SKIN LESION BINARY CLASSIFICATION

A dissertation submitted to Manchester Metropolitan University
for the degree of Master of Science
in the Faculty of Science and Engineering

2024

By
Tooba Zahid
Department of Computing and Mathematics

Contents

List of Tables

List of Figures

Abstract

Early and accurate diagnosis of melanoma, a life threatening form of skin cancer, is crucial to improve patient outcomes.. In this dissertation, a study of advanced deep learning approaches for predicting whether a skin lesion is benign or malignant is provided. This study seeks to enhance classification accuracy utilizing Deep Convolutional Neural Network (CNN) architectures: AlexNet, InceptionV3, ResNet, DenseNet and EfficientNet, by targeting optimizations.

An inherent class imbalance in the dataset, that was sufficiently mitigated using use of focal loss and hyperparameter tuning, yielded substantial performance improvements. Each of these models was tested on precision, recall, accuracy, and F1 score, for which ensemble learning uses the strengths of several architectures to again enhance performance.

Results showed that baseline models were able to achieve reasonable accuracy, though application of optimization techniques brings substantial improvement especially in malignant cases, where ensemble models maximize F1 scores and increase accuracy beyond 83%. It shows the efficacy of tailoring optimizer in the clinical diagnostic fields. Future studies aim to refine model comprehensibility and tackle limitations in datasets for greater clinical relevance.

Declaration

No part of this project has been submitted in support of an application for any other degree or qualification at this or any other institute of learning. Apart from those parts of the project containing citations to the work of others, this project is my own unaided work. This work has been carried out in accordance with the Manchester Metropolitan University research ethics procedures, and has received ethical approval number 68459.

Signed:

Tooba Zahid

Date:

03-10-2024

Acknowledgements

I want to acknowledge Dr. Nashid Alam for her incredible advice and motivation during this project. Her expertise in technical areas and valuable suggestions played a crucial role in concluding my research effectively.Her direction and patience during the experimental phase were invaluable, and his advice on the theoretical aspects helped shape my understanding of the subject matter.

It is rewarding to have experienced her understanding personality and knowledge in the area made the research journey both fruitful and pleasurable. Throughout the difficult moments her steady backing has stayed with me deeply and I am incredibly thankful.

Additionally I want to thank Manchester Metropolitan University for providing the resources and facilities that facilitated this research. I want to thank the professors and staff for their help and support during this work.

I want to express my sincere thanks to my family and friends for their ongoing encouragement and trust throughout this journey.

Abbreviations

CNNs	Convolutional Neural Networks
ResNet	Residual Networks
ISIC	International Skin Imaging Collaboration
AUC-ROC	Area Under the Receiver Operating Characteristic

Chapter 1

Introduction

1.1 Project Overview

Skin Cancer is a major health issue that affects a large number of individuals, and it is vital to recognize it at an early stage. The binary skin lesion classification project can tackle this by establishing a methodology for the analysis of the images, and classification of the lesions as either benign or malignant. This can be helpful in diagnosing skin cancer to the dermatologist on duty specially in complicated cases.

The Skin Lesion Binary Classification Project has a goal for constructing the more enhanced deep learning model for the fundamental binary classification of the lesion skin whether it is malignant or benign. This effort is rather necessary because melanoma and other skin cancers are considered among the diseases with the highest mortality rates, and timely identification of the pathology is life-saving. Utilizing the deep convolutional neural networks (DCNNs) the project employs the commonly reputed models like AlexNet and Inception V4 best suited for the image classification tasks. Some of these models will be trained with dermatoscopic type of images with good data preparation in terms of resizing, normalization and augmentation of the images. The project also carries out a study of adding segmentation techniques that may enhance the classification in a way by tackling issues like; lesion localization and if there would be a need to tackle class imbalance then it will be done by the aid of tools like the focal loss function.

1.2 Potential Problem

The primary issues that should be addressed when speaking about skin lesion binary classification with deep learning was data quality and quantity in analysing the data obtained from the field, there was a realization that the quality and quantity of data were critical success factors. This concept has a very wide range, however, for the targets of this project; it focused on the need for a large pool of dermatoscopic images with high quality to test the generality of the model. The other consideration proposed was the issue with lesion features' intricacy. Essentially, benign and malignant lesions may not look very different, which would pose a challenge to the model in terms of how it may differentiate (M Emre Celebi et al. 2007).

The other major problem is overfitting, which varies based on the complexity of models being used, such as AlexNet or Inception V3. This problem is born when the modelling of the process is executed in such a way that the model captures noise and outliers from training data, and a vastly poor performance is encountered on new data. It is worth mentioning that, techniques such as dropout, regularization, and cross-validation are helpful to solve the overfitting issue (Srivastava et al. 2014b).

Finally, ensuring the model's generalizability to different populations is crucial. Variations in skin type, age, and geographical location can affect the appearance of lesions, necessitating a diverse and representative dataset for training and validation (Brinker et al. 2019).

1.3 Aims and Objectives

Aim:

The main purpose of this project is to build a reliable deep learning model, which is able to identify the type of skin lesion image as benign or malignant. The idea behind the project is to enhance diagnostic performance and apply automated skin lesion classification systems to a greater extent by solving a set of the main problems in the analysis of medical images. Some of these difficulties include class imbalance where malignant instances are less in number, and the techniques that will be used to overcome them include focal loss or oversampling. It means that the nature of input data is to be normalized, free from noises and augmented to ensure improvement in training. To further minimize on overfitting the following techniques will be used; regularization, dropout and cross validation.

The project for skin lesion classification will employ individual deep learning models as AlexNet and Inception v3 besides utilizing consolidated ensembles from over other architectures like EfficientNet, DenseNet or ResNet. This ensemble approach is anticipated to enhance the classification accuracy since different models are built, and the errors which arise from such construction are usually minimized. This means that the detection rate, error rate, accuracy, sensitivity and specificity will be used to measures the performance of the system. Further, the validity of the model for different populations will be verified on the gathered data, thus its applicability in practical clinical setting.

Objectives:

1. Perform research on the existing deep learning models to classify skin lesions.
2. First step is to obtain and prepare the skin lesion image data set.
3. Optimize architectures based on the discovered DCNN models like AlexNet and Inception V3.
4. Combine segmentation with the classification model.
5. As far as the class balanced, initially apply standard binary cross-entropy loss, also consider focal loss if class imbalance becomes a concern.
6. ensemble approach is anticipated to enhance results
7. Assess the goodness of fit of the model on adequate indices like accuracy, sensitivity, and specificity.

1.4 Tools and Timeline

For a skin lesion binary classification project, various advanced tools and techniques are utilized to ensure effective image analysis and model training. Key tools include large datasets like ISIC Archive, and image processing libraries such as OpenCV for tasks like resizing, normalization, augmentation, and more complex operations. Deep learning frameworks like TensorFlow, Keras, and PyTorch are used for building and training models, leveraging pre-trained architectures like AlexNet, Inception V3, ResNet, and DenseNet through transfer learning.

Segmentation techniques like U-Net and SegNet are crucial for lesion localization. Model performance is evaluated using metrics like accuracy, precision, recall, F1-score, and AUC-ROC, often validated through k-fold cross-validation. Development environments like Jupyter Notebook and PyCharm facilitate model development, while deployment is managed using tools like TensorFlow. Additionally, cloud services such as Google Colab provide scalable resources for model training and deployment

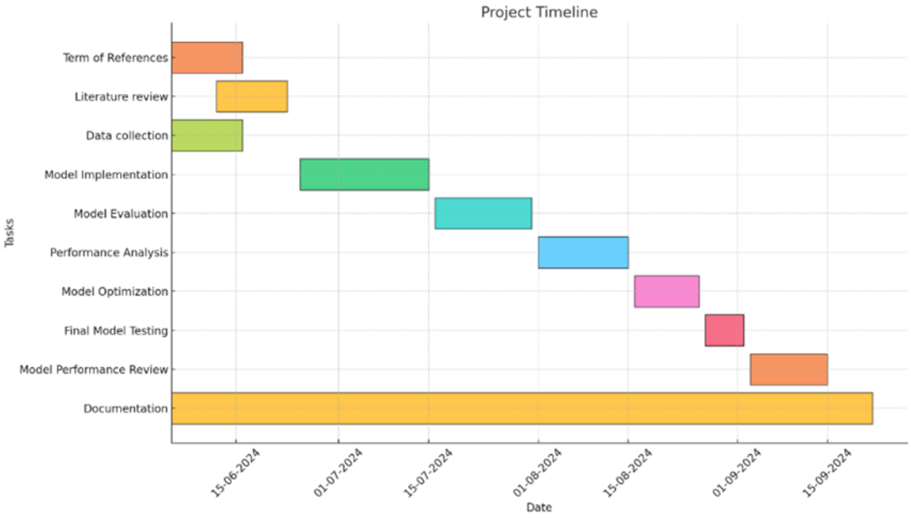


Figure 1.1: Project Pipeline Gantt chart

Chapter 2

Literature Review

2.1 Introduction

Before moving into the corpus collection process and the actual experimentation of this project, it was necessary to dedicate some pages to discuss in the literature the techniques utilized for this investigation. This was a crucial because it would assist in revealing the ideas that have been discussed before and give a keen list of the set of options that are good for useful experimenting. The other role done by the literature review was to present some of the ideas that were put into practice and tested that closely related to the question asked by this project.

The following is the list of many different subjects, containing information collected with the aim to shed light on this project. These were:

- Skin Lesion Analysis
- Image Processing Techniques
- Machine Learning Techniques
- Deep Learning for Skin Lesion Classification
- Transfer Learning
- Ensemble Models
- Evaluation techniques
- Challenges and Future Directions

It is concluded that the datasets used for skin lesion classification are useful for improving the effectiveness of diagnostic technologies.

