. The application of AI, primarily deep learning, generates optimistic results in ophthalmology diagnosis, as several algorithms, starting from the detection and grading of cataracts with slit lamp and colour fundus photos, may assist ophthalmologists in the recognition and subsequent grading of the examined ailment, thereby decreasing the workload on ophthalmologists. Such reference data derived formulas such as the Ladas Super Formula or the Hill-RBF appear to have better accuracy in comparison to conventional formulas. It is still, however, a promising technology and one that has certain limitations applicable to the consideration of atypical eyes or individuals who have had refractory surgery. It is often challenging and time-consuming to develop and validate trained AI systems and large clinical data sets are required. Cataract itself could be ameliorated by AI especially in low-middle income regions while cataract surgery education could similarly be boosted by the assistance of advanced AI. [26]

Moreover, in a cross-sectional study by Vasan et al. (2023), measured the ability of e-Paarvai, an AI smartphone app, to find and gauge the severity of cataracts compared with ophthalmologists guided by the slit-lamp exam. Carried out from January to April 2022 in South India, this study recruited 1,407 patients more than 40 years old and with a poor vision of their eyes recruited 2,619 eyes. The features identified by the app offered higher sensitivity at 96% but lower specificity at 25%, with an overall observed accuracy of 88%. Compared to the other types of lens ever observed, it was most accurate in identifying immature cataracts while least accurate in differentiating between mature cataracts, posterior chamber intraocular lenses, and clear lenses. Based on them, the authors of the study concluded that patient demographics and visual acuity could be useful for increasing the diagnostic accuracy of the application. Certainly, while they may provide limited benefits with short video consultations in cataract screening at the moment, the e-Paarvai app essentially demonstrated a feasible solution that is cost-effective that can be utilised to address the issue in distant areas. [45]

Another AI model for cataract detection using fundus pictures was suggested by Xu et al. (2022) with the goal of preventing interference. It developed a two-stage model, a quality recognition model for evaluating quality and a convolutional neural network (CNN) for classifying—to tackle the problem of low-quality photos impacting diagnosis. Compared to binary-class models, it significantly improved the cataract detection rate by 10%. It scored well in both internal and external validation while ensuring the disease is diagnosed from low quality images. This AI tool could assist in cataract triage through informing the insurance recommendation as well as policy development by governments. [38]

Due of their impact on the worldwide blindness rate, Gan et al. (2023) set out to develop an AI-powered diagnostic tool for cortical cataracts. Two platforms were developed using 647 high-quality anterior segment images: one that does data segmentation automatically and another that does it manually. The automatic platform achieved 94% in the assessment while the student-group achieved only 72%. 59% training and 84. This means that the automated form achieved 50% test accuracy while attaining 97% when the manual platform was used. 48% training and 90% test accuracy. On both the micro and the macro defined average, separately, both platforms demonstrated average AUCs of more than 95% and AUCs of each classification of 90% and more. The automatically generated platform proved faster in generating results, while the manually initiated platform yielded more accurate results. Among all the models the SVM models showed the highest accuracy both on the web and programming platform. Mentioned in the study interacts with AI as a tool that has the possibility of correctly diagnosing cataracts and providing timely and effective patient care. [40]

Furthermore the assessment determined the effectiveness of the Logy AI cataract screening solution, which gauges the presence of the disease from smartphone photos and compared the results to when the specialists used slit-lamps. The study by Pareek et al. (2024) included 437 patients from medical care and they analysed 794 eye images. As seen in In the previous section, the overall efficiency of the AI solution is 90.08%, with 88. Thereof, a PNN achieved 98,02% accuracy for immature cataracts, 97.71% for mixed cataracts, 96.32% for nuclear cataracts. 16% for mature cataracts, including 72% for patients with no prior surgery on the affected eye and 90% for patients with prior surgery on the affected eye. Sensitivity was 90.38%, specificity was 89. It showed the precision of 80% and recall of 87%, in addition to the F1 score of 87%. Accuracy: sensitivity, 98%; specificity, 94%; The positive predictive value was 85%. Sensitivity to was 71% and specificity was 57%, negative predictive value was 93%. The AUC of the AI module was 0.29%. Based on the findings, the authors established that the Logy AI cataract screening module is suitable for further community level tests especially in the scattered rural centres and can also be strong candidates for home use after the COVID-19 pandemic. [52]

With the use of the K-means algorithm and a radial basis function neural network, Acharya et al. (2010) conducted a superb study into eye diseases, including normal and abnormal optical eye pictures, and achieved a predicted accuracy of over 90%. A different research was covered that dealt with the use of a neural network based classifier to distinguish between normal, cataract, and post-operative cataract eye

pictures. The results indicated an overall accuracy of 90%, sensitivity of 97.7%, and specificity of 100%. [4]

Numerous studies have explored the application of AI in different stages of cataract surgery, showcasing promising results in IOL power calculation accuracy and complication risk prediction using models like neural networks (NNs), support vector machine regression (SVM-RM), and ensemble models. For example, as per Tognetto et al. (2022), the NNs trained on one hand had correlation coefficients of 0.72 to 0.83 in different datasets. Likewise, in terms of IOL power calculations, the results of the study demonstrated significantly enhanced accuracy for both the SVM-RM and the MLNN-EM models in comparison to the existing formulas, with more eyes located within acceptable prediction error margins. However, there are several issues; firstly, proposed models still require validations in actual clinical environment contexts and secondly, models need to be more interpretable. [37]

Khairudin et al. (2023) as far as the authors' knowledge has done a pioneering study that focuses on the pattern analysis of diabetic patient's cataract photos by using CNNs owing to smartphone cameras. It is applied with the purpose of addressing the increasing global threat of diabetes – an illness that is linked with cataract formation, thus causing blindness. There was a high level of reliability in the model as perfected by the scores in accuracy, recall, and F1-score measures being above 90%, thereby indicating that the model will be useful in the diagnosis of cataracts with regard to the length and severity of diabetes. Furthermore, the research saw a correlation between high blood sugar levels and other problems which cause the cataract to progress at a faster pace; maybe artificial intelligence may come in handy in the early diagnosis of diabetes complications. However, the researchers claim that in an effort to improve the ratings of the diagnostic abilities, the dataset must be expanded and altered in the AI model. [49]

Li et al. (2020) and Wu et al. (2019) used ResNet-CNN in slit-lamp images and indicated that the diagnostic accuracy rate can amount to 99.8% and sensitivity up to 99.5%. Yet, Gutierrez et. al (2022) found that contemporary formulas such as the Kane formula, which was derived from theoretical optics with regression analysis, has yielded better results than the traditional, giving a much better result for the refractive error. Moreover, in this paper the authors conceptualized a model that involved AI algorithms that incorporate deep learning algorithms especially target in screening or diagnosing cataracts through slit-lamp and the fundus photos that can easily allow telemedicine for remote diagnosis. To perform the calculations, the authors utilised a dataset retrieved from the WHO and Hill-radial basis function (RBF), an IOL calculator that employs regression analysis involving a large refractive outcome set coming from an ANN. In preoperative assessments, applications related to span screening, diagnosis, intraocular lens (IOL) power calculation, intraoperative management and postoperative care all provide one-stop solutions for better patient management. However, there are the potential issues that would work against further progression: data security issues, the clinicians' reluctance to accept EMRs, and the potential barriers in implementing EMRs en masse. These are some barriers which must be overcome especially in developing countries where resources are limited yet every human being deserves to have an equal chance of benefiting

from advanced technology such as artificial intelligence in managing cataracts and overall eye health problems.[27][22][33]

In another creative approach Shimizu et al. (2023), the author of the implementation study decided to take a closer look at cataract detection, through the lens of a combination of conventional slit-lamp practices and state-of-the-art machine learning. This was to determine whether a model based on AI could correctly detect nuclear cataracts from videos acquired over these examinations. They did it in a pretty neat way. To record videos of the eyes during slit-lamp exams, they used a portable Smart Eye Camera and then turned the machine learning algorithms loose on the footage. With an incredible 1,812 Asian eyes in their dataset,they had a wealth of data to sift through and ended up studying more than 38,000 frames to identify which were diagnosable. And boy, did their model deliver! The accuracy was off the charts, hitting a whopping 94.2%. That's pretty impressive, especially when you consider how closely it matched the performance of skilled ophthalmologists. It's like having a digital assistant right there in the clinic, helping to identify cataracts with incredible precision. But what's even more exciting is how this study underscores the power of simplicity and diversity in data. By keeping their diagnostic criteria straightforward and their dataset diverse, Shimizu and the team showed just how effective AI can be in diagnosing cataracts across different ethnicities and types. It's a big step forward in the quest to improve cataract diagnosis and ultimately, treatment outcomes for patients worldwide. [44]

Another notable contribution in this domain is the study by Kumar et al. is also a significant contribution in this direction. (2023) entitled "Cataract Disease Identification Using Transformer and Convolution Neural Network: A Novel Framework" is a paper proposing a hybrid architecture which uses the Convolutional neural networks (CNNs) with the Vision Transformers (ViTs) for improving the performance of cataract detection on ocular images [43]. The study introduces a two-stage framework that is formed by pre-trained CNNs (ResNet or VGG) for feature extraction and a Transformer module for enhanced representation learning. This is a hybrid model that seeks to leverage CNNs' strength in extracting local features and ViTs' efficiency in capturing global contexts. Our model's capability to jointly capture local texture and global long range dependencies results in a substantial performance gain over classical CNN-only techniques. In addition to architectural contributions, the authors emphasize the significance of dataset curation and preprocessing. The authors utilize multiple data augmentation techniques to mitigate class imbalance and improve model's generalization. The work achieves more than 98% accuracy on their data, and demonstrates the model's potential capacity and practicality of integrating CNNs and Transformer modules in the clinical practice.

Lastly, Lu et al. (2025) investigated a full deep learning based framework for cataract management, which has been considered as a major advancement beyond conventional diagnostic AI by a large margin.[53] Instead of only detecting the existence of cataracts, they designed a multi-dimensional model able to exactly detect diseases can predict the severity of cataracts, forecast surgical complications, or even provide recommendations on optimal intraocular lens (IOL) powers. They used CNN models pre-trained on large and diverse clinical data; slit-lamp and fundus images, and

surgical video sequences. Notably, their model was reaching diagnostic accuracies as high as 97%, nearly on par with expert ophthalmologists. Their deep learning model achieved an accuracy of 92% for surgical phase recognition, which provides significant reliability for use in real-time clinical assistance. Predictive analytics, including IOL power estimation achieved greater than 90% accuracy demonstrating a high potential for personalized surgical planning. In doing so, authors of these approaches increased both transparency and model trust and acceptance, by emphasizing the importance of model interpretability approaches like Grad-CAM. This confluence of integration and prediction is a significant move forward toward a future in which AI is a true partner in the clinic, enriching patient care and the efficiency of cataract service provision.

Chapter 3

Data Acquisition and Preparation

3.1 Preliminary Dataset Review

As part of our exploration of cataract datasets, we looked at several publicly available datasets containing images of cataract to both understand what type of data was out there for this task and to assess the quality and potential utility of this data to train the model. The datasets we primarily collected alongside our real-world collections were ITEC's Cataract-101 (ITECCataract101), Cataract-21 (ITECCataract21), CatInstSeg (CatInstSeg), Cataract-1K (ITECCataract1000) and Roboflow Dataset: Optiscan (Manganti, 2024). [50] In this thesis, however, we have concentrated on a custom-collected data, which was more tailored towards what we required with respect to quality, labelling and diversity. Still, the listed datasets were crucial for our understanding of the problem domain and have guided our design decisions during the whole spanning project.

Among the datasets analyzed, Primus et al (2018) introduced a deep CNN for automatic phase recognition in cataract surgery videos with the **Cataract-21** dataset, reporting a mean F1-score of 68% and a maximum of 75% even with temporal information. [18]. Schoeffmann et al. (2018) introduced the **Cataract-101** which is an annotated surgery video dataset for visual analytics and surgical quality assessment. [20] Fox et al. (2020) implemented Mask R-CNN for tool detection on the generated **CatInstSeg** dataset.[25] Ghamsarian et al. (2024) introduced the **Cataract-1K** database, which includes annotated surgical videos for various computer vision tasks in ophthalmic surgery. [48] These datasets were not used in our final training pipeline, they remain significant references in the domain of cataract classification.

3.2 Proposed Dataset Description

We have used the custom dataset for training, validation and testing purposes and our own dataset was collected from two of the well-reputed ophthalmology centers of Dhaka city namely Harun Eye Hospital and Vision Eye Hospital. It has been prepared as a multi-class medical image dataset and annotated under the clinical supervision of expert ophthalmologist, **Prof. M.A. Bashar Sheikh**. The main purpose of this database is to differentiate various types of cataract and normal with the help of anterior segment images. These pictures in fact were taken under real clinical conditions and provide those essential diagnostic features in terms of lens opacity, clarity and morphological abnormality related to cataract formation. There are 5 classes in the dataset, corresponding to genuine ocular states: Cortical Cataract, Hypermature Cataract, Nuclear Cataract, Posterior Cataract and Normal. Every image shows a different lighting, focus or presentation, which gives a realistic image of the visual variability in clinical practice. The data were processed after acquisition, cleaned, labelled and organised in a systematic way for deep-learning techniques. Overall dataset is composed of 692 high resolution eye images that were divided into 3 separate sets in a 62-25-11 split (training, validation, testing – in order to conduct a representative division the stratified sampling method is used) as such:

- **Training Set:** 432 images
- **Validation Set:** 171 images
- **Testing Set:** 89 images

This partitioning strategy has been followed in many standard deep learning workflows to provide adequate data for training a model along with independent sets for unbiased validation and testing. The second choice assures all subsets have a similar portion of all classes, and hence prevents the integrity and the proportionality required for robust multi-class classification. The distribution of samples for each class for each set is given below:

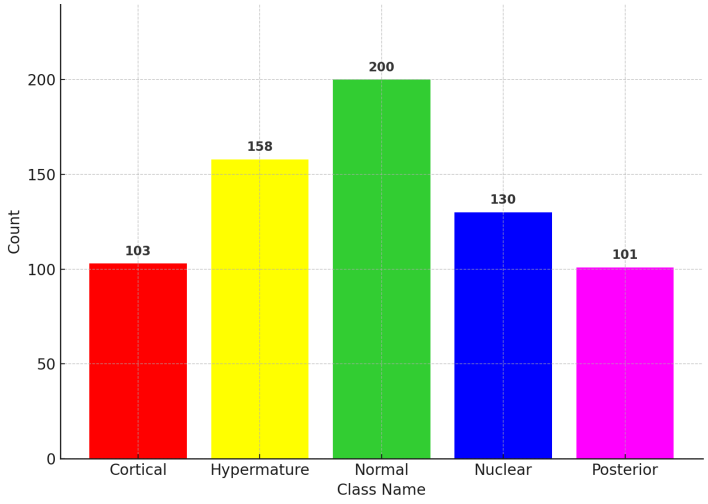

Figure 3.1: Distribution of Samples in Each Class

These numbers show a reasonably proportional distribution of classes between all subgroups and even though there are some differences in the amount of available samples from a class per group. For medical classification, not even mild imbalance can degrade model quality, hence this distribution was designed such that each class still is decently represented. Such a constitution provides meaningful learning and a reliable evaluation of the classification model under the different clinical situations.

3.3 Data Preprocessing Summary

Before training the deep learning models, some important preprocessing steps were performed on the eye image dataset to standardize the data, to aid in feature extraction and enhance the generalization ability of the model in unseen samples. As the dataset consists of anterior segment eye images, which are classified into one of the five categories of Cortical, Hypermature, Normal, Nuclear, and Posterior, careful preprocessing is required due to variability in acquisition procedures, lighting conditions, and demographics of patients.

Firstly, images were rescaled to a fixed size of 224×224. This was required for gratifying the input layer needs of convolutional networks (CNNs) including VGG19, Xception, and ResNet101, that the present study employed. By resizing, the data was normalized and computational load reduced while maintaining important visual details such as lens clarity and presence of opacities. After resizing, all images were labelled-using their visual diagnosis-to their corresponding classes. (Note that these labels were in text (e.g., "Normal", "Nuclear") but were recorded into numeric ones to be incorporated into deep learning classification models.) This mapping is indispensable for performing multi-class categorical classification with softmax activation in the output layer.

All images were subsequently normalized at the pixel-level, by rescaling the pixel intensity values from [0, 255] to [0.0, 1.0]. For pre-trained models, architectural-specific normalization functions were also used to further guide the input distribution to match the statistical properties of the ImageNet dataset used to train the weights of the model. This step is crucial for achieving stable gradients and meaningful feature transfer during training. Image orientation was also considered, and corrected as necessary using EXIF metadata from the original image files. Stability of the orientations is crucial for learning and orientation inconsistencies would be more critical in training especially in medical imaging where the anatomical alignment should be preserved. Automatic normalisation was used to ensure that all eye images were displayed in a uniform and anatomically correct manner.

For enhancing subtle intraocular structures like pupillary details, it is also added to the above mentioned preprocessing pipeline to process both overexposed and underexposed images by contrast stretching and brightness normalization. Histogram Stretching was used for further enhancing contrast, in particular near places of little opacification difference. Randomized brightness changes were also added to simulate differences in clinical light sources (e.g., from different examination rooms, or between smartphone and slit-lamp cameras). The brightness variants had the purpose to make sure that the models were able to detect relevant structures even

at low or high exposure, resulting in a higher robustness towards realistic conditions.

Overall, these preprocessing methods resizing, label encoding, normalization, orientation correction, contrast enhancement, and brightness modulation played an important role in data preparation for deep learning models in the performance. They standardized all images that were presented to the models to be consistent, diagnostically relevant, and robust against noise and variability that is typical of the real-world medical imaging, respectively.

3.3.1 Data Augmentation Techniques

An important part of the preprocessing pipeline included training image augmentation. These techniques include horizontal flipping, brightness (exposure) adjustment, Gaussian blurring and salt and pepper noise. All of these were done in **Roboflow** workspace where we manually labeled our data into classes. Below the augmentation steps are described:

1. Random Horizontal Flip
Simulates random flips on the input image. For each image the pixel values are flipped along the vertical axis using a certain possibility:

$$x_{i,j} \rightarrow x_{i,W-j-1}$$

where W is the image width.

2. Gaussian Blurring
Adds mild smoothening to simulate defocused images or sensor blur. This operation allows the model to generalize better, reducing fine-scaled image details. A Gaussian kernel is convolved with the image:

$$G(x,y) = \frac{1}{2\pi\sigma^2} \cdot e^{-\frac{x^2+y^2}{2\sigma^2}}$$

where σ is the standard deviation controlling the blur strength or the degree of smoothing.[2]

3. Salt and Pepper Noise
Randomly replaces pixels with white (salt) and black (pepper) values to mimic sensor noise or data corruption. This makes the model learn more about larger patterns rather than single pixel values.

$$x_{i,j} = \begin{cases} 0 & \text{p/2} \quad \text{(pepper)} \\ 255 & \text{p/2} \quad \text{(salt)} \\ x_{i,j} & \text{otherwise} \end{cases}$$

p is the total noise density or probability of corruption.[2] $x_{i,j}$ is the original pixel value at position (i,j) while 0 and 255 values indicates pepper and salt noise respectively.

4. Random Brightness Adjustment
Alters image brightness in order to mimic different illumination. This encourages the model to be invariant to varying lighting conditions by brightening or dimming pixel intensity. A uniform random value is added to each pixel:

$$x' = x + \Delta b, \quad \Delta b \sim \mathcal{U}(-\beta, \beta)$$

where Δb is the sampled brightness adjustment factor.

These type of augmentations are essential in medical image analysis to represent realistic variability of imaging conditions and objects and also the model from overfitting to certain textures or light conditions. Moreover, the preparation was essential for obtaining the high classification performance and low bias, therefore the reliability of the models in real screening application. After the augmenation our test increased to 1296 images.

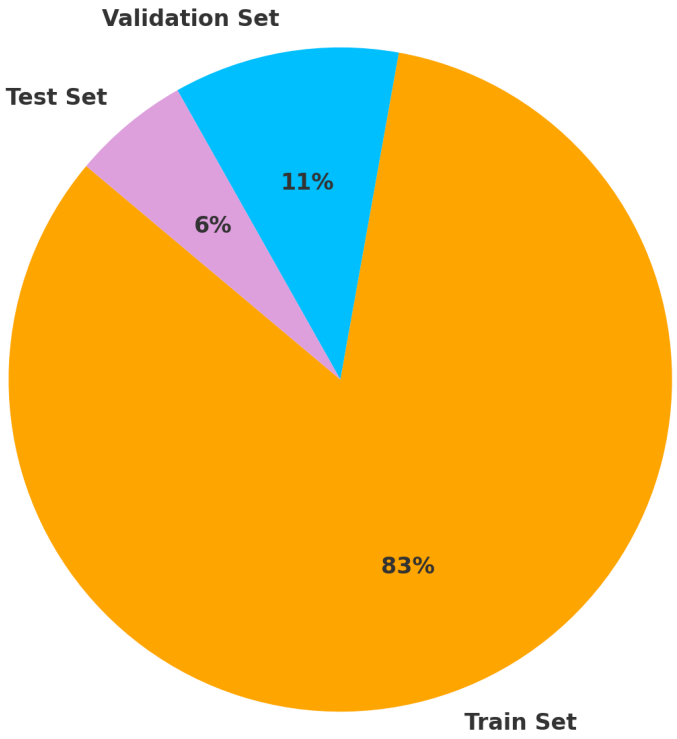

Figure 3.2: Distribution of Samples in Training, Validation and Testing Set

After that we employed Keras generators to integrate all the preprocessing steps (resizing, normalization) and augmentation techniques into single pipeline. Three **ImageDataGenerator** were used to load images from directories and then resize the images, load them into the chosen batch size, and add categorical label to them as categorical classification will help to classify multiple classes separately as well as gives us opportunity to use the custom Softmax layer that is widely used on the models we have chosen.

3.4 Dataset Limitations

Despite the rigorous clinical methodology and comprehensive preprocessing pipeline applied in this study, several limitations persist in the dataset which may influence the generalizability and scalability of the developed classification system.

First, although the dataset was custom-built under clinical supervision and includes five well-defined categories: Normal, Cortical Cataract, Hypermature Cataract, Nuclear Cataract, and Posterior Subcapsular Cataract.The total number of images remains relatively modest at 692. While stratified sampling was used to maintain proportional representation across training, validation, and test sets, the limited absolute number of samples per class (particularly Posterior Cataract and Cortical Cataract) may constrain the depth of feature representation learned by certain models. This is due to fact that in most cases eye samples was not possible collect from individuals due to hesitancy on privacy and security. Moreover, cataract cases are not listed in any of the hospitals therefore, we had to collect samples once a week when the hospitals arranged camps for free diagnose of cataract. Now, the Deep neural networks, particularly transformer-based models like ViT-B16 and FastViT, tend to benefit from larger, more varied datasets. Thus, the limited dataset size necessitated reliance on strong augmentation and transfer learning strategies, which, while effective, are not complete substitutes for broader data coverage.

Second, although the class distribution was balanced to the extent possible during sampling, minor disparities remain in the number of samples per class. In medical classification problems, even small imbalances can bias model learning toward majority classes. This is particularly important for conditions like Posterior Subcapsular Cataract, which had the fewest samples in the dataset and thus may be more prone to underrepresentation in predictions. Moreover, the dataset contains images of eyes that with has more than one class of cataracts which makes it difficult for the model to learn and give good results. While our ensemble learning method helps mitigate this issue by combining the strengths of multiple models, further data acquisition especially for underrepresented categories is essential to strengthen clinical reliability.

Lastly, while the applied preprocessing steps such as contrast stretching, brightness normalization, and salt-and-pepper noise simulation were carefully designed to improve real-world applicability, they still rely on the assumption that image artifacts introduced during augmentation adequately reflect real-world conditions. Although care was taken to match clinical variability (e.g. simulating camera noise and lighting shifts), there may still be unforeseen edge cases in live deployments that were not fully represented in the training data.

In summary, while this study introduced a clinically grounded and technically sound dataset with strong real-world potential, limitations in sample size, class distribution, geographic scope, and annotation subjectivity remain areas for future improvement. Addressing these limitations through larger, more diverse datasets and multi-expert validation would further strengthen the reliability and applicability of AI-based cataract screening systems in practice.

Chapter 4

Methodology

4.1 Model Description

4.1.1 VGG19

VGG19 is a 19 layer (including an input layer and an output layer) CNN developed by the Visual Geometry Group at University of Oxford, first introduced as an input to the competition: "Very Deep Convolutional Networks for Large-Scale Image Recognition" in 2014 by Simonyan and Zisserman.[11] The architecture is notable for its simplicity and uniform structure: there are 19 weight layers in total, and all of them are of learnable weights armed with 3×33×3 convolutions, and fix-sized max pooling. The power of VGG19 is its depth and hierarchical feature extraction provided by a uniform, deep convolutional pipeline. Even though VGG19 was originally designed for image classifications in the ImageNet dataset, it has been widely used in the field of medical image classification owing to its advantage in deep feature learning. For cataract classification, the sequential structure enables it the capture subtle changes of lens opacity and differentiate normal eyes from affected ones. This is attributed to the model's good generalization from pretrained weights and its amenability to transfer learning workflows that lends it well to the task of fine-grained medical diagnostics such as retinal, chest X-ray and cataract image analysis.

4.1.2 ResNet101

ResNet101 belongs to the Residual Network (ResNet) family that was proposed by Kaiming He et al. in 2015 in the landmark paper "Deep Residual Learning for Image Recognition."[13] The ResNet101 architecture is a 101 layer network, utilizing identity shortcut connections that hop over one or more layers, mitigating the vanishing gradient problem and making it possible to train very deep networks. ResNet101 has been extensively used for other classification tasks, such as tumour grading, chest X-ray classification, and ophthalmic disease detection. Its hierarchical architecture grants it the capacity to capture complex patterns and subtle differences, which is well-suited for cataract classification given that the information of a disease presence resides in the texture of the image. The model has already proven to produce state-of-the-art results across image datasets at scale and thus has become a lynchpin in deep learning pipelines, employed in advanced medical diagnosis amongst others, a

characteristic which aligns well with our work for improving cataract diagnosis via fine-grained classification.

4.1.3 InceptionV3

InceptionV3 is a convolutional neural net work architecture that has been optimised for image classification by Google and pre-trained on ImageNet. Szegady et al. (2016) first described its architecture in his paper in the use of computer vision.[14] The model extends on prior versions of the Inception architecture by incorporating factorized convolutions, and auxiliary classifiers as an additional regularizer of the main network. With 48 layers, the well-known characteristic of InceptionV3 is that the inception modules can process images at different scales in parallel. InceptionV3 is generally applied to the medical image analysis including the detection of diabetic retinopathy, breast cancer and pneumonia. Its property on mining high level features on different spatial resolutions is especially applicable to cataract diagnosis, where the progressive blurring and opacity at different levels need to be detected. Its deep and parallel architecture effectively provides the capability of exhibiting fine-grained features as well as eye images for detailed classification as we have emphasized in our contribution.

4.1.4 MobileNetV2

MobileNetV2 was developed by Google researchers and was published in the paper "MobileNetV2: Inverted Residuals and Linear Bottlenecks" in 2018.[19] The architecture is a small CNN specifically designed for mobile and embedded vision workloads. It also uses depthwise separable convolutions and inverted residual blocks to compute very efficient networks and achieve state of the art accuracy. In the medical field, MobileNetV2 has been widely used in real-time diagnosis and wearable applications (e.g., local retinal scan and dermatic testing). Its low latency and low memory footprint make it appropriate for deployment in resource limited models such as handheld cataract screening devices. As for our work, MobileNetV2 provides a compact backbone for cataract classification models designed for the use on mobile or field-based setups, where speed and limited resources are crucial.

4.1.5 EfficientNetB2

EfficientNetB2 is one of the variants of EfficientNet introduced after Google AI released their paper "EfficientNet: Rethinking Model Scaling for Convolutional Neural Network" by Quoc Le and Mingxing Tan in 2019.[21] EfficientNetB2 is a compound-scaled model that mainly considers the trade-off between resolution, depth,and width size for performance under the constraint that the computational cost should be limited. These models obviously achieve a higher performance than many standard benchmarks on practice with less parameters and FLOPS. In the medical field, EfficientNetB2 has been used for cancer detection, lung disease classification, and rabies retinal abnormality detection8. Its efficiency and scaling are maximized for high resolution medical images that it can process without consuming the entire GPU memory. EfficientNetB2 can play an important role in classification of cataracts, especially when compared to other network structures that become inaccurate when

the training set is small or the data is unbalanced, which is common in specific medical datasets, such as the cataract diagnosis image.

4.1.6 DenseNet121

DenseNet-121, introduced by Huang et al. (2017), adopts dense block connectivity architecture where each layer is connected with not only its preceding but also with all subsequent layers in feed-forward manner.[17] This encourages more efficient feature reuse, improved gradient flow, and reduced parameter redundancy. DenseNet-121 is known to work well for medical application classification including cataract detection, showing excellent performance (e.g., 92% accuracy rate on fundus images). In this work, we tuned DenseNet-121 to multiclass prediction of cataract types in our in-house dataset. It reached a very good accuracy and consistency, particularly between the classification of nuclear and cortical cataract, indicating good generalization even with a limited dataset size.

4.1.7 Xception

Xception is an abbreviation for "Extreme inception." The concept was introduced by François Chollet in 2017 paper "Xception: Deep Learning with Depth Wise Separable Convolutions."[15] It was designed as an enhancement over the Inception architecture, by replacing traditional Inception modules with depthwise separable convolutions to decrease comput- ational complexity, simultaneous improving the accuracy. Xception consists of 71 layers and is constituted of only depthwise separable convolution layers that can efficiently extract the features from images. In medical image classification, Xception has been used in skin cancer, pneumonia, and diabetic retinopathy diagnosis, successfully exploiting its capability to concentrate on local discriminative areas. It's fine-grained feature extraction is quite advantageous for cataract diagnosis since minor details makes normal eyes different from different levels of lens transmitted eyes. Due to Xception being very quick and yet extreme effective, that is very helpful when we want to improve the cataract classification tasks of our thesis focus.

4.1.8 ViT-B16

ViT-B16, proposed by Dosovitskiy et al. (2021), are the transformer architecture applied to images by decomposing them into 16×16 patches and processing with multi-head self-attention layers.[28] The model is particularly good at modeling global contextual relationships, which can also contribute to the fact that they are trained on a large-scale dataset like ImageNet-21k. In this study, we fine-tuned ViT-B16 for multiclass cataract classification. It provided attention-based representations, which facilitates the recognition of fine textures especially the posterior subcapsular and hypermature cataracts, supplementing the CNN outputs, and further enhances the overall performance in the stacked ensemble.

4.1.9 MobileViTV2

MobileViT model was originally proposed by Mehta et al. (2022) in their paper "MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer".[34] and its more improved version (MobileViTV2) that we applied in our work was proposed by the same authors in the paper "Separable Self-attention for Mobile Vision Transformers" where the they stated that the improved version outperforms its predecessor by 1% while being 3.2x times faster than the previous version when running on mobile devices.[35] The model is a hybrid between a CNN and a transformer, designed to preserve global context modeling while keeping the compactness and efficiency of CNNs. Seperated from the traditional transformers, there are alternating local feature extraction (by inverted residual CNN blocks) and global representation learning (by transformer-stype MobileViT blocks) in MobileViT. Such a dual mechanism enables learning of fine as well as wide spatial patterns in the cellstax efficiently. MobileViT has been applied to various vision tasks, e.g., image classification, semantic segmentation and object detection, with special focus on mobile scenarios where computation and power constraints are precious. Small parameter size and low latency make it suitable for edge AI deployment and applications: healthcare diagnostics. In this study, we integrated MobileViTV2 for the potential of a good competitive task accuracy with a reduced computational complexity than its previous version in image classification problem. Alongside the already constructed transformer-attention-based design and higher parameter count improves spatial perception — an indispensable quality to interpret ophthalmic images with cataract of different severity levels. The efficiency and interpretability of the model motivate its application in portable eye-care devices for early cataract detection.

4.1.10 FastViT

FastViT is a hybrid vision transformer model introduced by Anasosalu Vasu et al. in 2023 with a paper titled "FastViT: A Fast Hybrid Vision Transformer by Structural Reparameterization".[39] FastViT, developed by Apple researchers, was created to close the gap between performance and latency in vision models, especially as it applies to real-time applications on mobile and edge devices. The system includes architectural novelty that leverages convolutional and transformer-based ideas and includes a new module called RepMixer, that provides structural reparameterization to remove skip connections and improve inference efficiency. It is applied to replace early-stage self-attention with large kernel depthwise convolution which can capture spatial hierarchies with less computation. FastViT showed outstanding performance in standard vision benchmarks (e.g., ImageNet classification and object detection), with much lower mobile inference latency than that of prevalent architectures, e.g., EfficientNet and ConvNeXt. Its resilience to image corruptions is and out-of-data samples has made it appealing for safety-critical and mobile deployment applications. FastViT was used in our research for its fast inference speed, low memory cost and powerful classification ability, which is an appropriate candidate for real-time cataract screening in mobile diagnosis sites, especially those that computational resources are limited.

4.2 Model Implementation

4.2.1 Custom Layers

Here are the custom layers and parameters that we used in the models during the training phase:

1. GlobalAveragePooling2D
This layer averages all the spatial values of each feature map to scalar. It makes the network parameter-efficient by mitigating overfitting while retaining spatial information in a compact form.[7]

$$y_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_{i,j,c}$$

where:

- $x_{i,j,c}$ is the value at spatial location (i, j) in channel c for the feature map of the input
- H and W are the height and width of the feature map respectively
- y_c is the averaged scalar output.

2. Dense (Fully Connected Layer)
Performs a linear combination of the input and applies a non-linearity e.g. activation function such as ReLU or Softmax. It aids in mapping of learned features to class probabilities or representations.[12]

$$y = \phi(Wx + b)$$

where x is the input vector, W is the weight matrix, b is the bias vector and ϕ is typically ReLU or Softmax in this context.

3. BatchNormalization
Applies a normalization technique which maintains the mean outputs that are close to 0 and the output standard deviation that are close to 1.0. It normalizes and standardizes the activations to zero mean and unit variance over a batch.[9]

$$\hat{x} = \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}}, \quad y = \gamma \hat{x} + \beta$$

- x - input activation
- μ - batch mean
- σ^2 - batch variance
- ϵ - constant (for numerical stability)
- $\hat{x}$ - normalized activation
- γ and β - learnable scale and shift parameters

- y - the final output after batch normalization.

4. Dropout
A regularization method which randomly sets a fraction of the neuron inputs of the model to zero during training.[6] This also prevents network from relying too much on certain neurons, which would result in the network being overfit:

$$y_i = \begin{cases} 0 & \text{with probability } p \\ \frac{x_i}{1-p} & \text{otherwise} \end{cases}$$

x_i is divided by $(1-p)$ to maintain expected value.[8]

6. Conv2D
Applies convolutional filters to either image or feature maps. It allows to learn spatial hierarchy and convolves different distinguishing features to detect edges, textures and patterns.

$$y_{i,j,k} = \sum_{m=1}^{M} \sum_{n=1}^{N} \sum_{c=1}^{C} x_{i+m,j+n,c} \cdot w_{m,n,c,k} + b_k$$

7. Activation Function(ReLU)
Activation Function is a transformation that make the network capable of representing non-linear relationships. ReLU is most commonly used function, because it sets all negative inputs to zero while keep positive values unchanged.[5]

$$\text{ReLU}(x) = \max(0, x)$$

5. Softmax (Output Layer)
This probability function is applied to the output layer for multiclass generalization of the logistic sigmoid function.[3] It forces the model to produce a normalised probability or categorical distribution over all classes.

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

4.2.2 Training Parameters

Batch Size: 32
This is the number of samples that are going to be propagated through the network. A modest batch size will prevent excessive memory usage and excessive learning variance.

Epochs: 40
It refers to one pass forward and backward through the entire training set. More epochs the model gets to get better, up till a point of convergence or overfit.

Optimizer: Adam
Optimization algorithm is a method to find and adjust parameters such as weights to decrease loss function when training a neural network.[12] Adam optimizer is one of the renowned optimization algorithm which is designed specifically for training deep neural networks, requiring less memory.[10] It also offers rapid convergence and is fairly stable under various deep learning problems:

$$\theta_t = \theta_{t-1} - \alpha \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}$$

- $\hat{m}_t$ = bias-corrected first moment estimate (mean)
- $\hat{v}_t$ = bias-corrected second moment estimate (variance)
- α = learning rate
- ϵ = small constant to avoid division by zero

Learning Rate: 1×10^{-4}
This parameter determines how much the model weights are to be adjusted with the training process. A smaller learning rate provides a smoother convergence, but it may take longer to converge.

Loss Function: Categorical Crossentropy
It quantifies how much the predicted probability distribution diverges from the true distribution in a multi-class classification problem.[42] This helps the model concentrate its mass on the correct class:

$$\mathcal{L} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i)$$

where y_i is the true label (one-hot encoded), and $\hat{y}_i$ is the predicted probability for class i.

4.2.3 Transfer Learning Model Construction

Prior to the training phase, each convolutional neural network model is fine-tuned with a custom lightweight classification head. Before adding the custom head, firstly, the backbone is integrated by calling `(include_top=False, weights='imagenet', input_shape=(224,224,3))`, which ensures the model doesn't include top layer for fine-tuning, loads its pretrained weights from training on ImageNet dataset as well as the input shape for the model. The models weights are frozen by using (`base.trainable = False`). The backbone's output is then passed through the first custom layer by adding `GlobalAveragePooling2D` layer to pool spatial dimensions. This vector feeds a fully connected layer of 1024 units with ReLU activation and an ℓ_2 regularization term of strength 1×10^{-4}, followed by a dropout layer with rate 0.5 to reduce overfitting. A final `Dense(5, activation='softmax', kernel_regularizer=l2(1e-4))` layer produces class probabilities over the five target categories. The model is compiled with an `Adam` optimizer whose learning rate decays exponentially according to `decay rate of 0.9, decay steps of 1000` and uses categorical cross-entropy with label smoothing $\epsilon = 0.1$, and tracks accuracy. This modular setup preserves the pretrained convolutional filters while enabling rapid adaptation of the lightweight head during training.

4.2.4 Vision Transformer Model Construction

Since the ViT models doesn't follow the same pipeline as the Transfer learning models, we used a separate model constructor. Here, the `build_vit_head_model` function extends a frozen Vision Transformer backbone by appending a lightweight classification head. The MobileViTV2 and FastViT(ma36) model used for our work was implemented from the `keras_vision` library found in `veb101` [51] GitHub repository which is initially designed by Vaibhav Singh while ViT-B16 model was implemented from Vision Transformer library. Then, all pretrained weights were frozen. Only for FastViT model we integrated an intermediate feature map from the `final_mobileone` layer which is then extracted and passed through a `Conv2D` layer with 128 filters, followed by `BatchNormalization`, a ReLU activation, and `GlobalAveragePooling2D`. For other models, the backbone's output is either reduced via `GlobalAveragePooling2D` when it is four-dimensional like MobileViT model or flattened with `Flatten` for arbitrary tensor shapes like ViT-B16 model. The resulting feature vector then feeds a fully connected layer of 256 units for the all the models while for MobileViTV2's case this is followed by an additional batch normalization step. A `Dropout` layer with rate 0.5 reduces overfitting, and a final `Dense` layer produces class probabilities. Lastly, The model is compiled using the Adam optimizer, categorical cross-entropy loss, and accuracy as the evaluation metric.