2.2 Skin Lesion Analysis

Binary classification of skin lesion is one of the significant tasks in the dermatology area where dermatologists try to differentiate between the benign (non-neoplastic) and malignant (neoplastic) lesion. This process is essential for confirming early diagnosis of skin cancers most of which are melanomas as they are the deadliest. It greatly affects patients' prognosis; the five-year survival rate of melanoma recognized in stage I is about 98%, it becomes only 16% for stage IV. It also helps in turning more attention in planning resources so that premature biopsies and nonessential treatments do not happen, it only concentrates on the high-risk patients that require early attention (Esteva et al. 2017a). This besides enhancing the quality of patient care also helps in directly containing the skyrocketing costs of health care delivery.

The results are promising and they serve as an indication that with the new deep learning and machine learning the efficiency of skin lesion classification has significantly improved. As Deep Convolutional Neural Networks (DCNNs) are capable of providing high accuracy in analyzing dermoscopic images, which in many cases surpasses traditional diagnostic methods (Yu et al. 2023). Over time, these models are trained and developed on very large databases such that they can 'see' finer differences between a benign and malignant lesion.

Implementation of these AI based tools in the clinical practice is a great leap in the diagnosis and management of skin cancer, a development which holds the promise of improving patients' lot and enhance the delivery of healthcare services.

2.3 Image Processing Techniques

Techniques of image analysis are central when it comes to classification of skin lesions, this ensures that the diagnostics involved are accurate and dependable. This review delves into three critical aspects: They considered application of techniques, namely image preprocessing, feature extraction, and data augmentation which jointly can influence the effectiveness of algorithms for binary classification in dermatology.

2.3.1 Image Preprocessing

Image preprocessing is one of the most essential steps in the analysis of skin lesion images. It encompasses several processes whose objective is to enhance the image quality in order to facilitate subsequent procedures. It pre-processing methods for image include noise elimination, image contrast, and image normalization. Techniques such as Gaussian filtering as well as Median filtering are used in eradicating dispelled noises to enhance the image clarity (Gonzalez and Woods 2018). Traditional image processing procedures such as contrast enhancement that involves the use of histogram equalization to increase the visibility of lesions are used because of fear that some features of the lesions may be overlooked. Normalization standardizes the image values and is very important in bringing into scale images within datasets

2.3.2 Feature Extraction

Feature extraction is one of the primordial stages of image processing, where various attributes of skin lesion images are detected and measured. Features can be classified as colour patterns, texture, shape and the type of border that may be used. Shape features, like the mean of color channels, the standard deviation of color channels, are related to the pigmentation of lesions (Barata, M. Emre Celebi, and Marques 2014). Shape features, based on the developments of the new technologies such as LBP and Gabor filters, specify the spatial relationship of the pixel intensity and are beneficial in enrollment of benign and malignant lesion (Tang et al. 2018). A three shape morphological aspect of borders irregularity, asymmetry and compactness provides information of geometric characteristic of lesions and useful in distinguishing melanoma from non-melanama (Xie et al. 2017). Also, border features assess the margin of lesions so as to determine abnormalities that are characteristic of malignancy (M. Emre Celebi, Kingravi, and Uddin 2007)

2.3.3 Data Augmentation Techniques

This process is used in increasing the both quantity and variety of the training images as a way of handling the problem of limited datasets in medical imaging. Rotations, scaling, translation and flipping among others are usually employed in augmentation operations that create new images from existing ones and as a result enhancing the resilience of classification models. Some other sophisticated data augmentation techniques are elastic distortions and random erasing that do the enhancement of the variability of data for training purpose to decrease the chance of overfitting to unseen data (Simard, Steinkraus, and Platt 2003). Regarding the use of the generative adversarial networks (GANs) for synthetic data generation that has been successfully attempted in the creation of skin lesion images for better deep learning models training (Frid-Adar et al. 2018).

2.4 Machine Learning Techniques

This review covers several essential machine learning methodologies: supervised learning, unsupervised learning, decision trees, support vector machines and neural networks which are an individual input to the area of dermatological image analysis.

2.4.1 Supervised Learning

In supervised learning the model is trained on a training set, in which the training vectors, as well as the training classes, are well defined. Deep convolutional neural networks (CNNs) were employed for the skin cancer prediction as performed in a supervised manner. To perform the classification, it was trained with a dataset of more than 129,000 clinical images and the test model was proven to be near-dermatologist level with the AUC at 0. 96 (Esteva et al. 2017a). Likewise, (Litjens et al. 2017) presented a wide-range systematically organize survey on deep learning in medical image analysis, stressed that supervised learning methods were usually effective in several applications such as skin lesion classification.

2.4.2 Unsupervised Learning

Unsupervised learning methods are applied to find out unknown relationships and structures of the data that do not have any labels. Breast Cancer for instance, has been detected using whole-slide images through applying unsupervised learning with convolutional neural networks by (Cruz-Roa et al. 2014). Despite the fact that their respective focus was set on breast cancer detection, their approach is easily transferred to the skin lesion analysis context. They obtained the throughput sampling rate and further showed the effectiveness of the unsupervised learning to the medical image data analysis. Another related study by (Wang et al. 2018), who used the unsupervised deep learning techniques in the field of lung cancer with the feature extraction which can be concluded that this method can be used in medical images' classification.

2.4.3 Decision Trees

Decision trees are widely used because of their simplicity and understanding their results of the classification problem; therefore, they are useful in solving the medical image analysis, especially skin lesion classification. The first decision tree frameworks were established using Quinlan's ID3 algorithm proposed in 1986 with a top-down greedy approach to partitioning the dataset regarding the attribute that offered the biggest information entropy. This method is particularly applied in analyzing how clinical decisions are made when discriminating between benign or malignant skin lesions.. (J. R. Quinlan 1986)

Decision tree based methods were used to predict MGMT methylation status in glioblastomas based on MRI texture analysis to an accuracy of 85% . This confirms the competence of decision trees in medical imaging such as the analysis of skin lesions. Texture features obtained from skin lesion image databases can be classified using decision trees and the results are able to distinguish between benign and malignant cases. (Korfiatis et al. 2017)

However, Quinlan introduced C4. 5 in 1993 which is the enhancement of ID3; C4. 5 considers categorical and numerical attributes and deals with problems such as missing values (J. Quinlan 1993). This is especially true for skin lesion classification since the features (e. g. , color, texture, border irregularity) can be quite divers.

Random forest is an advanced version of the decision tree algorithm developed by Breiman in 2001 that is based on the creation of a multitude of classification trees and the provision of an overall decision through voting (Breiman 2001). Random Forest is particularly beneficial when decrease overfit and increase model stability which is quite important for classification of skin lesions.

When it comes to binary classification of skin lesions, decision trees method and its symbiotic extension, called Random Forests, can enhance the exploration of features obtained from dermoscopic images. For example, (Barata, M. Emre Celebi, and Marques 2014) introduced the method based on decision trees for classification of skin lesions using color and texture features; the same method proved its efficiency in classification of skin lesions as benign and malignant.

2.5 Neural Networks

Neural networks and specifically deep learning models are widely used for medical image analysis because they have high accuracy and the ability to learn high-level patterns. In the case of skin lesion classification they have demonstrated great promise in differentiating between the benign and malignant ones on the skin.

The theoretical basis for computational models derived from neural activity was first described by McCulloch and Pitts in 1943 on artificial neural networks or ANNs. But, the significant step in the development of the neural networks for image analysis was taken by LeCun et al. in 1998 with the Convolutional Neural Networks (CNNs). CNNs were initially developed to solve primarily grid-like data like images and they are quite efficient in solving these (LeCun et al. 1998)

The following are the comprehensive analysis of the different facets of neural networks and their utility in skin lesion binary classification.

2.5.1 Deep Learning for Skin Lesion Classification

Hence, the concept of deep learning is based on the neural networks having multiple layers, or deep neural networks (DNNs). Recent developments have made them more versatile by virtue of powerful computers as well as widespread and vast datasets (Goodfellow, Bengio, and Courville 2016) The principal constituents of deep learning models consist of layers, activation functions, loss functions, and optimization algorithms, all of which help the model knowing and often make forecasts on intricate data.

When applied in skin lesion classification Deep Learning has proven to be very useful. They are even more suitable for this purpose due to their ability of dealing with high-dimensional data and learning hierarchical feature representation (Esteva et al. 2017a). This is important in MRI and CT scans where the distinction of between malignant and benign tumors might be very thin and therefore needs better models for detection.

Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are currently a go-to solution for image analysis in deep learning, because the algorithm takes raw pixel data and learns different layers of abstraction on its own. The authors of the study by (Esteva et al. 2017a) proved that the CNN can correctly distinguish the skin lesions with the accuracy comparable to the professional dermatologists. Deriving ideas from the existing deep learning works, the network was trained on a large dataset of dermoscopic images where the authors successfully reached a high level accuracy in the task of binary classification of benign and malignant skin lesion. This study showed that CNNs can help clinicians in diagnosing skin cancer, nonetheless, such a system was dismissed for needing huge annotated data and ample processing power(Esteva et al. 2017a)

It is noted that numerous works have shown that CNNs can be useful when classifying skin lesions. The authors Esteva et al. , in their study, applied CNNs for the diagnosis of skin diseases with classifications' accuracy equivalent to that of dermatologists. Designed on a large dataset of more than 129,000 of the clinical images of the diseases, their model encompasses approximately 2,000 diseases. CNN was as accurate as 21 certified dermatologists, thus revealing its applicability in clinical practices (Esteva et al. 2017a). Deep learning framework using CNNs designed to differentiate dermoscopic images of skin lesions into seven classes, one of them being melanoma and the other benign nevi. Their model obtained the overall accuracy of 86.9%, superior to traditional machine learning corresponding models (Han et al. 2018). Classification performance enhancement was also touched upon by (Codella et al. 2018) who suggested an ensemble of CNN for skin lesion analysis. The presented approach showed the worst error rate on several public collections; moreover, it proved that models based on CNN outperform the state of the art. However, the CNNs specifically for the skin lesion classification come across several problems. Supervising CNNs entails huge amounts of annotated data, and in the medical field, the acquisition of images as well as annotation is a tiresome and costly affair that could call for the expertise of a doctor. Still, the access to large and high-quality databases still remains an issue (Esteva et al. 2017a).