4.3 Model Architecture

All of our implemented models architecture is shown below:

4.3.1 VGG19

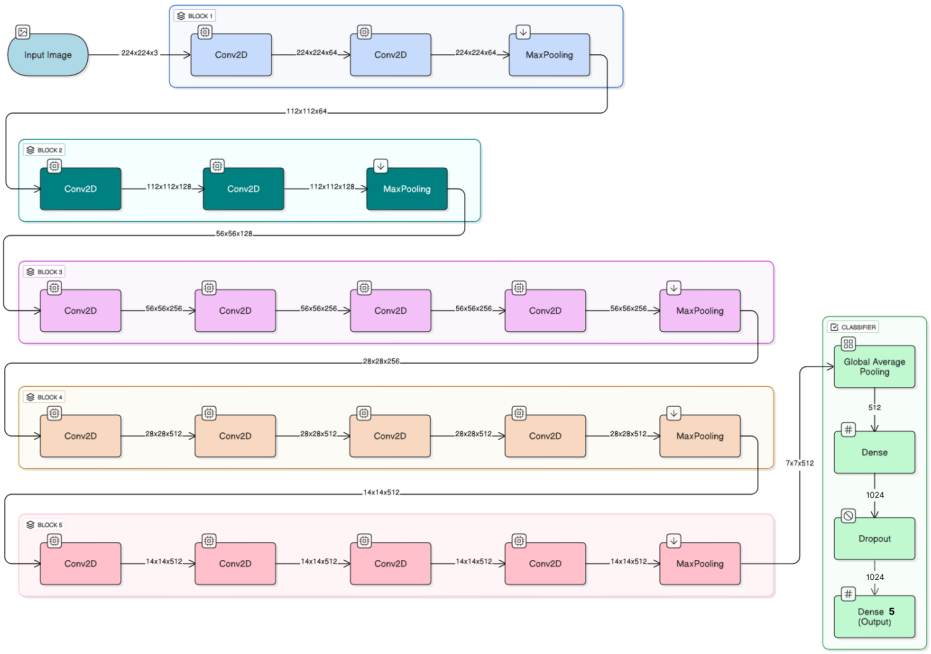

Figure 4.1: VGG19 Architecture

4.3.2 ResNet101

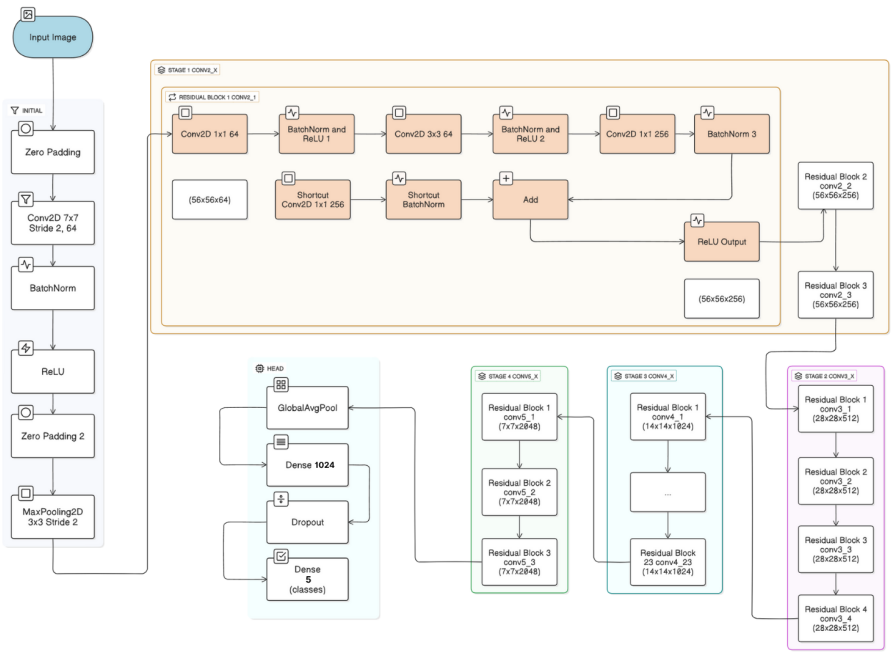

Figure 4.2: Resnet101 Architecture

4.3.3 InceptionV3

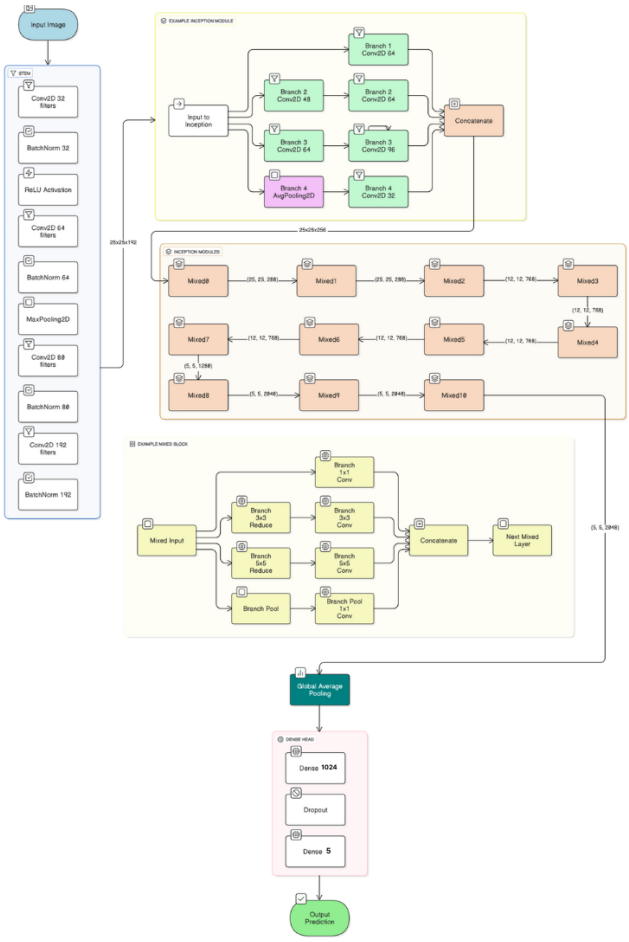

Figure 4.3: Inception V3 Architecture

4.3.4 MobileNetV2

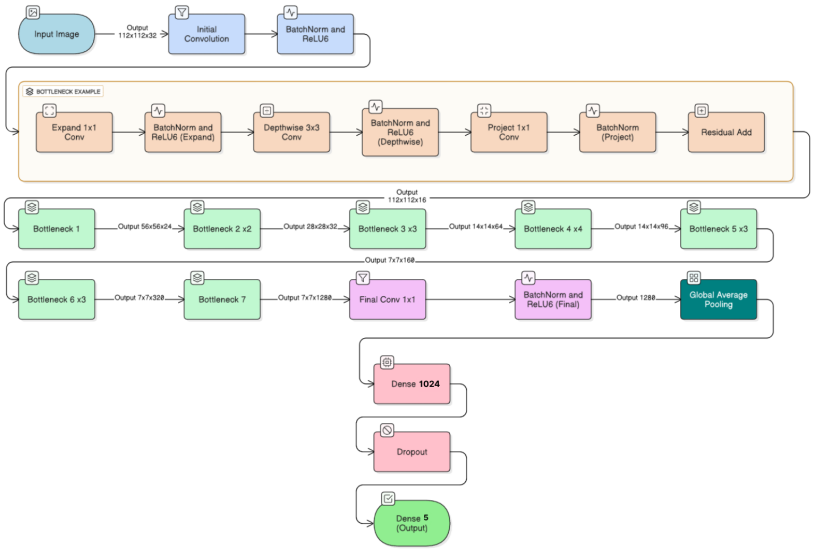

Figure 4.4: MobileNetV2 Architecture

4.3.5 EfficientNetB2

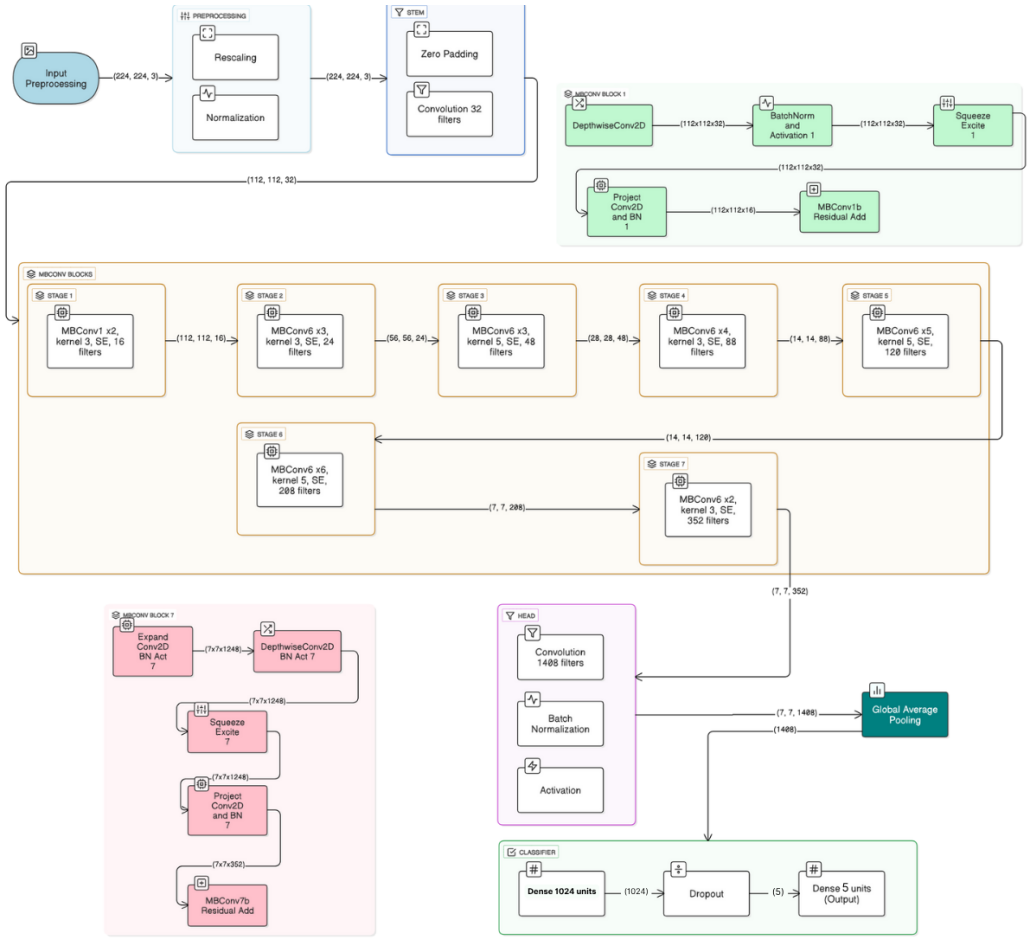

Figure 4.5: EfficientNetB2 Architecture

4.3.6 DenseNet121

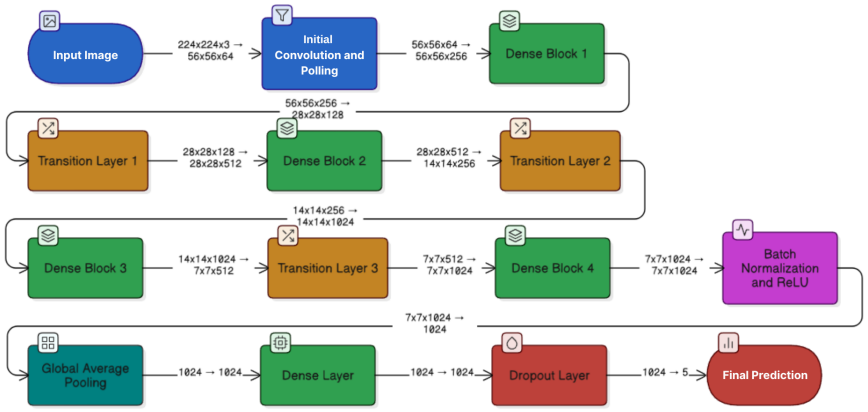

Figure 4.6: DenseNet121 Architecture

4.3.7 Xception

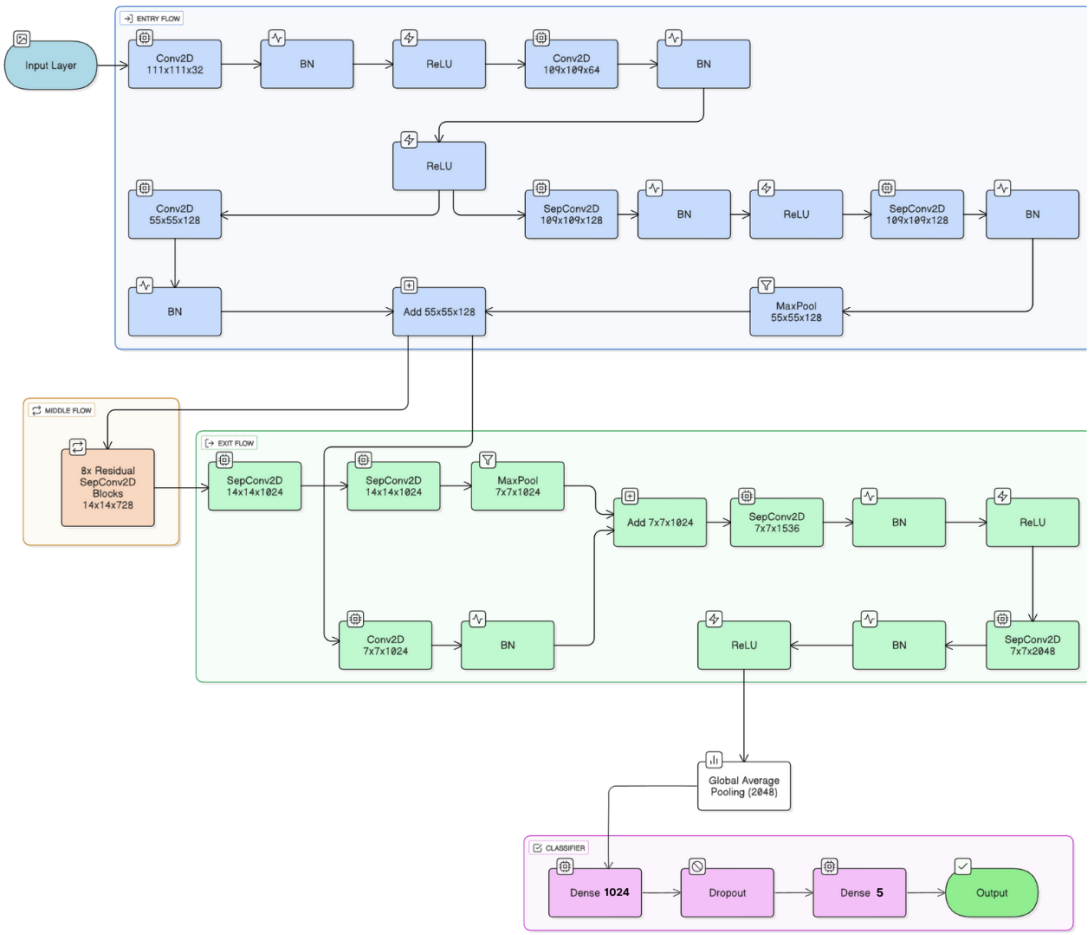

Figure 4.7: Xception Architecture

4.3.8 ViT-B16

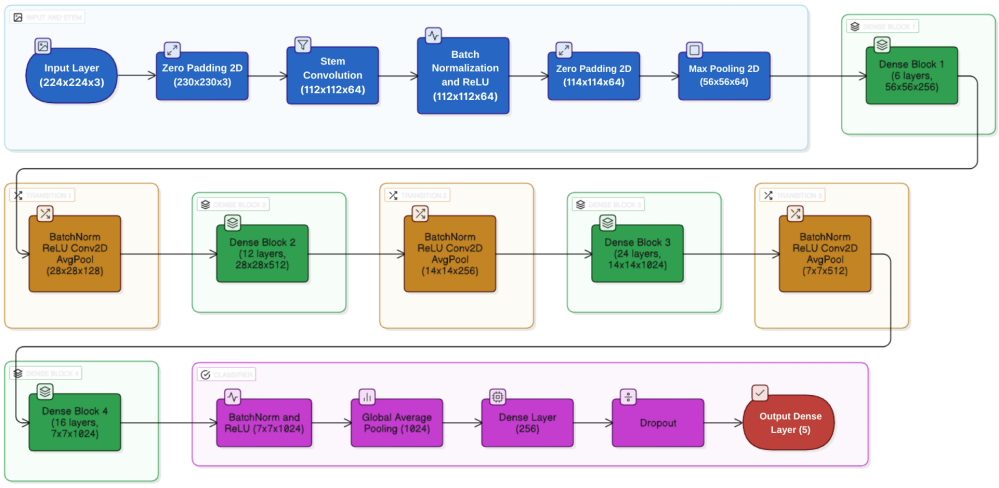

Figure 4.8: ViT-B16 Architecture

4.3.9 MobileViTV2

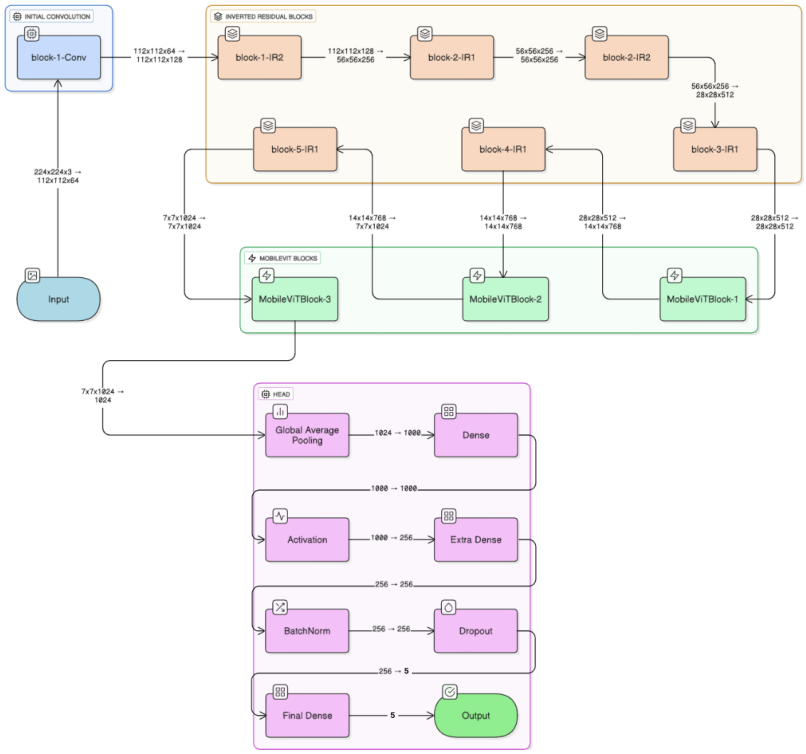

Figure 4.9: MobileViTV2 Architecture

4.3.10 FastViT

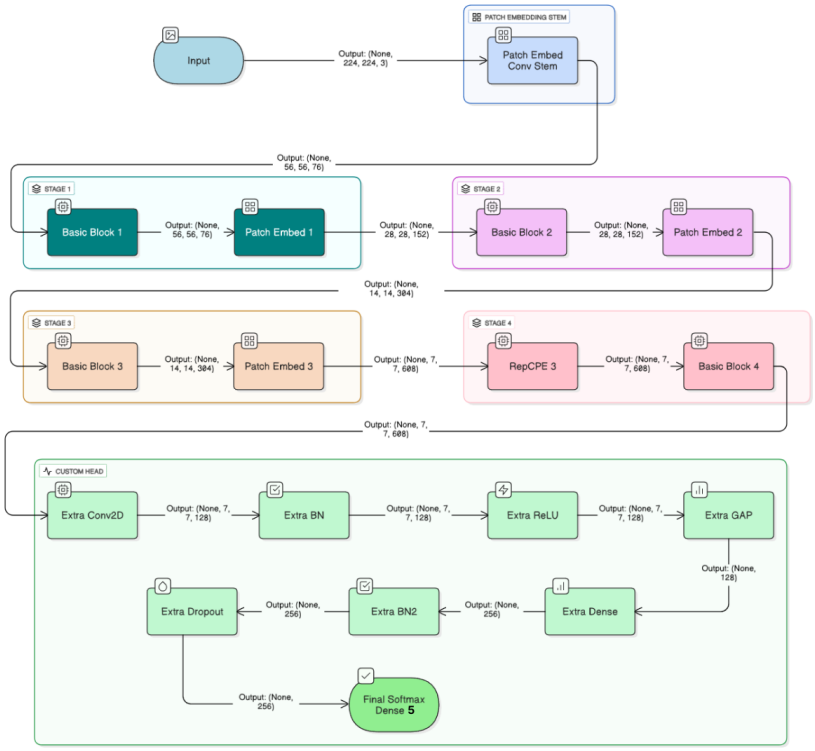

Figure 4.10: FastViT Architecture

4.4 Ensemble Learning

Ensemble learning is a machine learning technique wherein multiple individual models (learners or classifiers) are used to create a better and more accurate prediction than one would achieve using a single model alone. For our work we have used two types of Ensemble techniques:

4.4.1 Ensemble Bagging Aproach

The diverse set of good models are chosen and then their model predictions are combined to have better classification and reduce generalization error. In some embodiments, this method starts by selecting the highest validation accuracy model among two different families: CNNs and ViTs. The best CNN models and the best ViT model are selected to create a heterogeneous ensemble that capitalizes on the complementary properties of the two architectures.

For each model in the ensemble, predictions of class probabilities are made over the test dataset separately. These predictions are obtained by passing the input through model-specific preprocessing functions to be compatible with the input. An image generator feeds the test images to each model, where class probability distributions are obtained. The central idea of this bagging scheme is probability averaging: at the end, when all models have made predictions, the predicted probabilities are averaged elementwise over the ensemble. The class with the maximum mean probability of being predicted, is the final predicted class. Such averaging will help mitigate model biases and reducing the variance, which would solidify the overall classification stability.

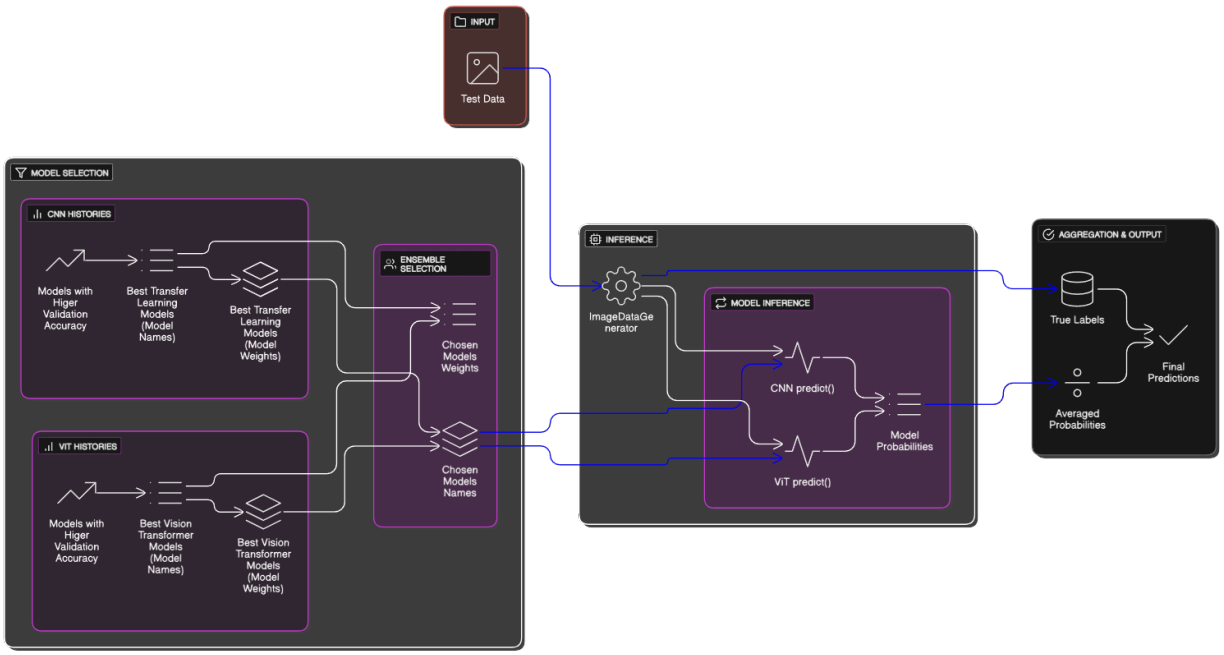

Figure 4.11: Ensemble Bagging Workflow Architecture

This prediction ensemble is then benchmarked using usual performance metrics (precision, recall, F1 score) and the production of a confusion matrix. The versatility of the models combined with their averaging in bagging becomes quite robust in terms of class imbalance and complex cataract presentations.

4.4.2 Ensemble Stacking Aproach

Ensemble stacking is a two-level machine learning procedure that integrates the predictions of several strong base models with a standalone meta-learner in order to produce better classification results.

This procedure first selects the best CNN models and the best Vision Transformer (ViT) models in terms of validation accuracy, so that a diverse and complementary set of classifiers is selected. Then, each base model is employed for making predictions on a held-out validation set. Namely, class probabilities are calculated by each model for every validation image. These probability vectors are concatenated horizontally to create a new feature matrix whereas each sample is represented by the pooled responses over all base models. This newly formed meta-features matrix is used as the input data for the second-level learner. A logistic regression classifier is used for the meta-learner. It is trained with the meta_X (stacked feature matrix), true validation labels. The logistic regression model is taught how to best weigh and combine the predictions of the base models so that, when combined, the final classification is better. This training technique constrains inter-model dependencies and forces conflicting predictions to be resolved in a wiser way than simple averaging. In testing, each base model predicts class level probabilities for the test dataset, and their outputs are concatenated to obtain the meta-level test features respectively. This input is then applied to the pre-trained logistic regression meta-learner and final class probabilities are produced. The category with the maximum value of predicted probability was further considered as the output.

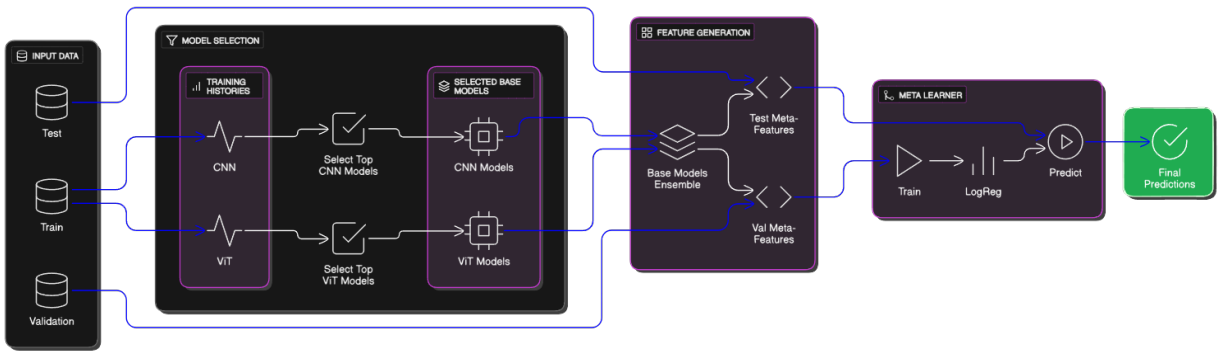

Figure 4.12: Ensemble Stacking Workflow Architecture

In this form of stacking, the strengths of the base models are fully exploited and the weaknesses are reduced as the meta-learner learns patterns of base-model's errors. Through the reconciliation of the heterogeneous decision boundaries of CNNs and ViTs, the stacked ensemble is in general more generalizable and robust, particularly in challenging multi-class classification settings, e.g., complex and imbalanced cataract subtypes detection.

Chapter 5

Results and Analysis

5.1 Accuracy and Loss Analysis

The performance of various deep learning models: VGG19, Xception, EfficientNetB2, MobileNetV2, InceptionV3, ResNet101, DenseNet121, ViT-B16, MobileViTV2, and FastViT are shown with the training curves for 40 training epochs in the following figures. The learning dynamics is observed with both accuracy and loss curves being drawn with training logs for each of the models. For every model, there are two lines on the graphs, the one for the training dataset and the one for the validation dataset, so that we can easily compare how models learn and generalize the best.

These curves are crucial for understanding the individual behavior of each model during training. In general, accuracy should go up and flatten out for both training and validation, and loss should go down and plateau over time. Concordant trend movements for the training and validation suggest such consistency is present. Departures or oscillations among the curves highlight differences in the behavior of the respective models on seen and unseen examples, especially indicating their strengths of generalization or the weakness in stability. This visualization is an indispensable diagnostic tool to assess whether the model is appropriate in the respect of medical image classification.

5.1.1 VGG19

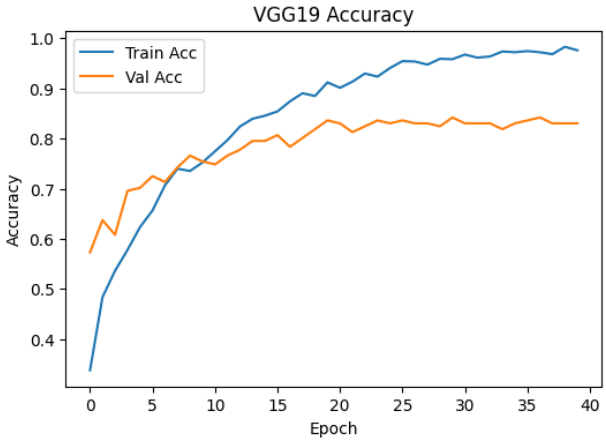

Figure 5.1: VGG19 Training Graphs(Accuracy)

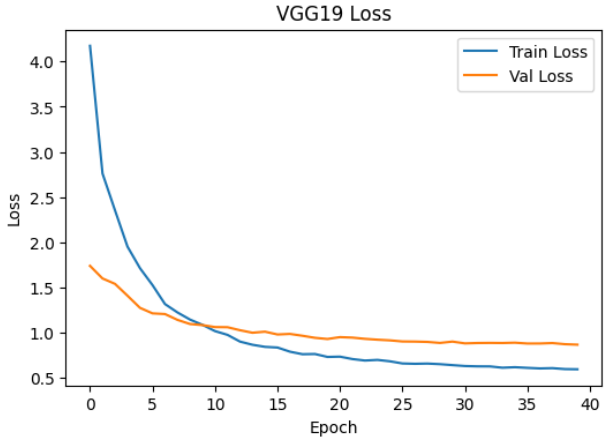

Figure 5.2: VGG19 Training Graphs(Loss)

The training accuracy of the VGG19 model had fast growth and was high (around 99%) near the end of training epochs. Validation accuracy was also good, holding steady at 83–85%. Training loss continued to drop to around 0.55. Validation loss trended lower smoothly with the training process, and settled around 0.9.

5.1.2 ResNet101

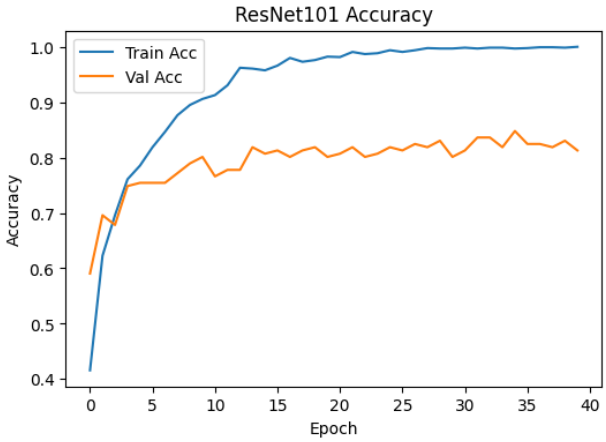

Figure 5.3: ResNet101 Training Graphs(Accuracy)

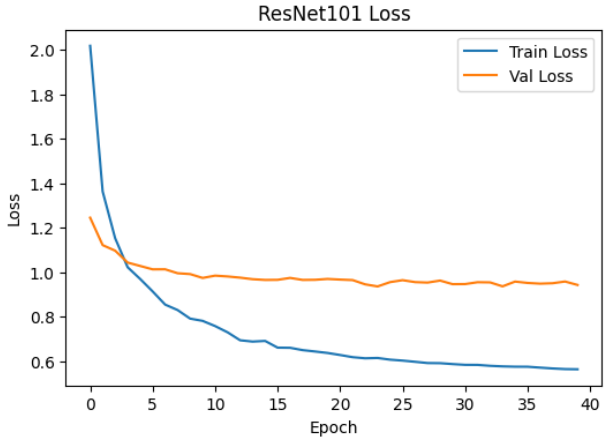

Figure 5.4: ResNet101 Training Graphs(Loss)

The ResNet101 performed well in training accuracy as it approached 100% with very few epochs. Validation accuracy remained around 83–85%, indicating a stable generalization. The training loss kept on decreasing, and went down to approximately 0.55. The validation loss stabilized around 0.95, indicating a maintained performance curve.

5.1.3 InceptionV3

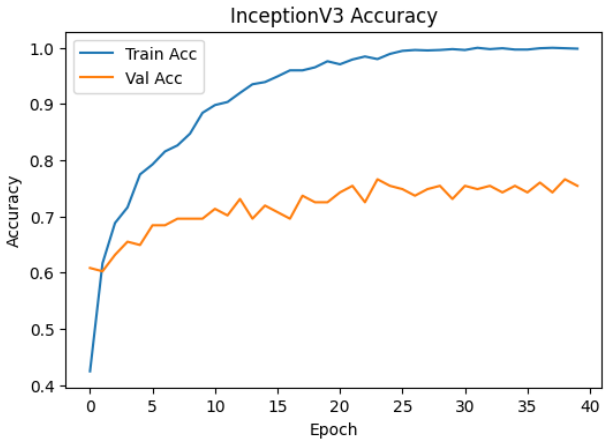

Figure 5.5: InceptionV3 Training Graphs(Accuracy)

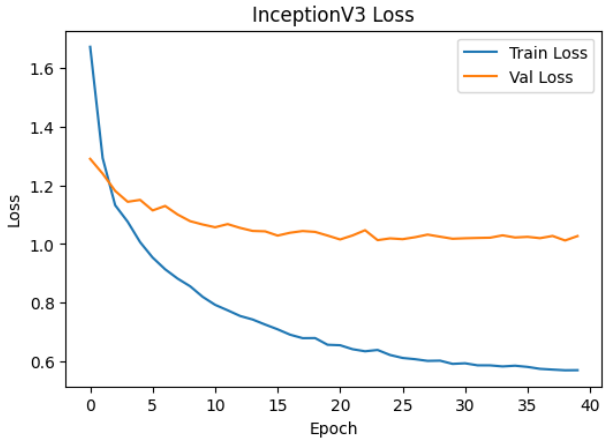

Figure 5.6: InceptionV3 Training Graphs(Loss)

InceptionV3 obtained almost perfect training accuracy by the last epoch of training. The accuracy of the validation data just hovered at around 75–77% after a few epochs. The training loss went down nicely to around 0.56. Validation loss after some initial training also stabilized around 1.03.

5.1.4 MobileNetV2

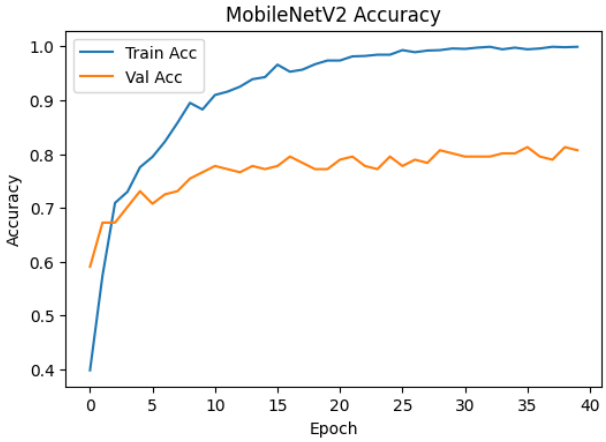

Figure 5.7: MobileNetV2 Training Graphs(Accuracy)

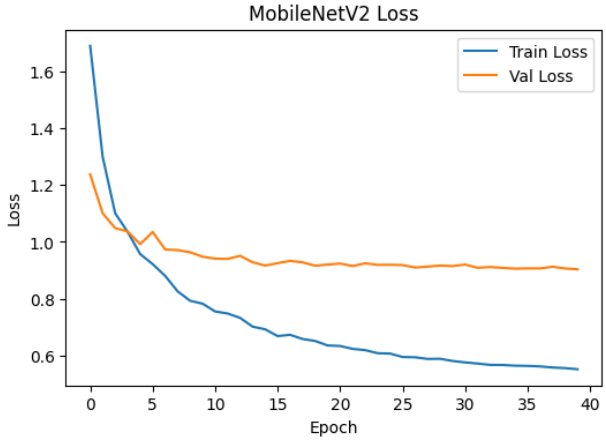

Figure 5.8: MobileNetV2 Training Graphs(Loss)

MobileNetV2 achieved 99–100% training accuracy and fast, effective convergence. The accuracy on validation remained between 80% and 82%, indicating a robust response in unseen test data. The training loss diminishes consistently to 0.55. Validation loss leveled off around 0.9 and stayed constant across the training period.

5.1.5 EfficientNetB2

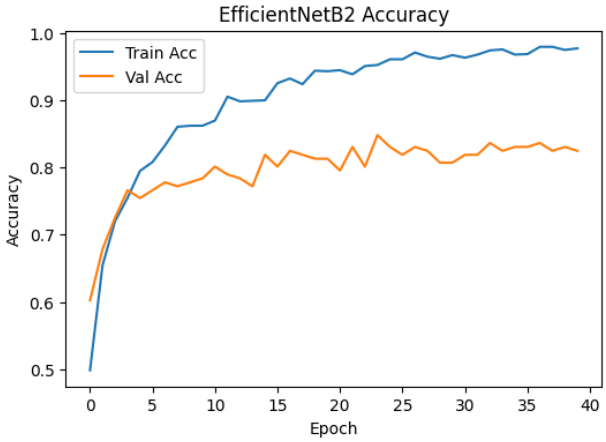

Figure 5.9: EfficientNetB2 Training Graphs(Accuracy)

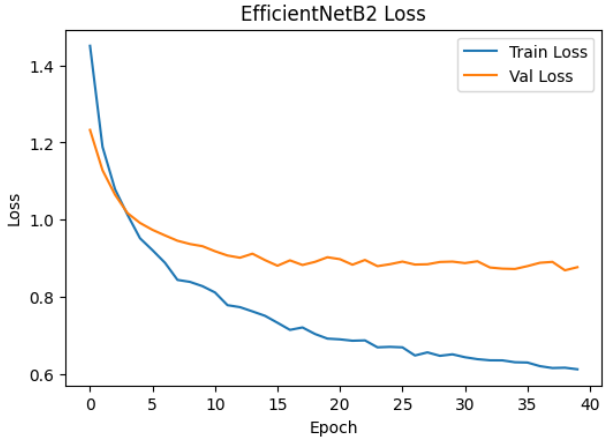

Figure 5.10: EfficientNetB2 Training Graphs(Loss)

EfficientNetB2 trained very well until around 98% accuracy. Convergence proved to be stable as the validation accuracy stayed relatively around 83–85%. The value of the training loss also decreased to about 0.6 during training. Validation loss converged around 0.87, indicating validation scores were accurate.