Moreover, CNNs are computationally expensive and need a lot of hardware: powerful GPUs, and large memory, which can be a problem for many small institutions still (Han et al. 2018). Deep networks also suffer the problem of overfitting whenever the training data is scanty. Methods including data augmentation, dropout and L2-regularization are usually used to tackle this problem, although these processes often require a lot of attention in order to be fine-tuned and validated in order to get models that generalize good enough (Srivastava et al. 2014a).

2.5.2 Comparative Analysis of CNN Architectures for Skin Lesion Classification

Some of the most well known architectures used in CNN have been used in skin lesion classification because of their effectiveness and capacity to capture feature from the images. Some of these architectures are ResNet, VGG and Inception among them being impressive architectures for a given neural network.

ResNet (Residual Networks)

ResNet further presented to employ the concept of residual connections to enable training of very deep networks without experiencing vanishing gradient. This architecture has been very efficient for many image classification tasks and skin lesion classification inclusive. Due to this, ResNet is able to train network with hundreds layers, hence it is able to learn the complex features needed in the classification of skin lesions. Among the new structures in ResNet, identity shortcut connections are used, which ‘jump over’ one or several layers. These shortcuts help to handle the degradation problem in which training error increases when the number of layers of a deep neural network is increased. Due to its capacity to train very deep networks, ResNet has emerged as one of the go-to solutions in many image analysis operations. (He et al. 2016). According to a study which revealed that using ResNet-101 made the classification performance for melanoma in the ISIC 2017 dataset better because multiple layers yielded hierarchical features. (Li Yu et al. 2017).

Furthermore, the existing research has shown that ResNet is equally good for the medical image classification. For example, a ResNet model for skin lesions classification showed that this improved its accuracy than the networks that are shallower. Due to its large architecture ResNet was able to learn high level features and learn the various and subtle differences and alterations in skin lesions making it an efficient tool for dermatological image analysis (Zhang et al. 2019b).

VGG (Visual Geometry Group Network)

VGG is an architecture standard issued from the Visual Geometry Group with three different models: VGG11, VGG16, VGG19. VGG was introduced by Simonyan and Zisserman in 2014 and is characterized by its simple and structure-oblivisent design. It has 16 to 19 layers, with small 3 x 3 convolutional filter and was extensively used for its simplicity and good result. In skin lesion classification, VGG has been effective because of the detail captured in its deep and simplicity of its architecture making it straight forward. Within the framework of the architecture and network topology, VGG's design is incredibly simple and efficient; it has a small Receptive Field, as well as it is built up from deep stacks of the convolutional layers. (Simonyan and Zisserman 2014).

Using VGG resulted in high accuracy and some of the best performances when applied in conjunction with transfer learning methodologies and particularly so when such methodologies incorporated the weights learnt from databases such as ImageNet. It largely improved the model's accuracy of the expert and narrow field of application, namely skin lesion classification (Menegola et al. 2017).

Inception (GoogLeNet and Inception v4)

GoogLeNet as nicknamed Inception v1 is the base model with inception architecture that aimed at costing less while delivering better result. Inception architecture brings into use network-in-network and includes the multiple filter sizes such as 1x1, 3x3, 5x5 and so on and thus establishes multiple layers within layers so as to formulate the multiple scale features all at once. On the second place, Inception v4 is somewhat better than its previous version but its structure is based on Inception modules with residual connections: It is even deeper and is more powerful in terms of feature extraction and training stability. This fact allows the model to process the visual information better, which makes it suitable for complex tasks, such as classification of skin lesions (Szegedy, W. Liu, et al. 2015). Inception-v3 is the modified version of the inception architecture and it includes improvements over inception-v1 including the factorized convolution and well-developed methods for decreasing the grid size. It also includes batch normalization and label smoothing which make the model better and increase its training rate. Inception v3 is used for skin lesion classification because besides being less computationally intensive as compared to the other models it is effective at discovering tiny features in medical images. (Szegedy, Vanhoucke, et al. 2016).

Inception v4 is the latest version of Inception structure and it comes with some improvements compared to the previous version as related to the efficiency of the network. Inception v4 is created by integrating inception modules of previous Inception versions and Residual Networks. This combination leverages the strengths of both architectures: such as the multi-scale feature extraction of Inception and the improved training stability of residual connections. There are certain enhancements of the structure of the inception v4 which make it more structured and less complex in order to improve on the accuracy of this model. (Szegedy, Ioffe, et al. 2016).

Inception v4 is suitable in the skin lesion classification since it was designed to provide more detailed features in the images. Inception v4 is thus very useful in providing a reliable structure when it comes to feature extraction from wildly different scales of the anatomy and overall reassurance that deep networks are well trained to differentiate between benign and malignant lesions.

AlexNet

AlexNet is one of the oldest networks in the high ranks of deep learning and has improved the performance of image recognition systems for a long time, designed by Krizhevsky, Sutskever, and Hinton (2012). It consists of eight layers: The network architecture used was five convolutional layers and passed through three fully connected layers. The architecture also has ReLU activation functions for non-linearity and max pooling layers for reducing dimensionality of the spatial features The last layers also contain dropout layers to help avoid overfitting. AlexNet provided a breakthrough for deep learning and Convolutional Neural Networks (CNNs) by winning the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in the year 2012 within a top-5 error rate of 15.3% from a picture, which was far better than the earlier accuracy of 25% established by the rivals. (Krizhevsky, Sutskever, and Hinton 2012).

Exploiting the AlexNet model trained once more on dermoscopic images gave them a classification performance higher from the traditional approaches of machine learning. The study achieved increased accuracy in classifying benign and malignant lesion, thereby, demonstrating the applicability of AlexNet to dermatologist's diagnostic endeavours (Kawahara, BenTaieb, and Hamarneh 2016).

Similar to other deep models like commonly used deep learning models, AlexNet is hard to analyze internally or might be referred to as a “black box”. Interpretability of AlexNet is also an active area of research, as clinicians require transparency to validate and trust the model's decisions (Ribeiro, Singh, and Guestrin 2016).

EfficientNet and DenseNet(Densely Connected Networks)

EfficientNet was proposed by Tan and Le in 2019, and proposed the idea of scaling the networks by the compounds of depth, width, and resolution. For example EfficientNet's models (EfficientNet – B0 through to B7) has been shown to outperform well-known CNN architectures like, ResNet, in various medical image applications. Liu et al. (2020) have employed it for skin lesion classification showing that it outperforms ResNet in terms of accuracy and model parameters which makes it very suitable for mobile and other low-resource scenarios. (Y. Liu et al. 2020). Another excellent family of deep models of medical images is DenseNet (Densely Connected Networks) also proposed by authors Huang et al(2017). The main differences are that the DenseNet has a Feed Forward connection between each layer and every other layer in the network which we reused features and many parameters that were needed were comparatively small than ResNet.DenseNet is more superior to ResNet especially when trained with limited data hence the use of feature reuse mechanism to enhance the flow of information. Namely it performs well in training high level features that are necessary for the differentiation in the top-down approach between microanatomical patterns of dermoscopy images of benign and malignant neoplasms.A study mentioned skin classification using DenseNet, this showcased that this network boosted the classification results over ResNet and VGG due to reusability of the features. (Zhang et al. 2019a). .

2.5.3 Transfer Learning

Pre-trained Models (ImageNet)

Transfer learning has been quite useful especially in the classification of medical images since annotated datasets are scarce. Therefore, through transfer learning, these models are pretrained on such data repositories including ImageNet that contains more than 14 million images as well as 1000 categories, and they can be fine-tuned for diversity tasks including skin lesion classification. It applies another parameter that utilizes the knowledge the model acquires from a large number of diverse images and from big datasets, and our experiments have shown that it improves the results in contexts concerning specific medical areas with fewer annotated samples.

While pre-trained networks and retrained them on dermoscopic image obtaining high classification rate are used to enhances the model performance by about 18% by adjusting the pre-trained weights to the configuration of dermoscopic images. Since transfer learning is done from the ImageNet, generalization is better and the features learnt from this smaller database is also relevant making the transfer learning strategy ideal for medical image classification (Menegola et al. 2017).

Fine-tuning vs. Feature Extraction

Fine-tuning and feature extraction are two primary methods of transfer learning

Fine-tuning

The final level of transfer learning is referred to as fine-tuning whereby new weights are used in retraining of the network to operate on the new data set. It usually involves retraining the whole network or the deeper layers of the network at a reduced learning rate. Therefore, fine-tuning enables the model to fit even closer to the new task in order to discover patterns and features from the target set. This method often produces higher performance improvement especially when its dataset has enough samples that can be labelled and adjust the model's weight. It can, therefore, assist the network in adjusting the parameters for the particularities of new data which is the key to enhancing the model efficiency and its capabilities of generalization in medical images classification (Tajbakhsh et al. 2016).

Feature Extraction

While, Feature extraction translates the features of the new dataset and passes them through the pre-trained network and then feed them to a new classifier of the choice like logistic regression or support vector machine. In this method, the weights of the pre-trained network are fixed and only the new classifier is trained with the extracted features only. Feature extraction is computationally simpler because many fewer weights must be updated. This approach a highly useful where the size of the target dataset is small because it only employs the representations of information derived from the features from a large scale pre-training without extensive retraining (Huang et al. 2017).

2.5.4 Ensemble Models

Thus, ensemble models create several learning algorithms to enhance the classification and generate better models. Ensemble methods are used in an attempt to bring together the results of different models that together can be stronger and produce fewer errors than the separate components. The above approach is more advantageous especially when dealing with functions such as skin lesion classification, where different models can learn different aspects of the given data.

Bagging (Bootstrap Aggregating):

The technique known as bagging is the process by which several instances of a single algorithm are created with the purpose of being trained using the different subsets of the training data. These subsets are generated by applying the technique known as random sampling with replacement. The last forecast is generated from the average of the quantitative predictions for regression or percentage for classification starting from all the models. Random forest is one of the most widely used methods in bagging, which entails growing many decision trees and combining their outputs in a way that increases the model's accuracy and reduces overfitting. (Breiman 2001).