5.1.6 Xception

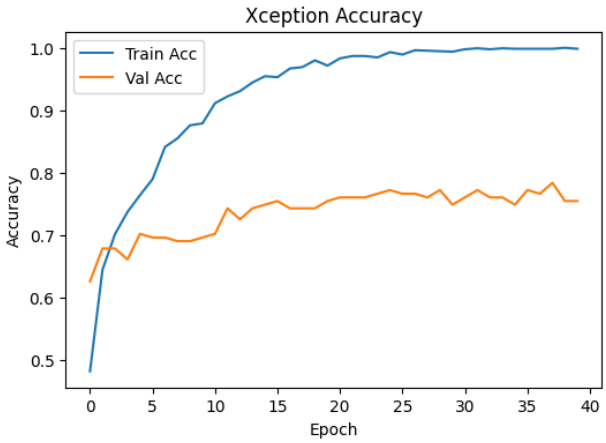

Figure 5.11: Xception Training Graphs(Accuracy)

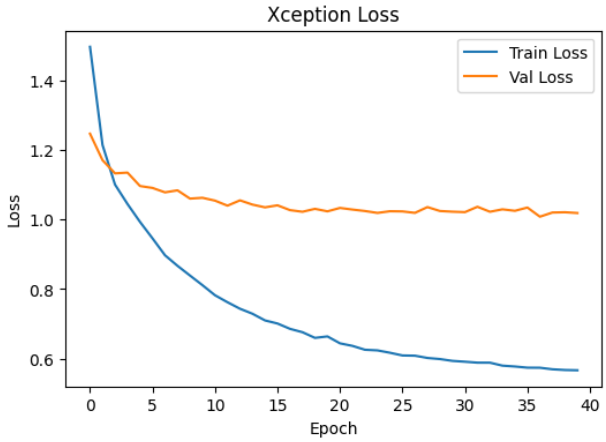

Figure 5.12: Xception Training Graphs(Loss)

Xception obtained almost perfection in terms of training accuracy, i.e., it learned the training data very well. Validation accuracy had already reach its peak afterbeing stable at 76-78%. The training loss consistently decreased to 0.56. Validation loss quickly converged and levelled off around 1.03 for the remaining epochs.

5.1.7 DenseNet121

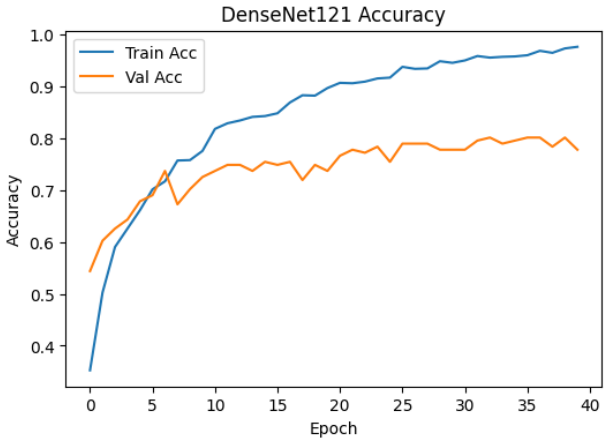

Figure 5.13: DenseNet121 Training Graphs(Accuracy)

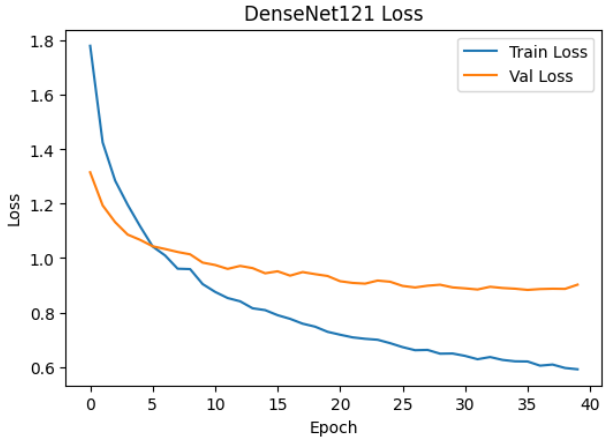

Figure 5.14: DenseNet121 Training Graphs(Loss)

The training accuracy of DenseNet121 kept increasing and finally converged to approximately 98%. The validation accuracy stayed flat between 78–80%. The training loss declined slightly to 0.6. The validation loss also had a smooth graph with the trend settling around 0.9 with no fluctuations.

5.1.8 MobileViTV2

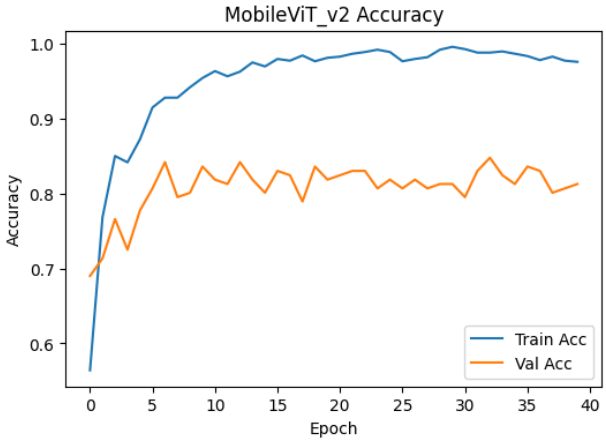

Figure 5.15: MobileViTV2 Training Graphs(Accuracy)

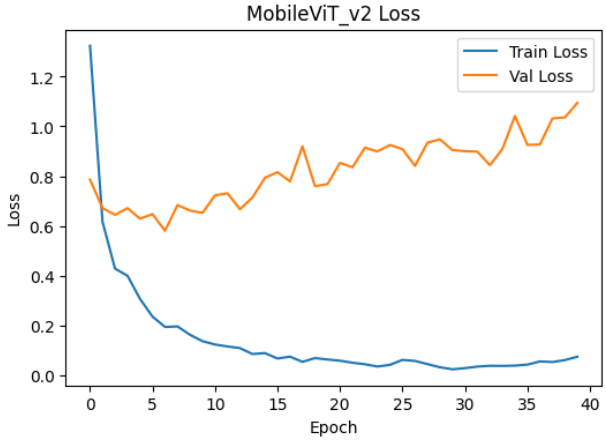

Figure 5.16: MobileViTV2 Training Graphs(Loss)

The training accuracy of MobileViTV2 improved quickly, exceeding 98% after just 15 epochs. Validation accuracy hovered around 82–85% during training. Training loss decreased sharply and approached zero. The validation loss hovered somewhat and in general ranged from 0.6 to 1.1 throughout epochs.

5.1.9 FastViT

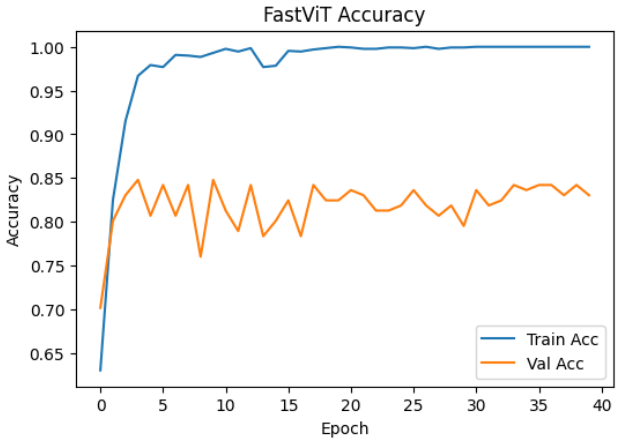

Figure 5.17: FastViT Training Graphs(Accuracy)

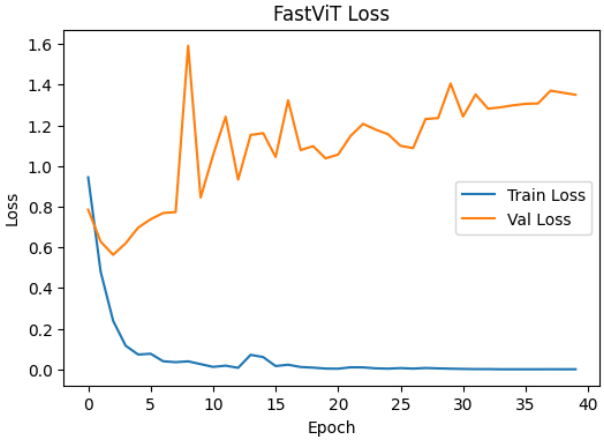

Figure 5.18: FastViT Training Graphs(Loss)

The FastViT model showed rapid increase in training accuracy and reached almost 100% starting from one of first epochs. The validation accuracy was fluctuating but on the most part, remained fairly constant around 80-85%. It is worth mentioning that training loss decreased rapidly and approached zero very quickly during training. The validation loss displayed, however, was more erratic and was from 0.8 to 1.4 across the epochs.

5.1.10 ViT-B16

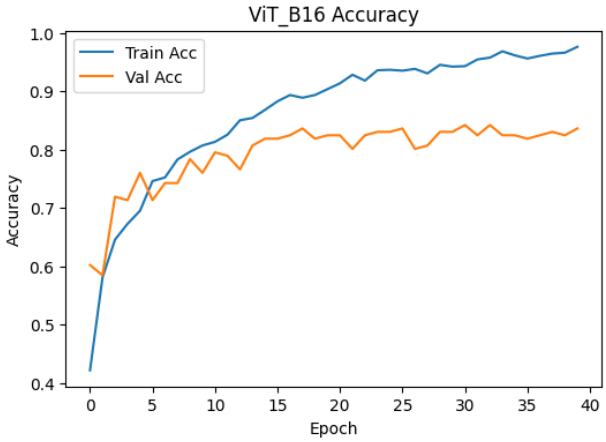

Figure 5.19: ViT-B16 Training Graphs(Accuracy)

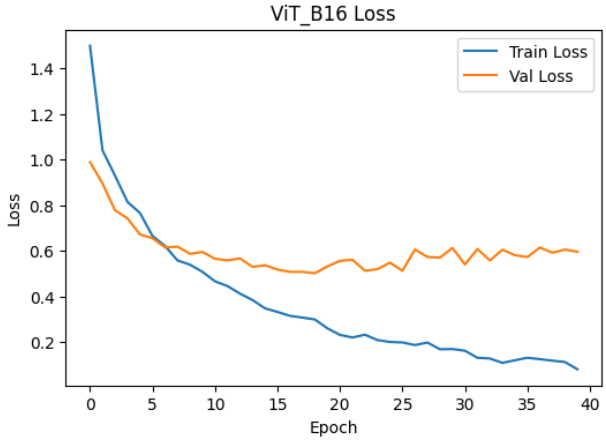

Figure 5.20: ViT-B16 Training Graphs(Loss)

The ViT-B16 model exhibited an improvement of training accuracy which monotonically increased with over 97% at the end of the experiment. Test accuracy kept increasing and fluctuating around 83–85% across the epochs. The corresponding training loss kept continuously decreasing and reached to below 0.1. The loss on the validation set decreased early and leveled off at the value of about 0.6, which was a sign of good generalization.

5.2 Performance Metrics Overview

The following chart shows our overall result for all the models and the ensemble techniques

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
VGG19	87%	87%	87%	87%	96%
EfficientNetB2	87%	86%	87%	86%	96%
InceptionV3	80%	81%	80%	80%	91%
ResNet101	87%	88%	87%	87%	96%
FastViT	88%	88%	88%	87%	97%
MobileViTV2	88%	89%	87%	88%	97%
ViT-B16	82%	83%	82%	82%	94%
Xception	83%	84%	83%	83%	95%
MobileNetV2	87%	87%	87%	87%	96%
DenseNet121	85%	85%	85%	85%	95%
Stacked Ensemble	90%	91%	90%	90%	**98**%
Bagging Ensemble	**92**%	**93**%	**92**%	**92**%	**98**%

Figure 5.21: Classification Report Results Across All Models

The confusion matrix is shown below for all the models as well as the two ensemble learning techniques on our custom test set:

5.2.1 VGG19

The VGG19 model achieved high classification accuracy rates on most of the five classes in the five-class test set. It performed best for Normal, correctly predicting 29 out of 30 samples (precision = 0.97, recall = 0.97), and consistently on Cortical Cataracts, with 12 out of 14 correct predictions. Posterior Cataracts were also well separated, being correctly identified 12 out of 15 times with only a little confusion with the nearest categories. Hypermature Cataracts obtained a recall of 0.85, but they presented some overlap with Cortical and Nuclear. The largest problem for the model was Nuclear Cataracts, which it correctly predicted 7 out of 10, while assigning others to Hypermature and Posterior classes. VGG19 had an accuracy of 0.87 and a macro F1-score 0.87 reflecting balanced and reliable performance on all categories. With its impressive performance on Normal, Cortical, and Posterior cataract types, VGG19 is especially ideal for baseline diagnostic settings as it promises reliable categorization for not only cataract-positive but also normal eyes.

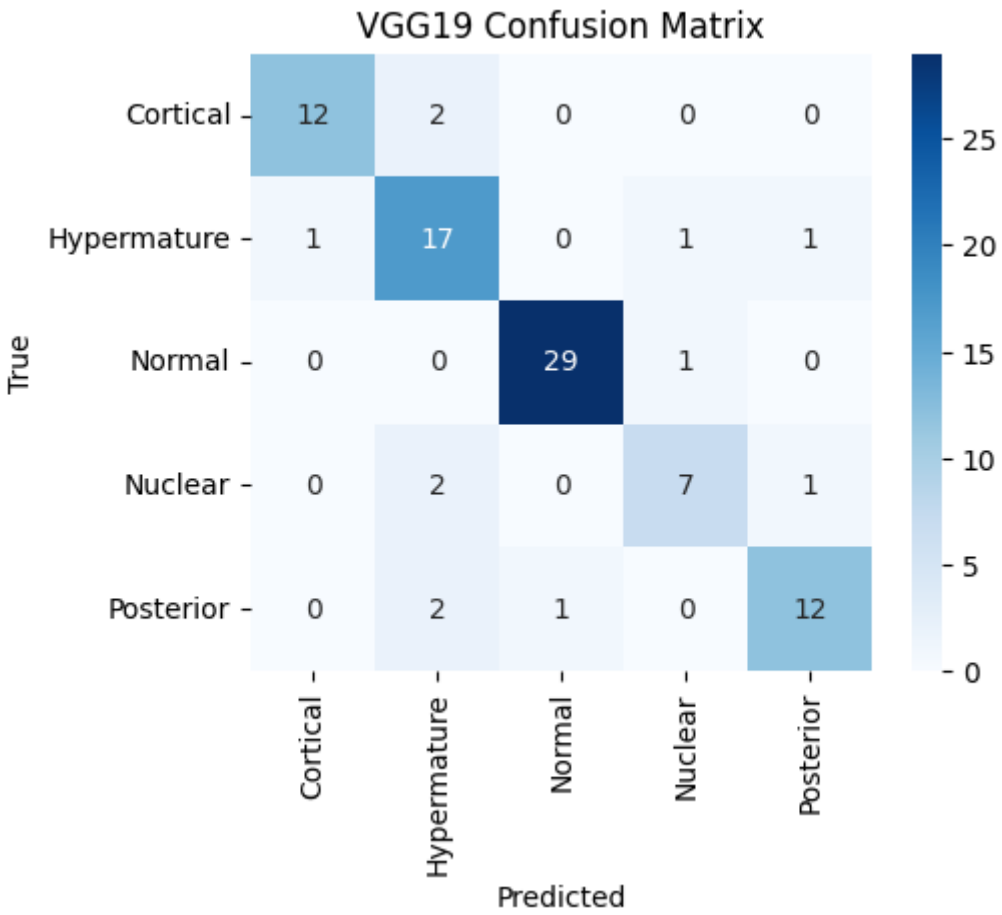

Figure 5.22: VGG19 Confusion Matrix

5.2.2 ResNet101

ResNet101 achieved solid class-wise results in various cataract categories. For Normal eyes, it obtained 100% recall, classifying correctly all 30 cases, and also showed a competitive performance on Cortical Cataracts with 13 out of 14 correct identifications. Posterior Cataracts were also well identified with 13 out of 15 in the correct class and little confusion with other classes. Hypermature Cataracts exhibited fair sensitivity with 14 of 20 identified correctly, but there were four misclassified as Nuclear type and one as Normal and cortical.

The most difficult category was Nuclear Cataracts, and only 7 of 10 were correctly predicted, all others being mistaken with Posterior, Hypermature or Normal classes. The value of the macro F1-score was 0.87 and accuracy was 0.87, which suggests that a good balance was achieved over all the classes of the test set. ResNet101, with a high recall on both Normal and Cortical eyes, and accurate prediction of Posterior Cataracts, is particularly suitable for clinical systems for which its focus is to reduce the number of missed diagnoses for the general cataract stages, thereby achieving an acceptable level of interpretability even for challenging cataract cases such as Nuclear and Hypermature Cataracts.

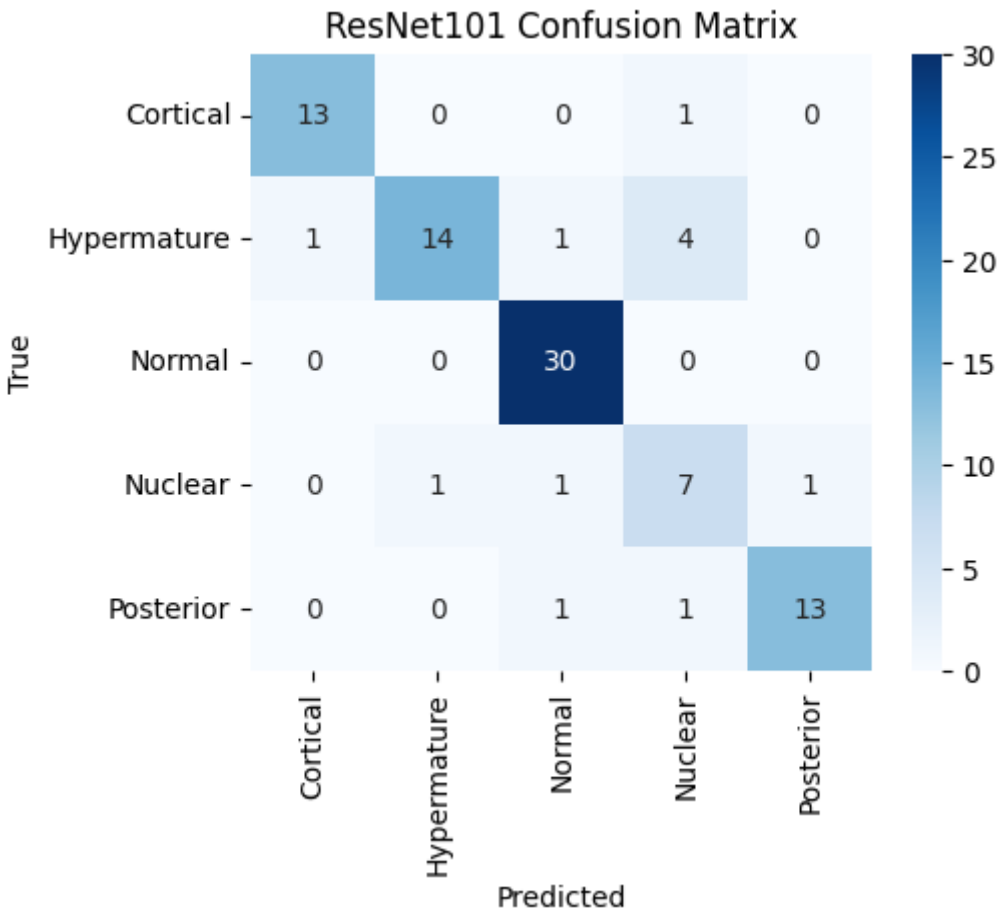

Figure 5.23: ResNet101 Confusion Matrix

5.2.3 InceptionV3

InceptionV3 performed differently when different cataract classes were presented, as five cataract classes were used for testing. The model performed well on Normal eyes (28/30) and on Posterior Cataracts (12/15). Cortical Cataracts were moderately reliably identified, with 11 correct and 3 misclassified, but 2 were classified as Hypermature. The most challenging class to predict was Hypermature Cataracts, where correct prediction was 13 out of 20 true Hypermature Cataracts among all classes, with confusion greatest between Nuclear (5 misclassifications) and Posterior types. Nuclear Cataracts were also difficult for the model: 7 out of 10 correct predictions were achieved, but a few misclassifications were scattered among Hypermature and Cortical. InceptionV3 had fairly balanced, but mottled performance across the dataset with a macro F1-score of 0.80 and accuracy 0.80. For that reason, its stable assignment of Normal and Posterior cases also makes it a viable alternative for screening systems which put more weight on definitive cataract-positive/negative eye discrimination, while additional models could be required to boost the performance for intermediate cataract stages.

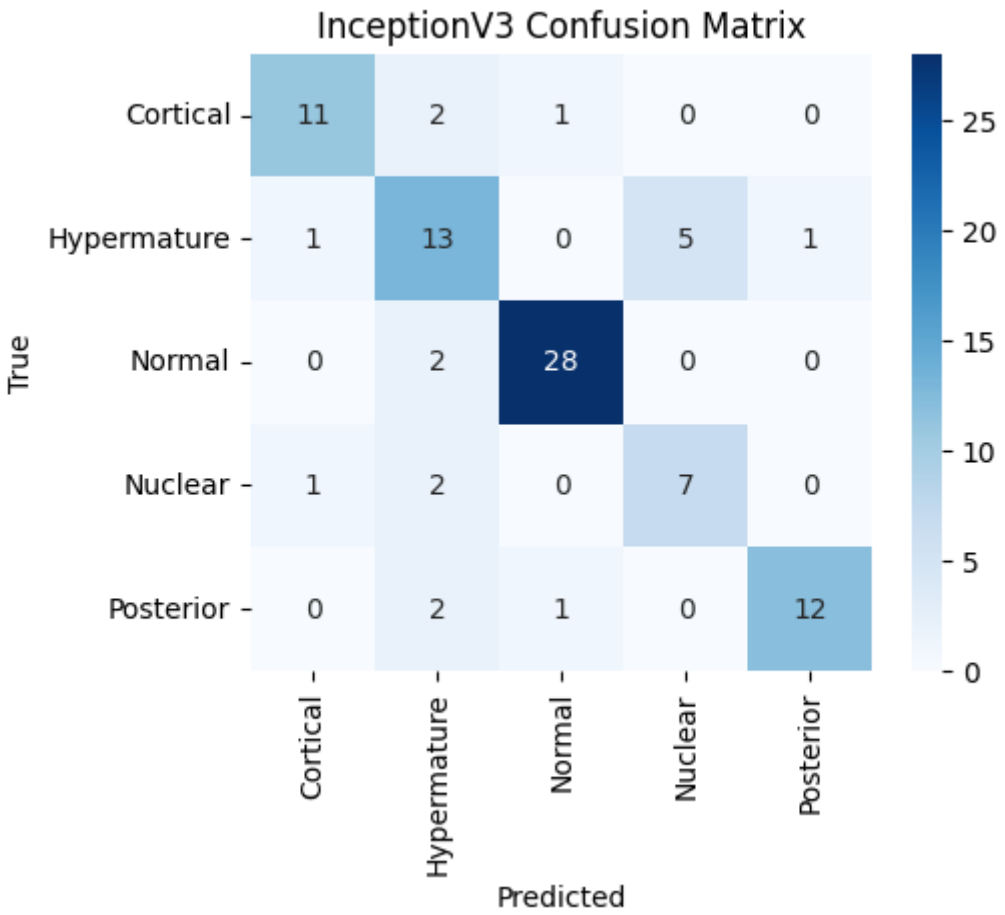

Figure 5.24: InceptionV3 Confusion Matrix

5.2.4 Densenet121

DenseNet121 exhibited good consistency for cataracts in different categories, particularly for the 30 Normal cases with perfect classification. The model did equally well on Cortical Cataracts, obtaining a precision of 13 out of 14 tests, with only one wrong classification into Nuclear. Hypermature Cataracts were relatively well-classified, yielding 16 correct from 20 , but two were incorrectly located with Cortical and one each with Nuclear and Posterior. Nuclear Cataracts were mildly challenging, with 7 out of 10 predicted correctly, where the main misclassifications occurred with the Posterior and Hypermature categories. Posterior Cataracts were characterized by more class overlapping, and 10 correct predictions were yielded among 15, with confusion among Cortical, Hypermature, and Normal classes. DenseNet121 demonstrated excellent generalization in all five classes with a macro F1-score of 0.85 and accuracy 0.85. Its robust performance on handling both normal and cataract-positive cases, particularly Cortical and Hypermature types, makes it a strong candidate for multiclass screening, where both sensitivity and structural detail extraction are desired.

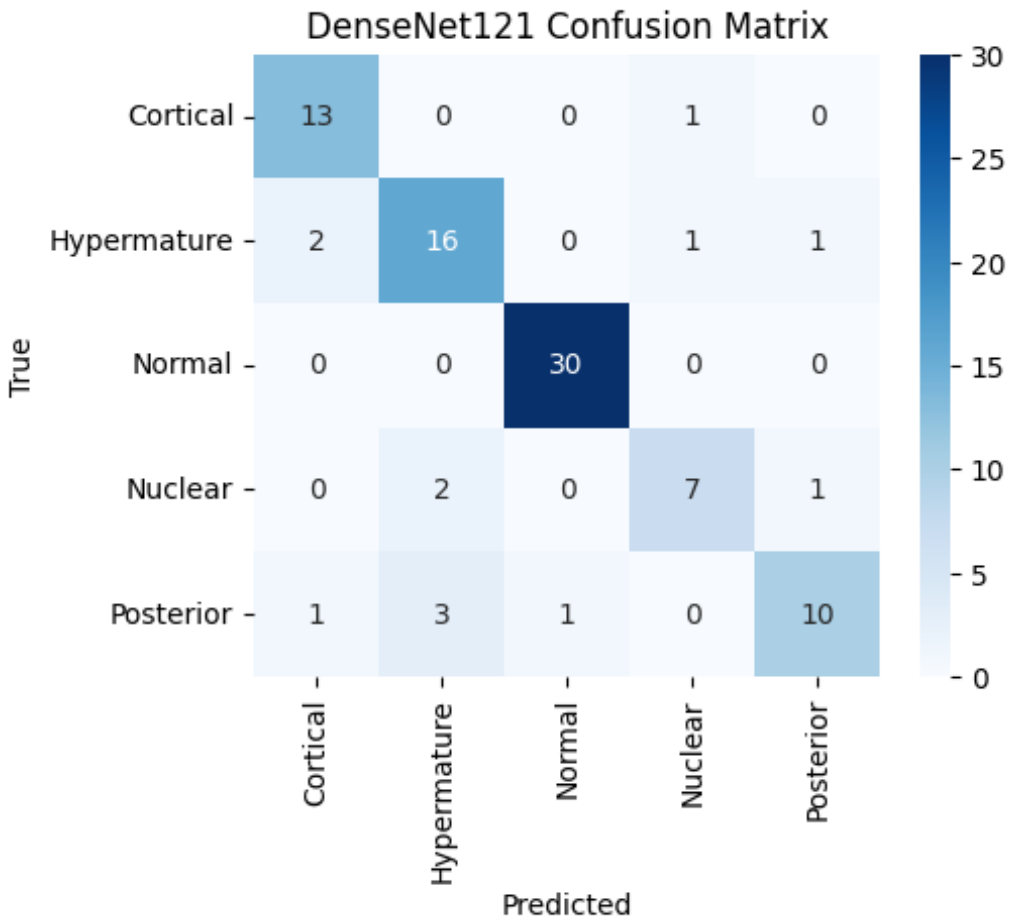

Figure 5.25: Densenet121 Confusion Matrix

5.2.5 MobileNetV2

MobileNetV2 exhibited consistently uniform performance across the classes with 29 of 30 precision rate on Normal cases and detected 12 of 14 Cortical Cataracts. The model had consistent detection of Posterior Cataracts, correctly classifying 12 out of 15 samples. For Nuclear Cataracts, 8 out of 10 predictions were correct. As for Hypermature Cataracts, MobileNetV2 correctly identified 16 of the 20 cases, however, there still was some confusion with Posterior and Cortical.

The macro F1-score was 0.87 and accuracy 0.87, indicating good reliability for all classes. Due to its reliable performance and computational efficiency, MobileNetV2 can be a good choice for real-time clinical systems or mobile diagnostic applications, particularly in clinics that demand stable classification for all types of cataract.

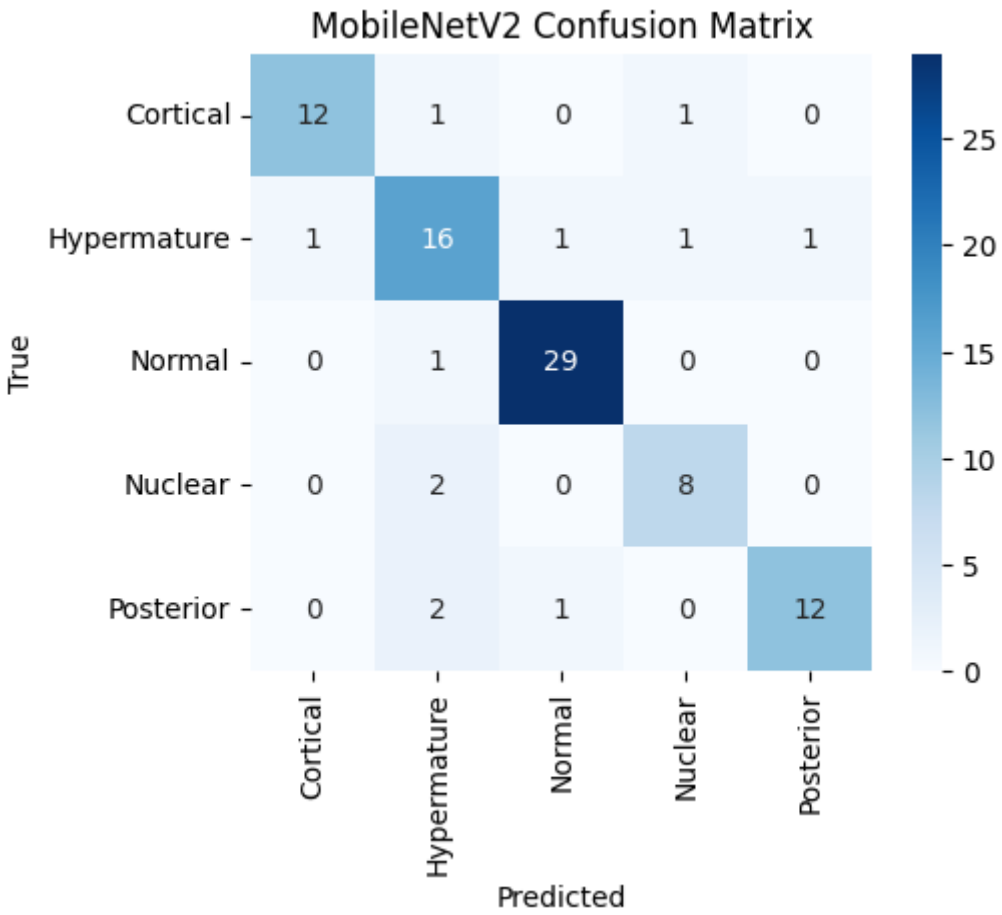

Figure 5.26: MobileNetV2 Confusion Matrix

5.2.6 EfficientNetB2

EfficientNetB2 had very good general performance for most of the cataract categories. The model exhibited a perfect classification for both the Cortical Cataracts (14/14) and for the Normal (30/30) eyes. The model's performance on the classification of Hypermature Cataracts was moderate, correctly classifying 15 out of 20, whereas the remainder of the samples were misclassified to the Nuclear, Posterior, or Normal classes. Posterior Cataracts were also foreseen with 12 out of 15 correct predictions. The hardest class was Nuclear Cataracts, for which the model made correct prediction of 6 out of 10 cases, and made almost all misclassifications to Hypermature and Posterior classes. Although such overlaps did occasionally occur, EfficientNetB2 obtained a macro F1-score of 0.86 and accuracy 0.87, representing good generalization among all the five categories. The model has high and uniform accuracy on Normal and Cortical images and reasonable accuracy on Posterior and Hypermature types, rendering it suitable for application in clinically-oriented systems which stress structural integrity and overall class balance.

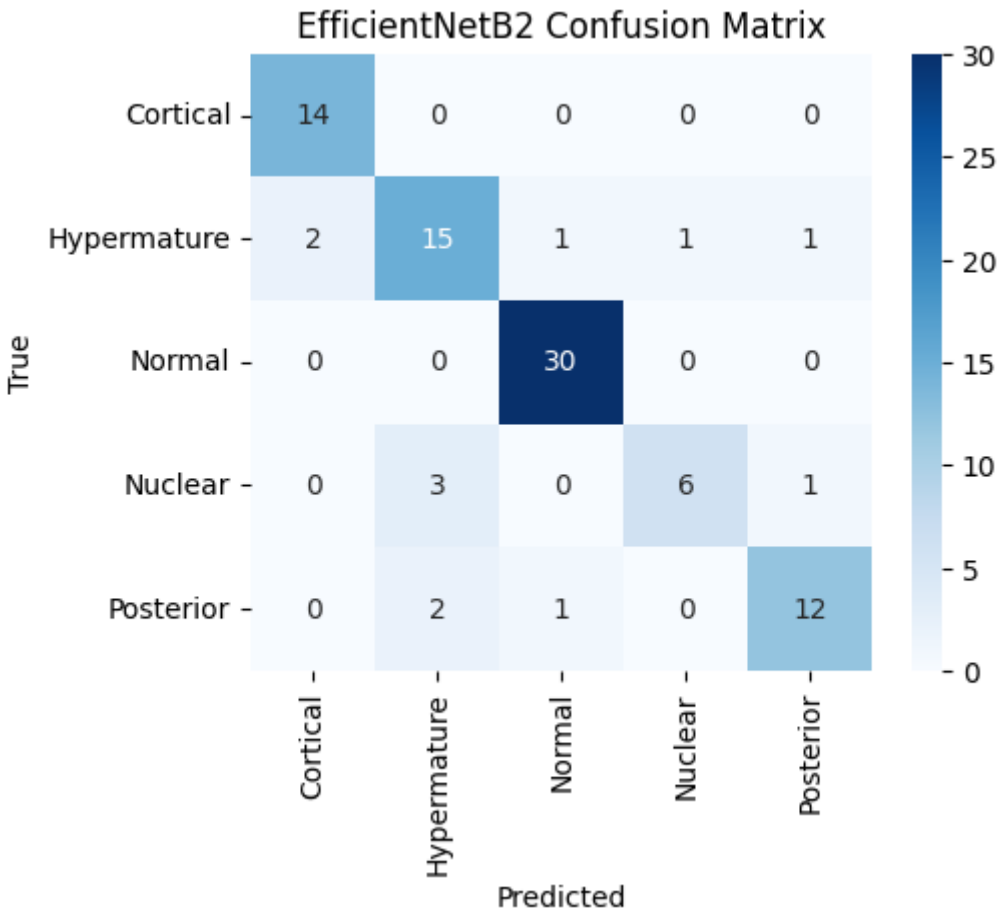

Figure 5.27: EfficientNetB2 Confusion Matrix

5.2.7 Xception

The Xception model provided a reliable performance for a few classes, especially for Normal eyes, where it achieved an accuracy of 29 out of 30 cases, and for Cortical Cataracts, where 12 out of 14 cases were correctly classified. Results were also good for Posterior Cataracts, where 12 out of the 15 cataracts were correctly classified. The model performed well on Nuclear Cataracts.The classification of Hypermature Cataracts was difficult, where we could classify only 12 out of 20 with high rates of being confused with Nuclear and Posterior. Since the classification errors come from the rest classes, such as an F1-score = 0.83 and accuracy was 0.83.Xception has reliable cataract type classifications, which is suitable in use cases that demand to be sensitive for having Normal and Nuclear.

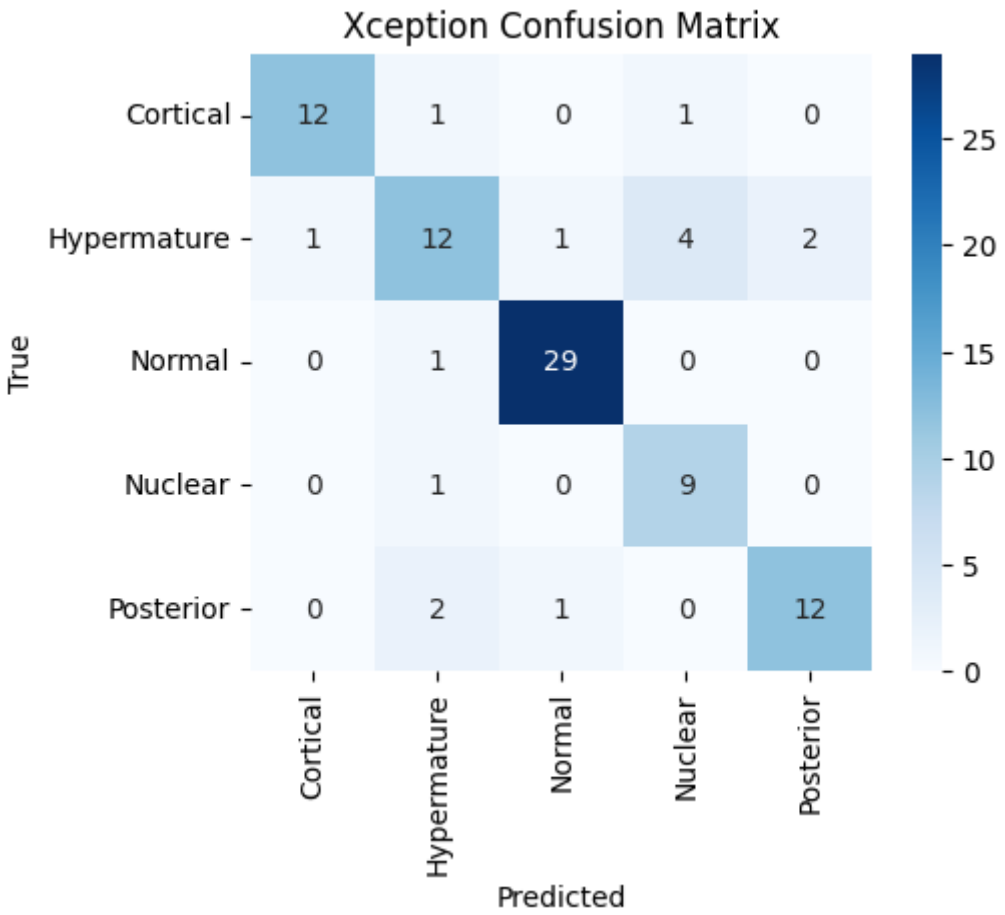

Figure 5.28: Xception Confusion Matrix

5.2.8 MobileViTV2

MobileViTV2 obtained good classification performance for almost all types of cataract compared to others, in particular identifying Normal samples with all 30 correctly predicted and Cortical Cataracts with a perfect recall rate (14/14). The model was equally successful for Posterior Cataracts, having accurately classified 13 scans out of 15. Nuclear Cataracts were reasonably well classified, with 8 out of 10 correct predictions, though 2 were shifted into other categories such as Hypermature and Posterior. The most uncertain category was Hypermature Cataracts, with only 14 out of 20 being correctly classified and the remaining six misclassified as either Cortical, Posterior, or Nuclear. MobileViTV2 had solid and robust performance with a macro F1-score of 0.89 and accuracy 0.89. Its lightweight architecture and high predictive reliability make it particularly suitable for deployment in mobile or edge-based diagnostic systems, especially in real-world environments where precision in Normal and Cortical classifications is paramount.