Boosting

It sequentially builds models whereby the next model is trained to correct the mistakes of the former model. The final output is produced either by using the weighted majority vote or summing up the predictions. Some of the boosting algorithms are AdaBoost and Gradient Boosting. Such techniques can reduce Bias and Variance and hence result in the improvement of various performances such as classification of Medical Images (Freund and Schapire 1995).

Stacking It is a technique in which multiple base learners are first trained and then the outcomes from all the base learners are combined using a meta-learner. The base models are applied to the original dataset and their predictions are the input features for the meta-learner which often can be a simple model as logistic regression. Stacking uses the strengths of various models and more often than not is more accurate than each of the models alone (Wolpert 1992).

The elaborate architecture together with an integration of different architectures in the ensemble methodology enables the best results of the classification of benign and malignant lesions in several public datasets. The analysis showed that the overall accuracy of the proposed ensemble model was higher than that of individual model; this we ascribe to the utilization of the various CNNs (Codella et al. 2018). This study also proved that use of ensemble with CNN could predict skin cancer to the level of dermatologists. The described ensemble approach enhanced the model's stability and raised the predictability and reliability of the outcomes, making the method beneficial in clinical practice when accurate diagnosis is required (Esteva et al. 2017b).

2.5.5 Challenges and Future Directions

Limited High-Quality Annotated Data:

At the present time, dermatoscopy image databases of sufficient size and quality are insufficient, and are often unmarked, thus restricting the opportunity to train and test deep models. This aspect becomes a big limitation to the performance of models and even their accuracy, when solving different sets of data. However, it has been recommended to use generative adversarial networks (GANs) to tackle this problem of augmenting the meager databases; however, the problem is still regarded as very difficult (Esteva et al. 2017b).

Generalizability Across Diverse Populations:

Still, the issue of how to prevent the degradation of the model's performance over learned subplets is unsolved as patients are distinctive and may have different skin color, age, and living area. These variations can therefore alter the appearance of the lesions which is not desired when it comes to generalization of the model. It is especially beneficial, if possible, to have a diverse set of patient's characteristic similar to real-life problems; for achieving the above, there must be considerations to enhance the models in health care industry (Brinker et al. 2019).

Overfitting:

The problem is even more pronounced in complex models such as AlexNet and Inception V4; more emphasis is placed on learning the input data's characteristics to the extent of over-learning and therefore, they perform poorly when the models are tested on other data sets. Preventing techniques that are used for overfitting include the following; Dropout, Regularization and cross validation and they are crucial if used appropriately although hard to set since it is a balance between complexity and performance (Srivastava et al. 2014b).

Feature Extraction:

It is also evident that the current status of feature extraction is inadequate to perform the discriminating analysis at the level of anatomic differentiation of the benign and malignant lesions. Another technique that is current is the Local Binary Patterns (LBP) and filters including Gabor filters have been employed, but what is however desirable as presented is their optimization particularly for the classification output and the variation in the discriminant of two lesions that have similar appearance (Barata, M. Emre Celebi, and Marques 2014).

Dataset Imbalance

In medical image databases, a large number of benign cases are often observed, which is why there is a data imbalance. This can make the model favor the majority class (benign), consequently, it will not be very sensitive to the second class (malignant). This means that the number of images of certain categories dominates the number of images of other groups, and methods like Focal Loss or sampling the minority class are required to solve such a problem (Esteva et al. 2017b).

Model Interpretability

CNNs are one of the most popular deep learning models; however, understanding how they make decisions is challenging, which earned them the “black box” nickname. This lack of transparency is a big issue in medical applications as the relationships between input and output (classifications, benignant/malignant) need to be trusted and explainable (Ribeiro, Singh, and Guestrin 2016).

The next chapter details the dataset used in this study and describes the structure thereof, describes preprocessing steps taken on it, and the challenges associated with the class imbalance of this dataset given our literature and model understanding.

Chapter 3

Dataset

3.1 Introduction

This chapter gives general information on the dataset employed in this work with a focus on the binary classification of skin lesions into malignant (melanoma) and benign forms. The dataset is used in the development and assessment of deep learning models used in diagnosis of skin cancer, particularly melanoma which is very dangerous skin cancer if not diagnosed in good time.

The chapter is divided into two sections based on the research activities of the two different faculty positions outlined in the previous chapter. First of all, the Section on the Dataset Structure is reviewed that explains origins, volume and division of the training, validation, and testing datasets. This section also discusses issues that arise due to class imbalance, which is a prevalent problem in medical datasets, and the strategies of having balanced classes.

The second part covers the preprocessing that occurs to the images before they are reached the deep learning models. Such preprocessing techniques are image resizing and normalization, and other data augmentation techniques like rotation, flipping, cropping, and others to mention so as to expand the type of training samples.

3.2 Dataset Structure:

The given dataset collects dermatoscopic images derived from standard dermatological image databases like the ISIC (International Skin Imaging Collaboration) challenges. These pictures are of different patients with different ages, skin color and characteristics of the lesions, that is why the given dataset can be considered appropriate to train models for clinical practice.

The dataset is structured into three subsets,

- **Train Set (train_balanced)**: For the training set, it possess of total of 7848 images. Where in the dataset a distribution similar to the Malignant-Benign image where there is 3924 images in Malignant class and another 3924 in the Benign class. This distribution is very crucial with a view to helping the model to avoid identifying either the positive or the negative class. The images are normally about 224 by 224 pixels but, it could be slightly off and are resized closer to the preferred sizes during the preprocessing. All the images are saved in the JPEG file format because JPEG is widely used for saving color images and each image is in the RGB color several.

-

- **Validation Set (val_balanced):**The validation set also consists of the similar number of samples as the training set and it includes 1962 samples whereby each class has 981 samples. This set is also useful in tuning of hyperparameters and while selecting a model since one can develop a check point on how the selected model is performing with the unseen data. The size, and format of the images to be used are also similar to the images used in the training set this is in a bid to avoid discrepancies between the results of the training phase and the results of the validation phase.

-

- **Test Set (test_ISIC_2017):** As for the test set 117 for oth 483 for mel images of 224 x 224 pixels are used which are extracted from the ISIC 2017. This is the list by which the final model can be tested and statistical indices such as the accuracy, sensitivity or specificity can be assessed. However the test set contains images with the same format and size as the one used for training and therefore the results are qualitative and can therefore be used with ease to deduce the best configuration

3.2.1 Class Distribution and Imbalance

Imbalance of class is a general problem in many medical datasets, where the number of negative patterns is significantly larger than the positive ones. The images in the training and validation dataset were sub-sampled to be balanced where the number of images for each class was the same. This is attained by a proper selection of images from the denoted ISIC Archive and increasing the amount of augmentation concerning the model's data set that enables to train the model under a real-world oriented simulation.

This balance is very important to avoid the creation of a bias toward the majority class, hence the sensitivity to identify the rare malignant lesions is minimized. To expand on this issue, techniques such as focal loss or using class-weighted losses were contemplated and incorporated if needed.

3.3 Data Preprocessing:

In order to clean the dataset and make it ready for training the model, several transformations were made on the downloaded dataset in order to make it easier for the model to learn properly from the data.

Image Resizing and Normalization

All the images in the dataset were preprocessed where all the images were resized to a dimensions of 224 by 224 pixels. It is very important to standardize the dimensions of the input data feed to each of the specified CNN architectures like AlexNet and Inception V3 in this study. The structure of the image plays a critical role in the generation of the feature data by the consistent image sizes so that the network may learn throughout the dataset.

Pixel values were also normalized so as to bring their values in the range of 0 and 1. Convergence is particularly beneficial during the learning phase because it makes the process more standardized, thus allowing for the stabilisation of the learning process. Thus, the model does not pay much attention to variations of pixel intensity, thus targeting only important features that distinguish between benign and malignant neoplasms.

Data Augmentation

Thus, data augmentation was used to enhance the range and versatility of the approach under consideration. These techniques included:

Rotation: Lesions were randomly rotated in the images by up to an angle of 20 degrees which gives different orientation of lesion. This is preferable for the model because the lesions' orientations may change in practice, and this retains the model's orientation invariance.

Flipping: The directions of horizontal and vertical flips were also applied to equalize the orientation of the lesions which may differ naturally. This technique makes it possible to replenish the training set twice as large, thus expanding the model's exposure to different examples.

Figure 3.1 demonstrates the application of various data augmentation techniques, including rotation, shifting, zooming, and flipping. The original skin lesion is shown on the left, while the other images display augmented versions.

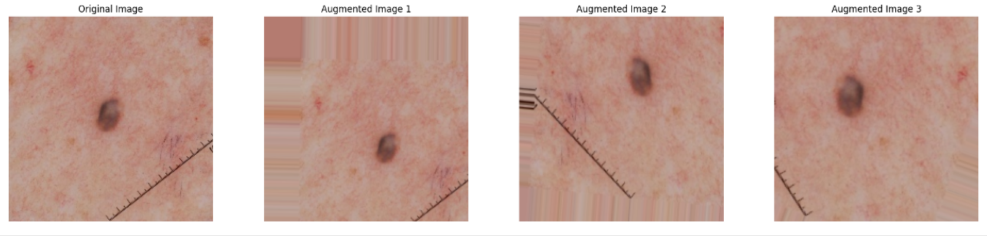


Figure 3.1: Augmented Skin Lesion Images

Scaling and Cropping: Normal scaling and cropping techniques were used in order to imitate different zoom factors and the focus of lesions in some particular areas of the images in order to help the model generalize the problem, because it has to differentiate the lesions at all scales and positions within the picture. This is especially pertinent in the medical imaging where the region of interest is usually small and may be appearing at different positions in each image.

Handling Class Imbalance

As the dataset was thoroughly selected to be almost equally divided by the classes, extra precautions were taken to prevent model bias to specific class. To speculate with the rest of the imbalance, Focal loss and class-weighted loss functions were employed. These methods assist in maintaining the sensitivity of the model with regards to the benign and malignant cases, hence increasing the overall sensitivity of diagnosis.

The methods used for training and evaluation of models including optimization of hyperparameters and the implementation of focal loss, in order to improve the ability of the model to classify skin lesions are discussed in the next chapter.

Chapter 4

Methods

4.1 Introduction

This chapter provides the details of technical aspects related to the classification in skin lesion images based on deep learning models. Concerning the particular architectures, the main work explains the configurations of InceptionV3, AlexNet, DenseNet, EfficientNet, ResNet and an ensemble model constructed to differentiate between benign and malignant tumors. Based on the introduced data preprocessing techniques, the chapters formats approaches to the architecture modification, optimization techniques, and the use of evaluation indicators to consider for improved model performance.The methodology flowchart for the approach used is shown in Fig. 4.1.

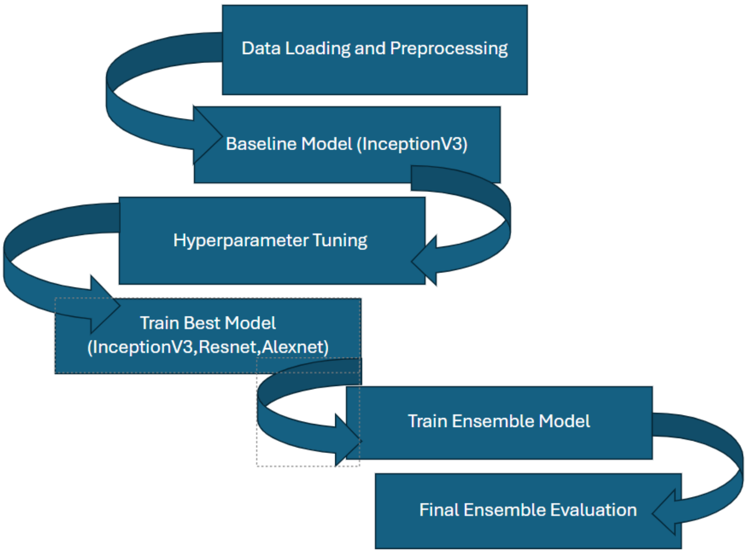

Data Loading and Preprocessing
Baseline Model (InceptionV3)
Hyperparameter Tuning
Train Best Model (InceptionV3,Resnet,Alexnet)
Train Ensemble Model
Final Ensemble Evaluation

Figure 4.1: Methodology Flowchart

4.2 Model Selection

In the utility of classifying skin lesions into benign and malignant categories, several models were used(ALexNet,InceptionV3,EfficientNet,ResNet,DenseNet). Specifically, each model was chosen according to the been proved efficiency in image classification as stated in the Chapter 2.

4.2.1 InceptionV3:

InceptionV3 architecture was chosen for basline model because it was pre-trained on the ImageNet dataset and was able to capture image features at multi-scales, which can be especially beneficial for classifying skin lesions, where the shape and size of lesions can widely differ. To enable further fine tuning on the binary classification task,the top layers of the classification structure were not utilized in this task. To this network as a part of the experiment, a GAP layer was incorporated to bring the dimensionality of the feature maps down; the final classification layer chosen was a Dense layer with sigmoid activation to produce the likelihood of a lesion being malignant or benign as recommended in the Literature Review section of this document.

Model Architecture

The InceptionV3 model was loaded while excluding the final layers of the model because of the decision to replace them with the layers that were optimized for the binary classification. The model was set initially to have an input dimension of 224, 224, 3 as the images were preprocessed to these dimensions and the pre-trained layers were preserved for the learned feature maps.

Global Average Pooling (GAP) Layer: Rather than employing fully connected layers right after convolutional sections, a Global Average Pooling layer was added. The GAP layer then averages the spatial dimensions of each feature map, effectively transforming a whole feature map into a single value, which furnishes a neater and easier interpretation of the feature vector. It also reduces the total number of parameters in the model which also helps it reduce overfitting.

Dense Layer: To make the feature extraction non-linear and post process the copious features from the GAP layer, Dense layer with units 2048 and ReLU activation function were included. This fully connected layer also gives the model the required capacity to make good predictions as the model is able to make sense of different patterns of data.

Dropout Regularization: A Dropout layer was applied with a dropout rate of 0.5 to minimize overfitting since during training, a certain fraction of neurons is set to zero randomly. This type of regularization actually constrains the apply of the model to overly depend on a specific neuron/feature, thereby making it learn more general features.

4.2.2 AlexNet:

From state-of-art convolutional neural networks, AlexNet which is proposed by Krizhevsky et al. (2012) was chosen due to its plain structures and efficient computation. , it was describe in 2.5.2 of the Literature Review that AlexNet architecture is appropriate for image classification problems where reduced computational cost is desirable. Just like the previous model, it has a series of convolutional and max pooling layer that is efficient in spatial feature extraction making it too ready for use in this research study. Specifically to skin lesion classification where the goal of the classification is to distinguish between benign and malignant skin lesions, modifications were made to AlexNet.

Model Architecture

In this study, AlexNet has been employed as a model of image categorization and consists of several layers aimed at extracting important features from the input images as classified into two categories.

Convolutional Layers:

Firstfully connected layer that uses 96 filters with large kernel size 11X11, and a stride of 4 to make the depth of the input image small and get low-level feature maps. This is succeeded by a max-pooling layer in order to down sample feature maps while adding concern for spatial locations. Further, fully convolutional layers with 256 filters and kernel size of 5 x 5 followed by 384 filters and 3 x 3 kernel size are implemented for deeper and abstract features in the input image.

Max and Fully Connected Layers

The operation of max-pooling is performed after particular convolutional layers, to decrease the parameters of a convolutional layer and the feature maps. In order to reduce its dimension the final output-passed through two further fully connected layers with 4096 characteristics, each of which has a ReLU activation function.

Output Layer:

The last layer is a Dense layer with one neuron and sigmoid activation function that returns probability that particular input lesion is malignant.

4.2.3 ResNet Model

In this section,a ResNet50 model is implemented, fine tuned for binary classification under hyperparameter tuning and a custom focal loss function that is robust to class imbalance.Residual connections were chosen due to their ability in training deep networks without vanishing gradient problem. These connections permit information to be able to by pass some layers so that deeper networks can be trained effectively. For medical imaging, which features can often be subtle, require deep layers to be extracted, ResNet's deep architecture is key. There is good reason to include this in this project, as its proven performance on a large range of image classification tasks predominantly relating to subtle pattern discovery.

Model Architecture

ResNet50 model is used to provide the backbone of the feature extraction. The portions of ResNet50 downstream from the top layers are discarded and custom layers are appended to convert it to work for binary classification of skin lesions (benign vs. malignant). Where ResNet50 has generated the feature maps, Global Average Pooling layer is applied to shrank dimensionality. Then on top of this, a Fully connected Dense layer with the amount of units between 512 and 2048 tuned is added.A final output layer (single neuron with sigmoid activation), giving a probability in binary classification is used.

Model compilation

The ResNet50 model is compiled using the Adam optimizer with a learning rate of 0.0001. This optimizer is chosen for its adaptive learning rate capabilities, which contribute to faster convergence and improved stability during training. To address class imbalance, the binary cross-entropy loss function is employed, as it is well-suited for binary classification tasks.

4.2.4 DenseNet Model

Due to its innovative architecture, the architecture of DenseNet121 used in this project, in which each layer is densely connected to every other layer. Such kind of structure allows for feature reuse which is also parameter and gradient flow efficient. More importantly, DenseNet excels at identifying intricate details through its reuse of features across layers for which capturing the difference between benign and malignant skin lesions requires. This task being one of precision and computational efficiency make its compactness and efficiency a perfect fit.

Model Architecture

We pretrain the DenseNet121 model as ResNet50 model are used. It allows for feature reuse from all previous layers, and each layer gets input from all previous layers. To reduce dimensionality, the top classification layers were removed and replace the last ones with a Global Average Pooling layer. A Dense layer of width 1024 and activation ReLU is added to process the features following this. To try preventing overfitting, a Dropout layer with a rate of 0.5 is used. The last layer in the form of a single neuron with the sigmoid activation has been taken, which will classify the lesion as benign or malignant.

Model Compilation

The model is then compiled with the Adam optimizer learning rate = 0.0001. The loss function is binary cross entropy which can be used for binary classification. The evaluation metric is accuracy and is how well the model predicts if a lesion is or isn't classed as malignant.

4.2.5 EfficientNetB0 Model

EfficientNetB0 model, pretrained on the ImageNet dataset for binary classification of skin lesions (benign vs malignant) is used. The highly optimized balance for the accuracy and computational efficiency was EfficientNetB0. EfficientNet scales depth, width, and resolution uniformly, and this not only improves the accuracy, but also helps reduces the number of parameters required overall, which is especially appealing for medical image classification. It allows the model to achieve state of the art performance without sacrificing efficiency.

Model Architecture

For EfficientNetB0 to be adapted to binary classification, the top classification layers are removed and the feature maps extracted are reduced using a Global Average Pooling layer to obtain a lower dimensional output.

The extracted features are then joined with a Dense layer of 1024 units and ReLU activation to introduce non linearity and further process the extracted features. In order to avoid overfitting, we insert a Dropout layer with rate 0.5 that randomly ignores neurons in the training and a RATE of 0.5 in the testing. Finally input layer with 30 neurons sigmoid activation to ensure zero mean and unit variance, output layer has a single neuron sigmoid activation, and generates a probability score between 0 and 1 to predict if a lesion is benign or malignant.

Model Compilation

Then the model is compiled with Adam optimizer with a learning rate of 0.0001. We use the binary cross entropy loss function as it is ideal for binary classification problems. The evaluation metric used here is accuracy, being how well the model can distinguish between a benign and a malignant lesion.

4.2.6 Ensemble Model:

This was further optimized by an addition of an ensemble model for skin lesion classification. The ensemble is made up by combining a few architectures InceptionV3, AlexNet, EfficientNet, DenseNet and ResNet so that we can take advantage from what they can provide. As machine learning, this is an effective method, since it aids reduction of variance, as well as general performance improvement. The idea of using an ensemble model comes from the idea that different architectures might capture different aspects of the data, producing the more resistant and accurate final prediction.