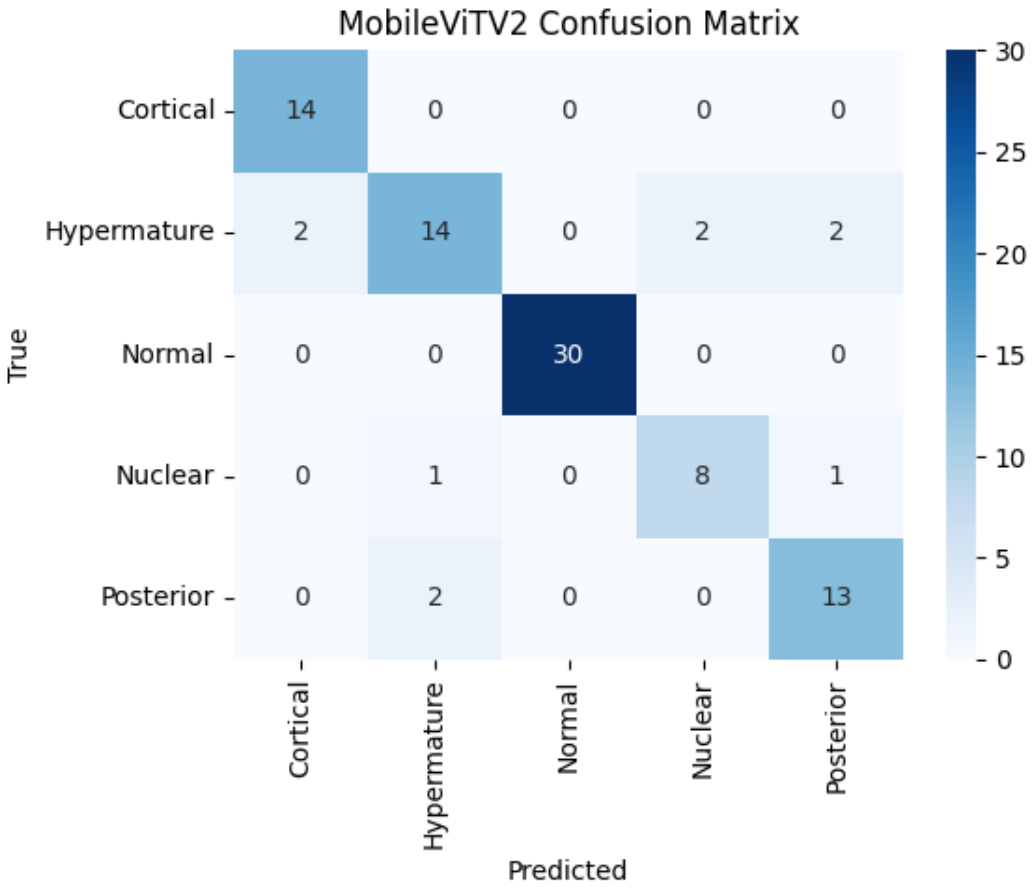

Figure 5.29: MobileViTV2 Confusion Matrix

5.2.9 FastViT

FastViT showed good performance for classification in most categories, with 100% recall for all 30 Normal and Cortical Cataract cases. The model did equally well on Posterior Cataracts, with 14 out of 15 correctly classified and low confusion. Nuclear Cataracts were well classified, with 8 out of 10 correctly identified , but 2 were misclassified as Hypermature. It was evident that the maximum variation occurred in the prediction of the Hypermature Cataract class, in which only 12 out of 20 cases were predicted accurately. The wrongly classified predictions were scattered across Nuclear, Normal, and Posterior types. FastViT achieved a macro F1-score of 0.87 and accuracy 0.88, which indicates that it works well in generalizing to both cataract and non-cataract classes. The exact auto-classification of Cortical, Posterior, and Normal categories, combined with strong spatial reasoning via its hybrid model, offers great advantage in clinical real-time tools, where clear and advanced cataract stages need to be distinguished with high precision.

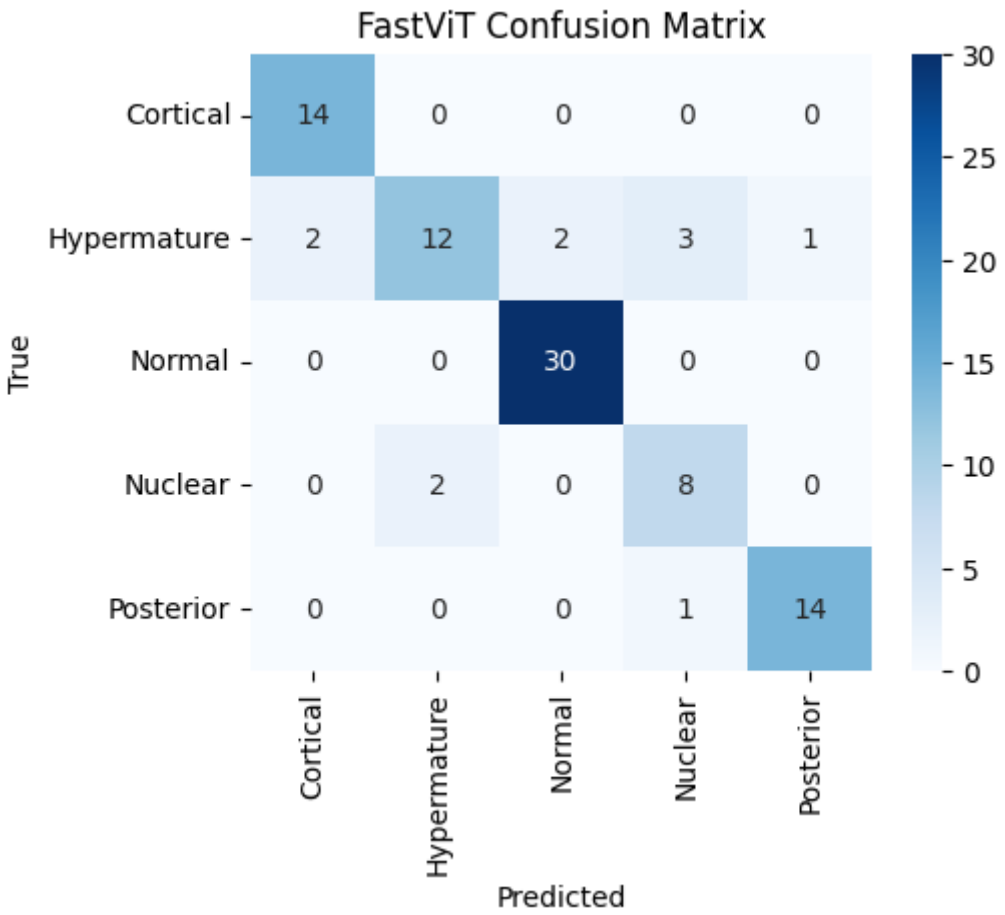

Figure 5.30: FastViT Confusion Matrix

5.2.10 ViT-B16

ViT-B16 excelled in recognising Normal eyes, with 29 out of 30 correctly predicted , and Posterior Cataracts, with 13 correct predictions out of 15. Nuclear Cataracts were reliably detected, with 8 of 10 cases correctly identified, while two cases were incorrectly classified as Cortical and Hypermature. Cortical Cataracts exhibited 12 out of 14 correct classifications but showed confusion with Hypermature and Nuclear types. The most difficult class was Hypermature Cataracts, where only 11 out of 20 were correctly identified, with misclassifications distributed among Posterior, Cortical, and Nuclear categories. Regardless of these difficulties, ViT-B16 reached a macro F1-score of 0.82 and accuracy 0.82, indicating overall balanced classification with slight underperformance in advanced cataract categories. Having transformer-based global attention, the ViT-B16 is particularly apt to capture global visual patterns and complex features, representing an effective tool for diagnostic systems with high precision in Normal, Posterior, and Nuclear cataracts, and is highly interpretable with tools like Grad-CAM.

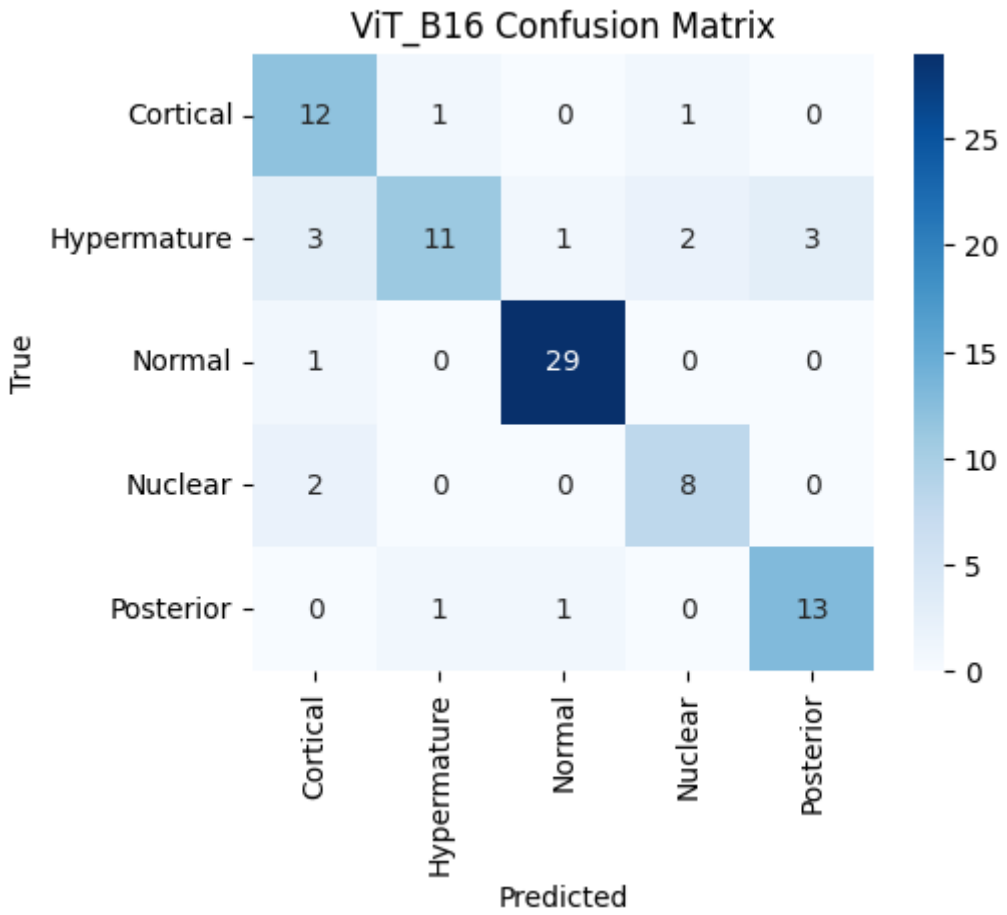

Figure 5.31: ViT-B16 Confusion Matrix

5.2.11 Stacked Ensemble Confusion Matrix

Out of all the models, the proposed stacked ensemble model — including the five top-performing CNN architectures (VGG19, EfficientNetB2, ResNet101, MobileNetV2, DenseNet121) and three transformer-based models (ViT-B16, FastViT, MobileViTV2) — exhibited the best overall trustworthiness and clinical consistency. The ensemble demonstrated improved stability of predictions for all five classes of cataract, leveraging architectural diversity and complementary capabilities through a logistic regression meta-learner.

As seen in the confusion matrix, Normal (30/30) and Cortical Cataracts (14/14) were classified perfectly by the model. Posterior Cataracts were properly identified 14 out of 15. Nuclear Cataracts were detected with a balanced precision was higher than various base learners. Hypermature Cataracts exhibited the most confusion, with 14 of 20 correctly classified and misclassifications spread across Nuclear, Posterior, and Cortical classes.

The ensemble achieved a macro F1-score of 0.90 and accuracy 0.90, showing excellent class separation power and discriminative ability in general. Standard parameters for sensitivity and specificity were also calculated for each class to evaluate the strength of the diagnosis. It should be noted that Normal and Cortical Cataracts

had 100% sensitivity, respectively. Posterior Cataracts detected 14/15, indicating high true positive detection and low false alarms. Nuclear and Hypermature classes both were detected 8/19, 14/20, leaving much room for improvement over individual models.

The results demonstrate the benefit of using ensemble learning in reducing the individual model biases and increasing the robustness over visually overlapped cataract stages. Furthermore, the stacking approach can benefit from local spatial sensitivity (acquired from CNNs) as well as global contextual attention (from ViT models), resulting in more meaningful diagnostic behavior from the clinical point of view.

Due to its robust classification performance, interpretable heat maps via Grad-CAM, and positive ophthalmologist feedback, the stacked ensemble model provides the most practically feasible solution for clinical implementation. It is especially well suited for high-throughput screening environments and mobile diagnostic platforms where diagnostic confidence and generalization over acquisition types are important.

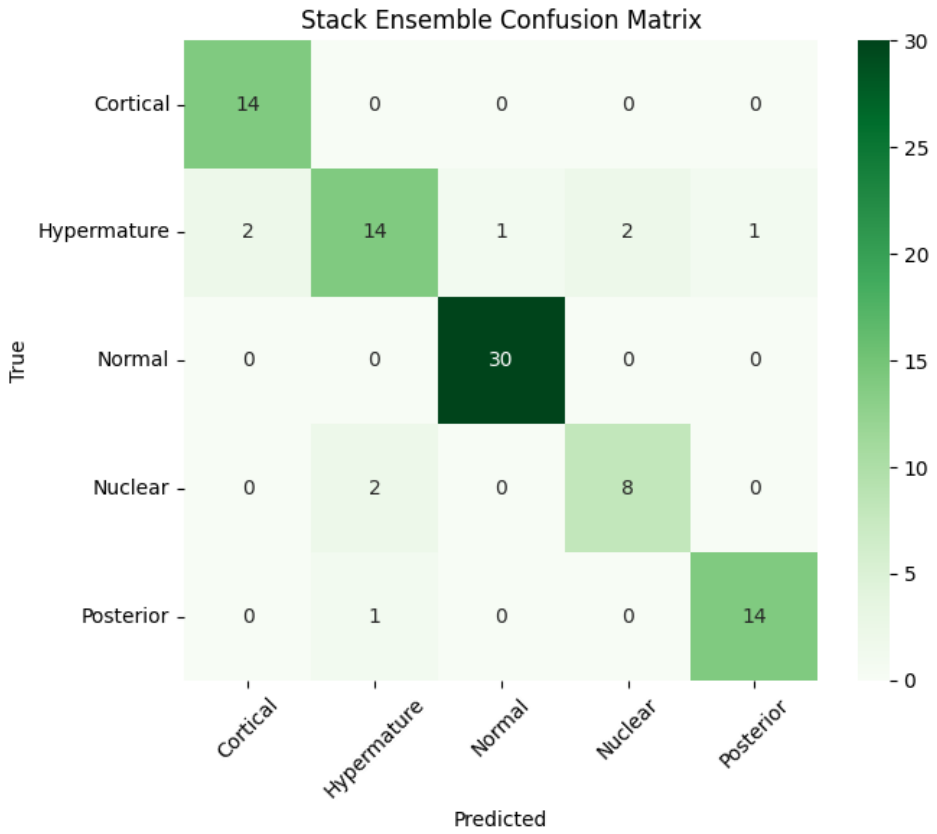

Figure 5.32: Stacked Ensemble Confusion Matrix

5.2.12 Bagging Ensemble Confusion matrix

The bagging ensemble model, used to combine multiple base learners obtained via randomized sampling and averaging techniques, provided a powerful and well-generalized performance in all five cataract categories. In contrast to stacking, which combines model outputs using an additional meta-learner, bagging reduces variance by training different instances of the same algorithm on bootstrapped subsets with a stable aggregate prediction.

As seen in the confusion matrix, the model classified all 30 Normal cases, and all 14 Cortical Cataracts correctly, with a recall of 1.00 for each class. It detected 9/10 Nuclear classes. For Posterior Cataracts, 14 out of 15 were correctly predicted, and Hypermature Cataracts showed moderate classification difficulty with 15 out of 20 correctly classified. These results demonstrate mastery of class-wise variability and the ensemble's ability to generalize across visually overlapping situations. The accuracy was 0.92 and F-1 Score was 0.92 further underlines the effective discriminative power of the ensemble across all categories, supporting high-confidence prediction when distinguishing normal from pathological classes. In summary, the bagging ensemble model provided a strong benchmark with high recall, high precision, high F-1 Score and excellent ROC performance, making it a useful candidate or complement to the stacking strategy. Its simplicity and robustness also make it suitable for applications where determinism and low variance are essential. Alongside stable Grad-CAM attention and clinical relevance, it serves as a reliable AI-assisted diagnostic tool in real-world cataract screening tasks.

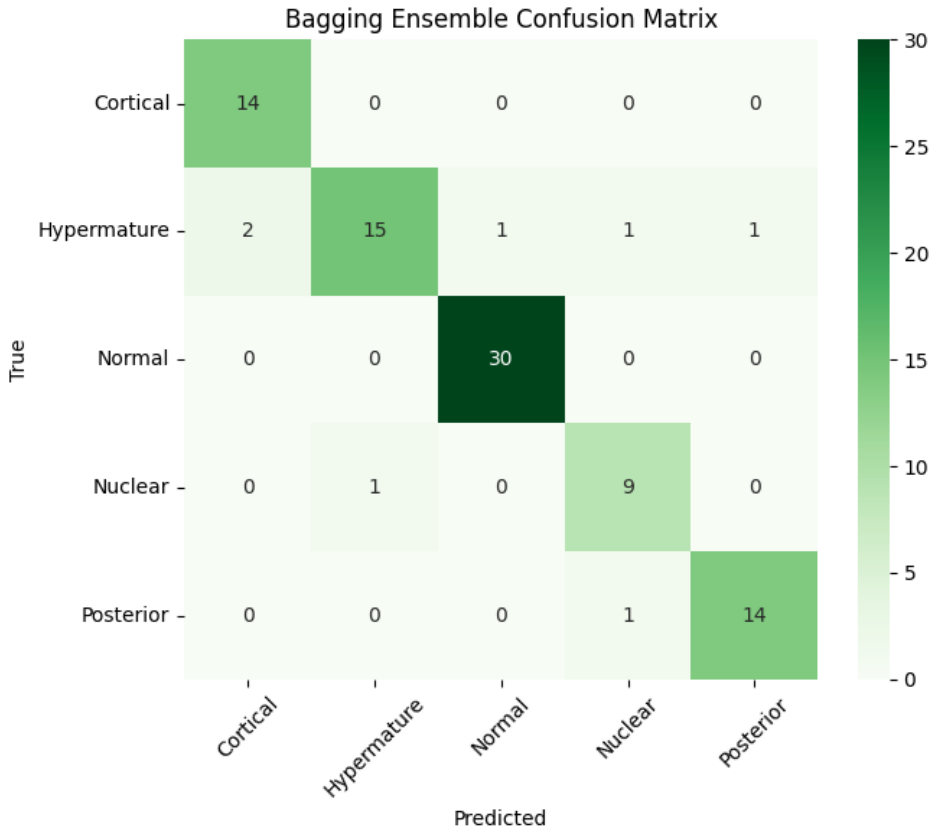

Figure 5.33: Bagging Ensemble Confusion matrix

5.3 Model Interpretability Using Grad-CAM

To interpret how our deep learning models predict, we employed the Grad-CAM (Gradient-weighted Class Activation Mapping). Specifically, we use the Grad-CAM technique to visually explain model decisions to identify if the input image contains a cataract or not. The procedure begins with a forward pass through the network by giving an input image (e.g., with shape of $224 \times 224 \times 3$) through the network to receive a classification prediction.