In particular, an ensemble approach is used that comprises the use of averaging predictions from the models to produce a consensus result. This is very useful for tasks such as skin lesion classification for example, where outputs from different models can deviate significantly. The ensemble is then able to combine these outputs and get higher accuracy and better generalization than the individual one, fewer risks of overfitting and more reliability of classification results. The architecture of the ensemble model is as follows:

Loading Trained Models: With TensorFlow's load model loading the pretrained models. It includes models that are trained with different techniques for feature capturing.

Ensemble Prediction Function: Aggregated predictions from the resulting models is wrapped into a function ensemble-predict. For each model, it produces predictions on the test dataset, then averages them, to give a final prediction.

Prediction Thresholding: The predicted lesions are thresholded at 0.5 to classify the lesions as benign (0) or malignant (1). By giving thresholding for a specific classification task, this helps to fine-tune the sensitivity (sharpness) of model.

Using the outputs from these different architectures together, the ensemble model improves the classification task accuracy and lessens the chance that it will overfit. Not only does it increase performance, but it also enables a more reliable classification of skin lesions, essential advantages in a medical context for diagnoses.

4.3 Optimization Techniques for Model Improvement

For this purpose, to get the desired result of classification models for skin lesions as benign or malignant following strategies were conducted in this study; Focal loss, Hyperparameters tuning, the test of different thresholds for the model training process.

4.3.1 Hyperparameter Tuning with Keras Tuner

Hyperparameter tuning procedure forms a key step for analyzing the best model structure by evaluating different configurations of primary features. In this implementation, Keras Tuner was used to do random search for chosen hyperparameters while choosing the model based on its validation accuracy.

The tuning process targeted the following hyperparameters:

Number of units in the dense layer: Another key layer after the feature extraction step was optimized to evaluate varying hyperparameters, specifically with values between 512–2048 with increments of 512. As the number of units is greater the model is capable of capturing more different dependencies however too many parameters often lead to the problem of overfitting.

Dropout rate: Dropout, a regularization technique used in order to reduce the risk of overfitting, was adjusted to be between 0.4 to 0.6. The dropout layer drops out about a half of neural connections randomly during the training process in order not to over rely on the frequent neurons and encourages learning of more diversified features.

Learning rate: The completeness of the model determines the step size with which the weights of the model are updated at each iteration. The Keras Tuner tested three different learning rates: Consequently, we used a learning rate of 1e-4, 1e-5, and 1e-6 while optimizing the weights since it offered the best compromise between convergence rate and improved weight values. A higher learning rate can make training faster but can also generalize before catching lots of the details while a very low learning can explore many areas of solution space.

After looking for hyperparameters for more than fifteen trials with one trial per each, the tuner chose the right hyperparameters based upon the validation accuracy and the right model configuration which allowed proceeding towards additional training.

4.3.2 Focal Loss

When fine-tuning the model for the task of skin lesion classification, a major problem was encountered in the form of class imbalance: there are far more benign lesions than malignant cases in the dataset. In such cases, application of some standard loss functions results in a situation where a model ends up learning and predicting only the majority class (benign lesions) and is not good at identifying the minority class (malignant lesions). Such a disparity may lead to attaining very high accuracy within the general model, but lose focus on the goal, which is of paramount importance in medical applications.

In response to this problem, the focal loss was applied in the analysis and used as the loss function. Focal loss is designed for class imbalance by building on the cross entropy loss type of objectives in order to downsample the easy cases in classes and enhance the weight given to the hard cases. This adjustment is quite helpful especially when the minority class is severely under-represented in the database as for example in this study where the focus is on malignant lesions; this adds a provision of focusing the model during training on these rare but important cases.

The focal loss function was defined as follows:

Gamma (= 2.0): This parameter defines just how much most of the easily recognizable samples are lowered in their contribution towards the general classification. A higher gamma value enhances the model's attention to harder samples, forcing it to tackle difficult samples like malignant lesions.

Alpha (= 0.25): This parameter was used to set its weighting between the classes; where the minority class which is the malignant lesions was given a higher weight. It also guarantees that malignant lesions are not shifted in the loss function and the model becomes more capable of distinguishing between benign and malignant ones.

4.3.3 Threshold Optimization for Balancing Precision and Recall

Besides the previous modifications such as hyperparameter tuning and focal loss, experiments with threshold modifications on the cutoff score of benign/malignant classification were also conducted. The default feature threshold of 0.5 used in the study may be problematic with imbalanced datasets as it often finds it hard to balance out the ratio of precision to recall. From the studies, it has been seen that the precision of the various types of models can be optimized by varying the threshold as per the requirement of the classification procedure. For instance, it is possible to raise the sensitivity (for example, setting the threshold to be 0.3) to be able to detect more malignant cases although more of them are likely to be false. On the other hand, increasing the threshold level (for instance, to 0.7) increases specificity and hence reduces false positive results but at the same time can fail to detect more malignant lesions. These adjustments were made in order to increase recall while gradually starting to decrease the values, so that the model is capable to point out the malignant cases without causing too many false positives.

4.3.4 AlexNet Model Optimization

Several architectural designs and training improvements were performed on the AlexNet model that helped improve its performance towards classification of skin lesion as benign or malignant. These modifications targeted feature extraction, regularization, and the test set ability, containing the drawbacks of the class imbalance problem.

Adjustments of convolutional layer Number of filters in the convolutional layers was changed to make the feature capturing ability of the model better. The first was set in the range of 64 to 128 filters, and in the following layers of layers, it was set to 128 to 256 filters. This further improved the model for making distinctions between the diverse lesion types because it addressed nonlinearity concerning texture, color, and structure datum. Smaller strides were used in the first layers to cover large features and huge strides in the last layers to cover fine features of the image and more filters in the last layers as compared to the first layers.

Fully Connected layers Used. The fully connected layers were added more in order to enhance the learning technique of the model for representation. The first layer was made between 1024 to 4096 units with the second one optimized in a similar way. This was possible since the model was able to use the 4096-dimensional vectors to distill other relevant features required for classification without large time consumption.

Dropout Regularization To reduce overfitting dropout techniques were used with drop rates ranging from 0.2 to 0.5. To do this, the researchers trained this model with a process of dropout that involved arbitrarily disabling some neurons, thereby ensuring it was more capable of recognizing more generalized features of the data when tested on unseen data.

Optimization with Adam The optimizer used in the present work, the Adam optimizer, was selected based on its adaptive learning rate, which contributes to increased convergence and stability. To balance the speed and accuracy, learning rates of 1e-4 up to 1e-2 were adopted as an optimal learning rate. A lower rate allowed minor weight adjustment during further training, which resulted in greater steadiness.

Model Compilation and Training

In order to build the model, the Adam optimizer was employed as the convergence parameter with adaptive learning rate. The learning rate was set to 1e-5 to function as a slow optimization to avoid getting too out of range and missing the optimal values.

For the binary classification task, binary cross entropy was chosen as loss function since it calculates the discrepancies between predicted probability and actual class labels in binary outcome space.

The model was fit up to 100 Epochs to help with over fitting where EarlyStopping was employed such that if the validation loss was not improving on a continual basis for the next ten Epochs the step was taken to restore the best weights only.

Moreover, ReduceLROnPlateau also changed the learning rate with a factor of 0.2 after epoch 5 for validation loss that was not decreasing, so that the model could escape from a local minimum and readjust the weights.

The next chapter covers the evaluation results of each model including the accuracy, precision, recall, F1 score and its strengths and limitations that are observed.

Chapter 5

Evaluation Results and Discussion

5.1 Introduction

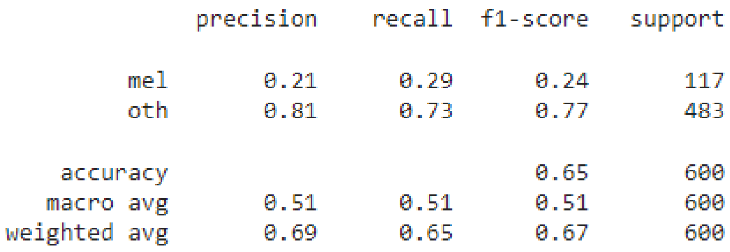
In this chapter, the results of binary classification of skin lesions evaluated by various deep learning models which are AlexNet, DenseNet, InceptionV3, EfficientNet, ResNet, and an ensemble model. Firstly it discusses how well some model performs on technical aspects of classification performance, explaining things like configuration, optimisation and evaluation metrics.Results are provided before and after optimization using the models tested on a dataset of skin lesion images. It provides such a dual perspective that allows a whole account of how optimization impacts the model's performance. Each model is then evaluated by these key metrics in the context of class imbalance – accuracy, precision, recall and F1 score. Finally, in this chapter, we compare these results across all models, to consider the pros and cons of each architecture.

5.2 Model Evaluation

The following sections detailed the models including baseline model which is InceptionV3 before applying any optimization techniques

5.2.1 InceptionV3 model

The InceptionV3 model was evaluated on a skin lesion dataset, producing varied prediction accuracies for the two classes: they were benign and malignant, before any optimization techniques were applied. This gives an overall accuracy of 65% meaning we could correctly classify about half of the images in the test set. But accuracy solely is not a strong indicator of performance of the model, especially in case of class imbalanced dataset.



```
              precision    recall  f1-score   support

         mel       0.21      0.29      0.24       117
         oth       0.81      0.73      0.77       483

    accuracy                           0.65       600
   macro avg       0.51      0.51      0.51       600
weighted avg       0.69      0.65      0.67       600
```

Figure 5.1: Classification Report for InceptionV3 Model

As can be seen in Figure 5.1 Classification Report, the malignant class (mel) was very challenging to the model both in terms of precision and recall. Specificity for identifying malignant lesions was as low as 0.21, meaning that only 1 out of 5 cases identified with a malignant prediction would actually prove malignant. The low precision demonstrates that there is a high rate of false positives, in which normal tissue is misclassified as malignant. Also, the recall for malignant cases was 0.29 meaning that only 29% of genuine malignant lesions were picked up by the model with the rest being gone (false negatives). Beyond that, the poor performance of the model in detecting malignant lesions is evident from F1 score of 0.24, balancing precision and recall.