Next, we trigger a backward pass, during which we calculate the gradient of score we found for the predicted class in relation to the output feature maps of the final convolutional layer. This gradient informs us of the significance of each spatial location in the feature map to the final decision. We usually refer to the last convolutional layer of our model as `model.get_layer('block5_conv4').output` for VGG19, or a similarly named layer if using a different model.

We calculate the gradients using TensorFlow's `GradientTape`, with the input image being watched and the gradient of the output class score being calculated. These gradients are pooled together—averaged over width and height—to obtain a weight for each channel. These are then multiplied by the feature maps from the last convolutional layer, and by summing these up, we obtain a heatmap that is passed through a ReLU activation, keeping only positive influences. Finally, we resize this heatmap to the size of the original image, and then overlay it on the original image using a heatmap opacity (alpha) and a colormap (such as `jet` or `viridis`) to indicate where the model was looking when making predictions.

This visualization helps confirm whether the model is learning the correct features. In our case, Grad-CAMs for the prediction of cataract consistently indicated the cloudy blurred part in various areas of eyes, in accordance with clinical exam findings. By contrast, in the normal eyes, the attention was distributed rather evenly, without any particular focus suggesting they were free of defects. Overall, these methods not only enhance transparency in our models but also increase trust in AI-aided medical diagnosis by visually verifying that the model's focus aligns with medically relevant regions of the image, even later confirmed by ophthalmologists.

On the next page we showed the visual examples of GradCam results for the ensemble methods that gave us the best results.

5.3.1 Grad-Cam Results from Ensemble Bagging Approach

Here, for each testing image we processed the top models (both ViTs and CNNs) through a same Grad-CAM pipeline where we preprocessed the input, computed the class specific gradient heatmap with resized the heatmap to the input resolution, and finally we averaged these heatmaps `avg_hm = np. mean(heatmaps, axis=0)`. The averaged map was converted into an 8-bit jet colormap, overlaid at 40% opacity onto the raw image and the final prediction was taken by averaging the softmax outputs of self model. In our five cases, the bagging ensemble demonstrated 100% accuracy, and their smoothed Grad-CAM overlays consistently uncover each cataract subtype's key area:

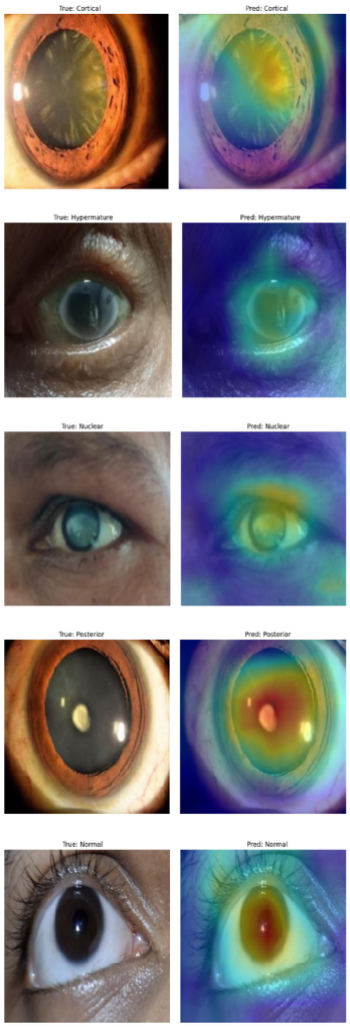

Figure 5.34: Ensemble Bagging Grad-Cam Results

- **Cortical**: Detects a soft yet specific activation along the peripheral, spoke-like opacities.
- **Hypermature**: Detects a wide, equatorial spread with an evenly hazy opaque lens.
- **Nuclear**: Detects a pronounced central hotspot contained entirely on top of the compact nucleus.

- **Posterior**: Noted a localized focus of posterior capsule plaque.
- **Normal**: Uniform soft attention, no hot spots of attention.

5.3.2 Grad-Cam Results from Ensemble Stacking Approach

For this case, we simply appended individual model's predicted probability in to a single feature and passed through a meta-classifier `meta_clf`, as follows. The predicted class determined which row of *metaclf.coef* we used to extract per-model weights. These weights which are absolved and normalized were then used as a coefficient for a weighted sum of the single Grad-CAM maps. The heatmap was colored and overlaid, and the final label was selected from the highest probability class of the meta-classifier. This stacking also achieved 100% accuracy on the five holdouts, but its attention maps are crisper than the previous method with every model contributes attention that is tuned by the meta-leaner

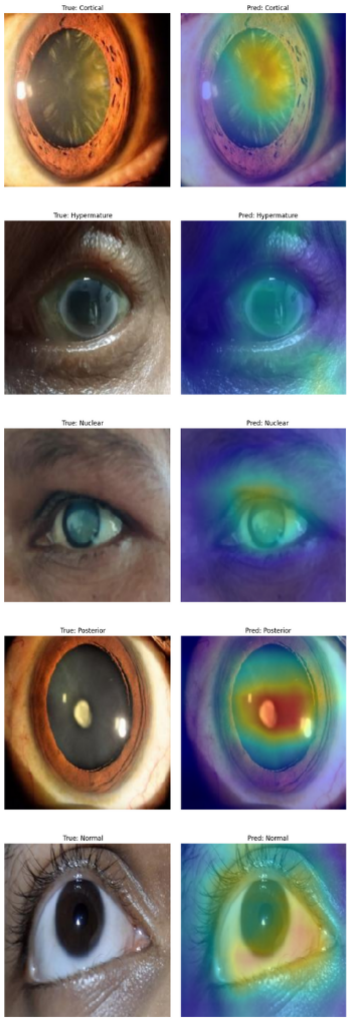

Figure 5.35: Ensemble Stacking Grad-Cam Results

- **Cortical**: Sharply focused on the outer cortical spokes of the eye.

- **Hypermature**: Detects matching figures of uniformly opaque and bright cortex.
- **Nuclear**: Pinpoint central activation over the nucleus.
- **Posterior**: Tight hotspot at the posterior capsule position.
- **Normal**: Shows low-intensity attention across the clear lens.

The Grad-CAM overlays in both ensembles are aligned with the clinical features. However, bagging results in smoother, consensus-style maps, while the learned weights in stacking produce stronger, class-specific focus but couldn't produce as good heatmap as the other technique. In the future, stacking can be improved using various techniques such as training the meta classifier with an added fine tuned custom head.

5.4 Real-life Professionals' Feedback

To make the results clinically applicable, the research was continuously supported by the senior ophthalmologist Prof. M. A. Bashar Sheikh from the beginning of the project (data collection) till the conclusion of the whole study to make it medically sound. Drawing upon his extensive experience in the clinic and a long career in medical research, Prof Bashar was able to share his perspective on cataract diagnosis, interpretation of images, and the on-the-ground application of AI in screening.

The Grad-CAM heatmaps and confusion matrices of each individual model that we implemented in our thesis, as well as the final stacked ensemble, bagging ensemble were personally examined by Prof. Bashar. He also emphasized that the ensemble method, combining both of the CNN and ViT models, showed highly stable and reliable classification results in all cataract types. His skepticism was based on the fact that the sensitivity(recall) for Normal and Cortical cases were maintained by the ensemble and that confusion was essentially avoided in the more difficult categories: Posterior and Hypermature Cataracts. The predictions of the Stacking and Bagging Ensemble learning showed, he said, a high degree of agreement with the clinical expectation and the low misclassification rate would validate its role in aiding diagnostic workflows.

Regarding interpretability, Prof. Bashar confirmed that MobileVitV2, Stacked Ensemble and Bagging Ensemble were the most diagnostically interpretable models, in that both consistently highlighted clinically relevant regions on attention maps for our classification task. Nonetheless, he also finds it satisfactory that at this stage the models effectively identified most cataract areas in both slit-lamp imaging and retained robust performance in different illumination and quality circumstances as commonly observed in mobile-captured photos.

He also acknowledges that the possibility offered by lightweight models like Vit-B16, FastVit and MobileNetV2 for deployment in a mobile environment might be huge in settings where specialized care is not available: rural or resource-constrained environments. He thus provided input in driving the clinical side of this research and giving validations regarding its applicability.

Chapter 6

Conclusion

In this paper, we propose a clinically validated deep learning-powered model capable of classifying cataracts in five classes: Nuclear, Cortical, Posterior Subcapsular, Hypermature, and Normal. The system is trained on a large dataset of slit lamp and mobile (captured outside a slit lamp environment) eye images, recorded under ophthalmologist supervision and is designed to handle real world variation due to factors including lighting, focus, and device noise. Various CNN and transformer models such as VGG19, ResNet101, EfficientNetB2, DenseNet121, MobileNetV2, ViT-B16, FastViT, and MobileViTV2 were fine-tuned based on transfer learning, and ensemble techniques were utilized. Bagging and stacking (with logistic regression as meta-learner) were also performed and resulted in better performance and consistency; the best accuracy was up to 92%. The visualizations by Grad-CAM demonstrated practical interpretability by focusing attention on the clinically significant areas. Performance characteristics of the device were determined using clinical testing, and its feasibility for use in rural and resource-limited settings was evaluated. Mobile light-weight models like FastViT, ViT-B16, and MobileViTV2 promote mobile-based real-time screening. This work opens the door to a cost-effective, scalable solution for cataract diagnosis, making preventable blindness, that which is curable or preventable, a manageable disease.

Bibliography

[1] C. Kupfer, "Bowman lecture. the conquest of cataract: A global challenge," *Trans Ophthalmol Soc UK*, vol. 104, pp. 1–10, 1984.

[2] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 4th. Prentice Hall, 2002, ISBN: 9780131687288.

[3] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006, Chapter 4: Linear Models for Classification, Section 4.3.2 (The Softmax Function), ISBN: 978-0-387-31073-2.

[4] R. U. Acharya, W. Yu, K. Zhu, J. Nayak, T.-C. Lim, and J. Y. Chan, "Identification of cataract and post-cataract surgery optical images using artificial intelligence techniques," *Journal of medical systems*, vol. 34, pp. 619–628, 2010.

[5] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 2010, pp. 807–814.

[6] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[7] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2014. [Online]. Available: https://arxiv.org/abs/1312.4400.

[8] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014. [Online]. Available: http://jmlr.org/papers/v15/srivastava14a.html.

[9] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, vol. 37, PMLR, 2015, pp. 448–456. [Online]. Available: https://arxiv.org/abs/1502.03167.

[10] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations (ICLR)*, 2015. arXiv: 1412.6980 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1412.6980.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015. [Online]. Available: https://arxiv.org/abs/1409.1556.

[12] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, ISBN: 9780262035613. [Online]. Available: https://www.deeplearningbook.org/.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90. [Online]. Available: https://doi.org/10.1109/CVPR.2016.90.

[14] 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 (CVPR)*, 2016, pp. 2818–2826. DOI: 10.1109/CVPR.2016.308. [Online]. Available: https://doi.org/10.1109/CVPR.2016.308.

[15] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1251–1258. DOI: 10.1109/CVPR.2017.195. [Online]. Available: https://doi.org/10.1109/CVPR.2017.195.

[16] N. Egejuru, J. Balogun, P. Mhambe, F. Asahiah, and P. Idowu, "Model for prediction of cataracts using supervised machine learning algorithms," 2017.

[17] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269. DOI: 10.1109/CVPR.2017.243.

[18] M. J. Primus, D. Putzgruber-Adamitsch, M. Taschwer, *et al.*, "Frame-based classification of operation phases in cataract surgery videos," in *MultiMedia Modeling: 24th International Conference, MMM 2018, Bangkok, Thailand, February 5-7, 2018, Proceedings, Part I 24*, Springer, 2018, pp. 241–253.

[19] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4510–4520. DOI: 10.1109/CVPR.2018.00474. [Online]. Available: https://doi.org/10.1109/CVPR.2018.00474.

[20] K. Schoeffmann, M. Taschwer, S. Sarny, B. Münzer, M. J. Primus, and D. Putzgruber, "Cataract-101: Video dataset of 101 cataract surgeries," in *Proceedings of the 9th ACM multimedia systems conference*, 2018, pp. 421–425.

[21] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 97, PMLR, 2019, pp. 6105–6114. [Online]. Available: http://proceedings.mlr.press/v97/tan19a.html.

[22] X. Wu, Y. Huang, Z. Liu, *et al.*, "A universal artificial intelligence platform for collaborative management of cataracts," *The Lancet*, vol. 394, S22, 2019.

[23] F. Yu, G. S. Croso, T. S. Kim, *et al.*, "Assessment of automated identification of phases in videos of cataract surgery using machine learning and deep learning techniques," *JAMA network open*, vol. 2, no. 4, e191860–e191860, 2019.

[24] H. Zhang, K. Niu, Y. Xiong, W. Yang, Z. He, and H. Song, "Automatic cataract grading methods based on deep learning," *Computer methods and programs in biomedicine*, vol. 182, p. 104 978, 2019.

[25] M. Fox, M. Taschwer, and K. Schoeffmann, "Pixel-based tool segmentation in cataract surgery videos with mask R-CNN," in *33rd IEEE International Symposium on Computer-Based Medical Systems, CBMS 2020, Rochester, MN, USA, July 28-30, 2020*, A. G. S. de Herrera, A. R. González, K. C. Santosh, Z. Temesgen, B. Kane, and P. Soda, Eds., IEEE, 2020, pp. 565–568. DOI: 10.1109/CBMS49503.2020.00112.

[26] J. H. L. Goh, Z. W. Lim, X. Fang, *et al.*, "Artificial intelligence for cataract detection and management," *Asia-Pacific journal of ophthalmology*, vol. 9, no. 2, pp. 88–95, 2020.

[27] W. Li, Y. Yang, K. Zhang, *et al.*, "Dense anatomical annotation of slit-lamp images improves the performance of deep learning for the diagnosis of ophthalmic disorders," *Nature Biomedical Engineering*, vol. 4, no. 8, pp. 767–777, 2020.

[28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2021.

[29] S. Hu, X. Luan, H. Wu, *et al.*, "Accv: Automatic classification algorithm of cataract video based on deep learning," *BioMedical Engineering OnLine*, vol. 20, pp. 1–17, 2021.

[30] M. S. Junayed, M. B. Islam, A. Sadeghzadeh, and S. Rahman, "Cataractnet: An automated cataract detection system using deep learning for fundus images," *IEEE Access*, vol. 9, pp. 128 799–128 808, 2021. DOI: 10.1109/ACCESS.2021.3112938.

[31] J. Ladas, D. Ladas, S. R. Lin, U. Devgan, A. A. Siddiqui, and A. S. Jun, "Improvement of multiple generations of intraocular lens calculation formulae with a novel approach using artificial intelligence," *Translational Vision Science & Technology*, vol. 10, no. 3, pp. 7–7, 2021.

[32] M. D. Toro, A. P. Brézin, M. Burdon, *et al.*, *Early impact of covid-19 outbreak on eye care: Insights from eurocovcat group*, 2021.

[33] L. Gutierrez, J. S. Lim, L. L. Foo, *et al.*, "Application of artificial intelligence in cataract management: Current and future directions," *Eye and Vision*, vol. 9, no. 1, p. 3, 2022.

[34] S. Mehta and M. Rastegari, "Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer," *arXiv preprint arXiv:2206.02680*, 2022. [Online]. Available: https://arxiv.org/abs/2206.02680.

[35] S. Mehta and M. Rastegari, "Separable self-attention for mobile vision transformers," *arXiv preprint arXiv:2206.02680*, 2022.

[36] K. Y. Son, J. Ko, E. Kim, *et al.*, "Deep learning-based cataract detection and grading from slit-lamp and retro-illumination photographs: Model development and validation study," *Ophthalmology Science*, vol. 2, no. 2, p. 100 147, 2022.

[37] D. Tognetto, R. Giglio, A. L. Vinciguerra, *et al.*, "Artificial intelligence applications and cataract management: A systematic review," *survey of ophthalmology*, vol. 67, no. 3, pp. 817–829, 2022.

[38] X. Wu, D. Xu, T. Ma, *et al.*, "Artificial intelligence model for antiinterference cataract automatic diagnosis: A diagnostic accuracy study," *Frontiers in Cell and Developmental Biology*, vol. 10, p. 906 042, 2022.

[39] P. K. Anasosalu Vasu, J. G. J. Zhu, O. Tuzel, and A. Ranjan, "Fastvit: A fast hybrid vision transformer using structural reparameterization," *arXiv preprint arXiv:2303.14189*, 2023. [Online]. Available: https://arxiv.org/abs/2303.14189.

[40] F. Gan, H. Liu, W.-G. Qin, and S.-L. Zhou, "Application of artificial intelligence for automatic cataract staging based on anterior segment images: Comparing automatic segmentation approaches to manual segmentation," *Frontiers in Neuroscience*, vol. 17, p. 1 182 388, 2023.

[41] T. Ganokratanaa, M. Ketcham, and P. Pramkeaw, "Advancements in cataract detection: The systematic development of lenet-convolutional neural network models," *Journal of Imaging*, vol. 9, no. 10, p. 197, 2023.

[42] M. Ghayoumi, *Generative Adversarial Networks in Practice*. Boca Raton, FL: CRC Press, 2023, ISBN: 9781032458494.

[43] D. Kumar, B. Bakariya, C. Verma, and Z. Illes, "Cataract disease identification using transformer and convolution neural network: A novel framework," in *2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS)*, IEEE, 2023, pp. 1230–1235.

[44] E. Shimizu, M. Tanji, S. Nakayama, *et al.*, "Ai-based diagnosis of nuclear cataract from slit-lamp videos," *Scientific reports*, vol. 13, no. 1, p. 22 046, 2023.

[45] C. S. Vasan, S. Gupta, M. Shekhar, *et al.*, "Accuracy of an artificial intelligence-based mobile application for detecting cataracts: Results from a field study," *Indian Journal of Ophthalmology*, vol. 71, no. 8, pp. 2984–2989, 2023.

[46] S. Yadav and J. K. P. Singh Yadav, "Enhancing cataract detection precision: A deep learning approach.," *Traitement du Signal*, vol. 40, no. 4, 2023.

[47] S. Yadav, J. K. P. S. Yadav, *et al.*, "Automatic cataract severity detection and grading using deep learning," *Journal of Sensors*, vol. 2023, 2023.

[48] N. Ghamsarian, Y. El-Shabrawi, S. Nasirihaghighi, *et al.*, "Cataract-1k dataset for deep-learning-assisted analysis of cataract surgery videos," *Scientific data*, vol. 11, no. 1, p. 373, 2024.

[49] M. Khairudin, R. Mahaputra, W. Wiharto, *et al.*, "Early detection of diabetes potential using cataract image processing approach," *SINERGI*, vol. 28, no. 1, pp. 55–62, 2024.

[50] C. C. Manganti, *Optiscan - cataract detection dataset*, Open Source Dataset, Apr. 2024. [Online]. Available: https://universe.roboflow.com/christopher-clark-manganti/optiscan-cataract-detection.

[51] V. Singh, *keras-vision: A collection of vision models in Keras*, 2024. [Online]. Available: https://pypi.org/project/keras-vision/.

[52] M. A. VM, N. Tiwari, V. Khobragade, *et al.*, "Clinical validation of artificial intelligence-based cataract screening solution with smartphone images (logy ai cataract screening module)," *International Journal of Advances in Medicine*, vol. 11, no. 2, p. 71, 2024.

[53] S. Lu, L. Ba, J. Wang, *et al.*, "Deep learning-driven approach for cataract management: Towards precise identification and predictive analytics," *Frontiers in Cell and Developmental Biology*, vol. 13, p. 1 611 216, 2025.