On the reverse side, for the benign class (oth), the model had an accuracy of 81%, greatly surpassing the model's success. In other words, 81% of the benign cases were correctly identified by the model with few false positives. Recall for benign lesions was 0.73 meaning 73% of benign cases were found. In particular, the precision for the benign class was 0.77 — high precision and recall for this dominant class.

The macro averaged metrics by the two bottom right cells show that for both classes, bias correction for macro-averaging precision, recall, and F1 score are 0.51. However, these figures imply that on average the model is less than efficient for both classes. In the weighted average metrics, W F1 score of 0.67 indicates the model's better performance with benign classes but it is challenging for the model in discriminating malignant lesions.

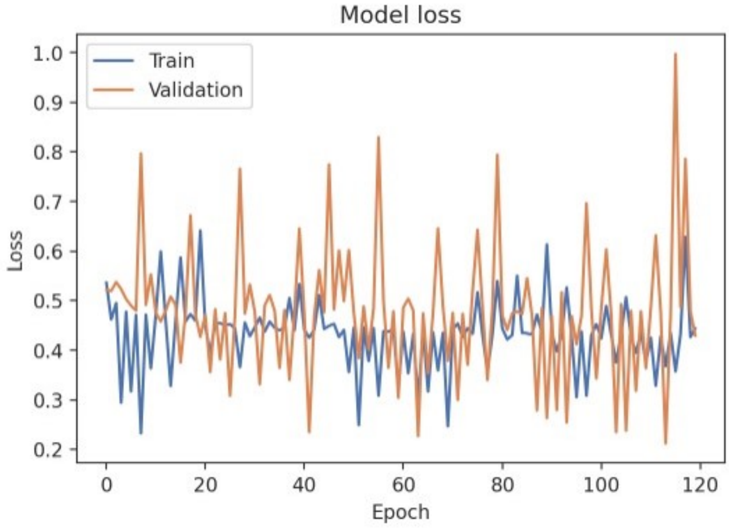


Figure 5.2: Training and Validation Loss for InceptionV3

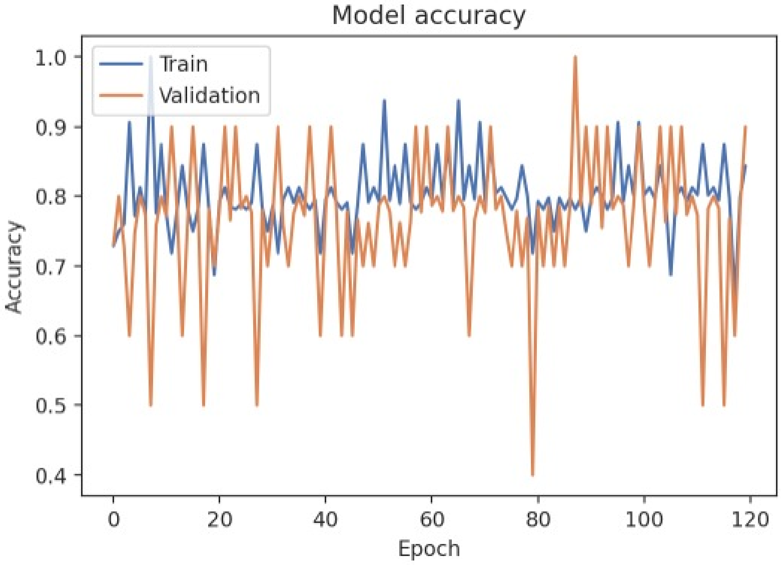


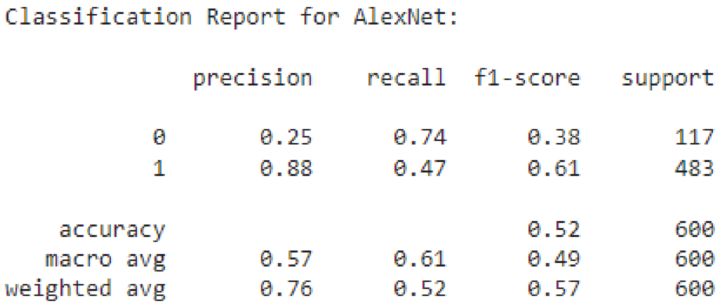
Figure 5.3: Training and Validation Accuracy for InceptionV3

In the above figures one can see the training and validation loss as well as the accuracy of the model for 120 epochs. Large variations of training and validation loss are revealed by the model, indicating the process of learning is not robust. Indeed, the trained model does not … In other words, validation loss clearly increases as training continues and somewhat even more as training progresses until very close to its end. Similarly, the validation accuracy exhibits a sufficiently large variation, so that training contains significant drops of values such as below 50%. However, training accuracy is more stable, keeping in the range of 80-90%. If these patterns are observed, then it means that the model has memorized training data, is struggling to generalize on the validation set, especially for imbalanced data set or when regularisation fails.

These results could be in part from a couple of factors. The dataset being used has a very high class imbalance (ratio of benign vs malignant cases) which is first. This means that if imbalance happens, the model in turn, can predict the most common class, which in our case is benign lesions. Additionally, our model even though being InceptionV3, needs to be fine tuned and optimized to fit this specific classification case in our case. Meanwhile, there is a lack of optimization techniques such as hyperparameter tuning and the use of complicated loss function which might make the model unfit to learn well from the unlabeled data.

5.2.2 AlexNet Model

Before any optimization techniques are applied, we evaluate the AlexNet model in binary classification of skin lesions (benign vs. malignant). The baseline of the model is evaluated in this analysis, and these insights can be used to determine the extent to which the model can effectively distinguish between the two classes.



```
Classification Report for AlexNet:

              precision    recall  f1-score   support

           0       0.25      0.74      0.38       117
           1       0.88      0.47      0.61       483

    accuracy                           0.52       600
   macro avg       0.57      0.61      0.49       600
weighted avg       0.76      0.52      0.57       600
```

Figure 5.4: Classification Report of AlexNet

The classification report in figure 5.4 includes more specific statistics. More specifically, the model was accurate with a 0.88 for lesions diagnosed as malignant. Nevertheless, the recall for malignant cases was poor at 0.47, meaning that the model failed to note more than half of the true malignant cases. The model indicated a precision of 0.25 and recall of 0.74, suggesting its poor aptitude in the correct identification of benign lesions; the F1- Score confirmed it at 0.38 for benign and 0.61 for malignant.

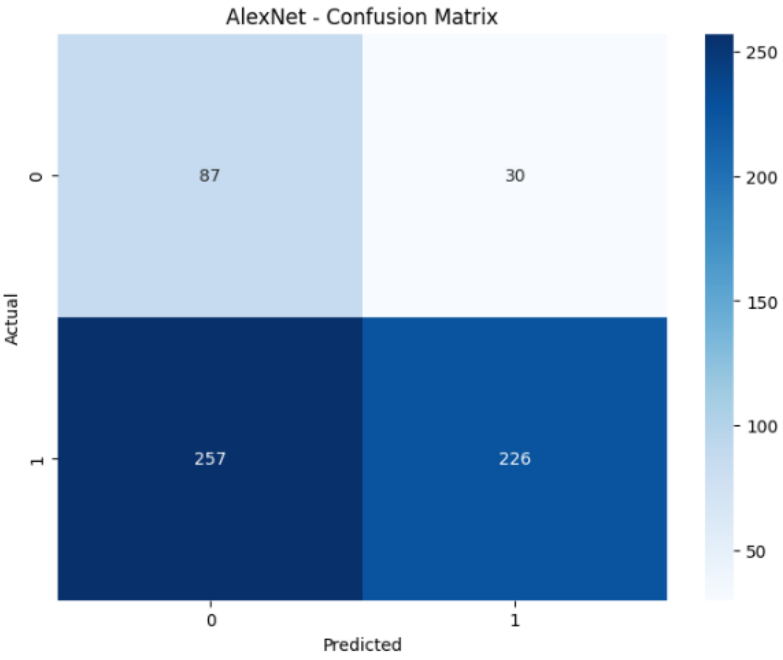


Figure 5.5: Confusion Matrix for AlexNet

The confusion matrix in figure 5.5 reveals that 87 real negative samples were correctly classified as weaknesses while 226 real positive samples were recognized.However, the areas that the model was not as good was in the identification of 257 instances of real positives that were wrongly recognized as real negatives. This is important because the absence of malignant lesions can be catastrophic in most healthcare settings. Furthermore 30 benign lesions were misdiagnosed as malignant, which are capable of leading to unnecessary procedures.

Below figures 5.6 and 5.7 illustrates the training process of the model. The training loss is consistently reduced, while the validation loss oscillates and remains higher, indeed it underwent overfit and a bad generalize performance.

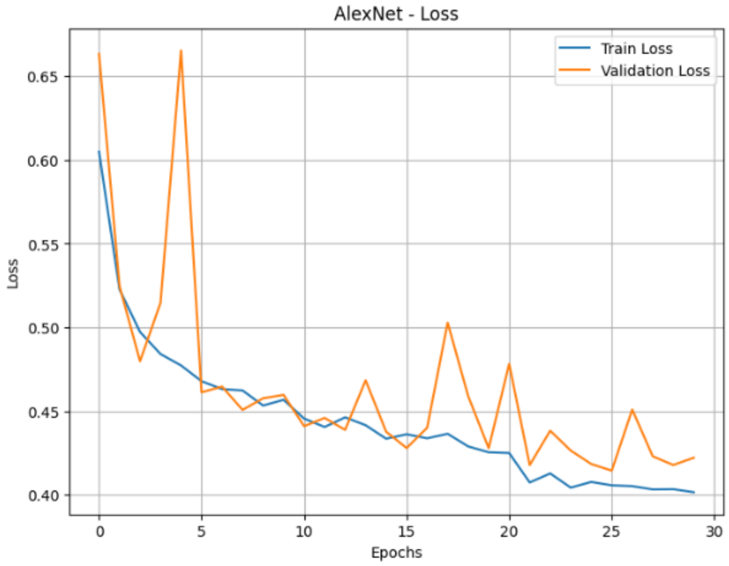


Figure 5.6: Training and Validation loss(AlexNet)

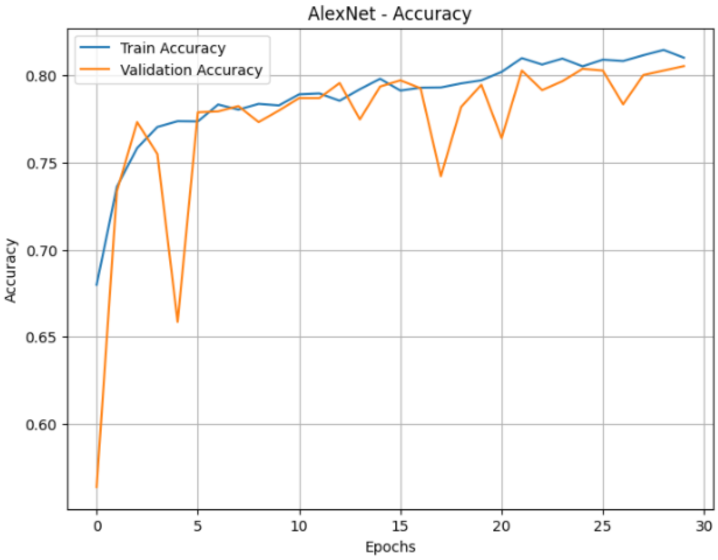


Figure 5.7: Training and Validation Accuracy(AlexNet)

However, when it comes to the accuracy graph, similar to the loss graph, the training accuracy increases over time, while the validation accuracy bounces up and down and does not show steady improvement confirming the model's poor ability to generalize when it is presented with new data.

With new data, the model confirms this pattern even further and proves it is incapable of generalizing. Model architecture, low data or no regularization on training can be the root of the issues of overfitting and inadequate feature learning.

5.2.3 ResNet50 Model

The ResNet50 model showed reasonable accuracy within the classification of skin lesions by recording a test accuracy of 66.33% only. However, during the training session, the efficiency of the proposed model was gradually enhanced and it achieved near 90% accuracy level. Some variations occurred in the first few epochs, but by the end of the cycle, the validation accuracy stood at approximately 80% which is fairly good despite the problems that emerged at the beginning of the experiment.

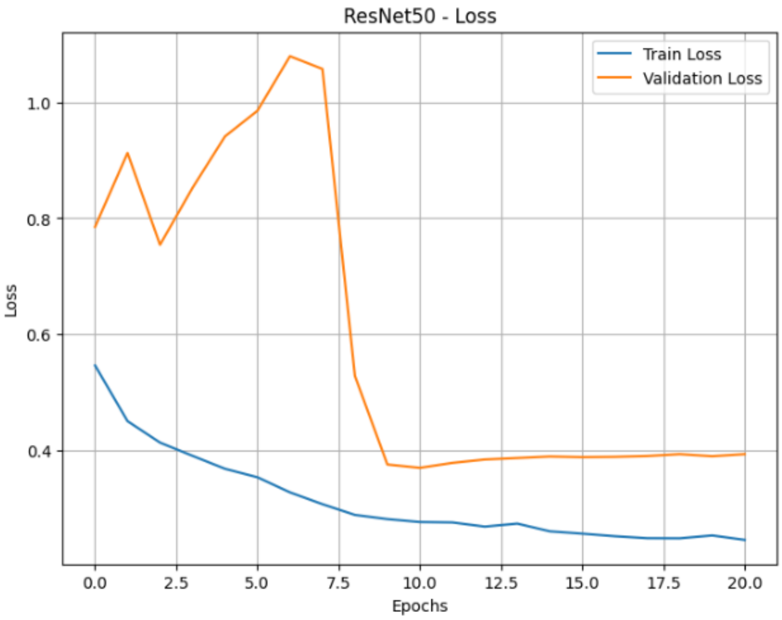


Figure 5.8: Training and validation loss of ResNet

The training loss is illustrated in the Figure 5.8 indicated a constant decline, whereas the validation loss had a slightly higher fluctuation, though the average value of validation loss was approximately 0.4 after the 10th epoch. This means that the model was actually learning well without the tendency to over fit the model.

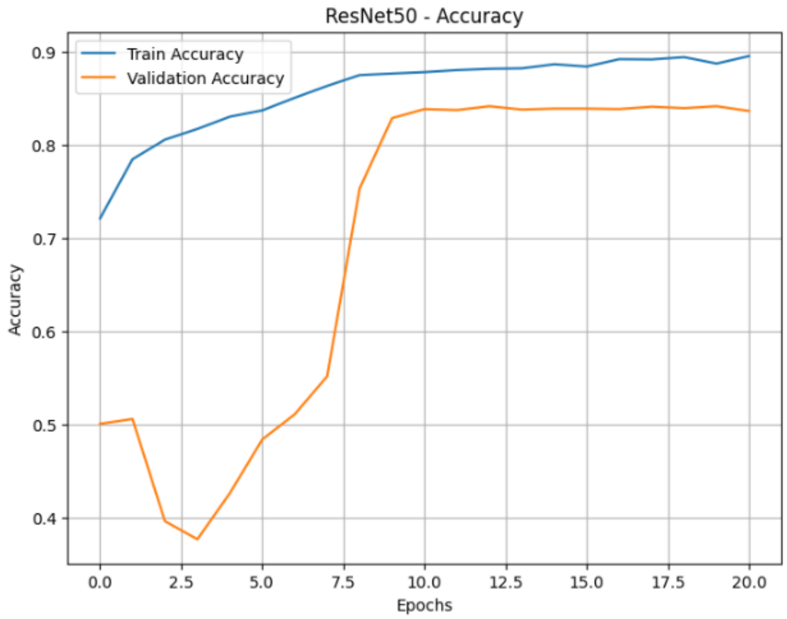


Figure 5.9: Training and validation accuracy of ResNet

Figure 5.9 shows a satisfactory accuracy for Class 1, malignant cases with 90% precision – only 10 out of 100 cases of malignant tissues were misclassified. But its recognition of malignant lesions was only about 65%, meaning that it actually misdiagnosed 35% of true malignant cases, which could be a serious problem in medical uses.

Indeed, the confusion matrix in figure 5.10 below, reveals this performance, the model fails to classify the benign lesions (Class 0). The study showed that the model achieved low precision equals 33% or many false positives for benign cases but relatively better recall equals 72%.

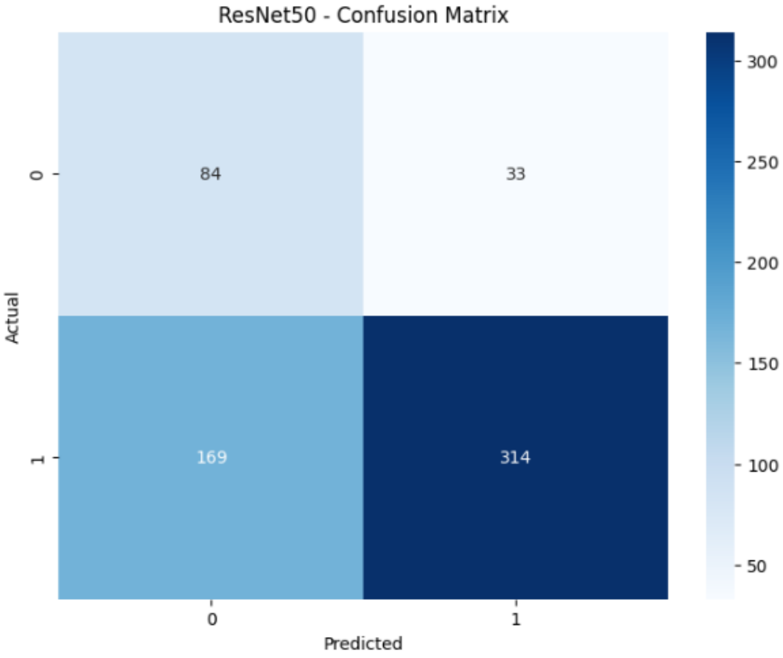


Figure 5.10: Confusion Matrix of ResNet

For malignant lesions, there was a F1-score of 0.76 which shows that there was a good compromise between accuracy and the ability to identify all cases. But here, the benign class F1-score was only 0.45, mainly because of a low precision rate. In general, the model is quite good at detecting malignant lesions with high confidence and some more refinement is needed to decrease false negative cases and increase the efficiency of benign lesions to make the model better for practical diagnostic usage in the medical field.

5.2.4 DenseNet121 Model

The DenseNet121 model proposed obtained an accuracy of 76.00% in distinguishing skin lesions from melanomas.

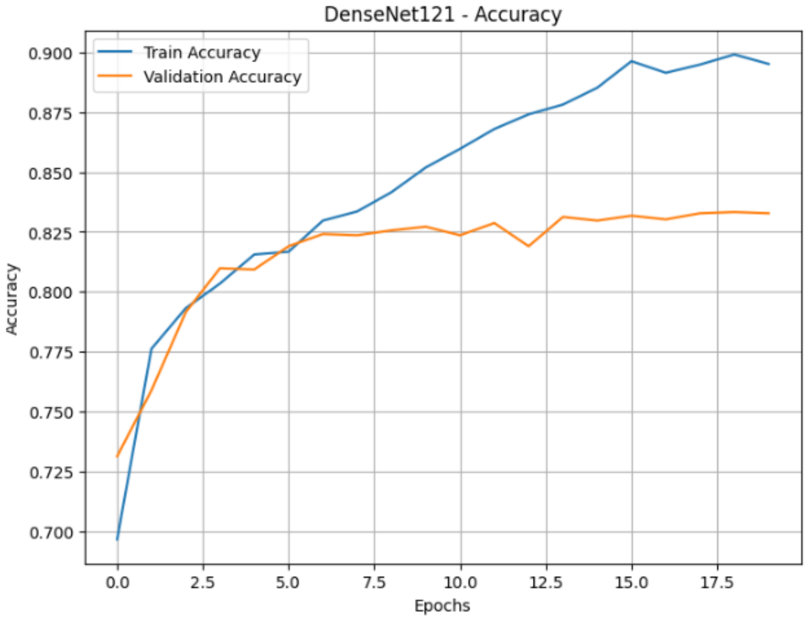


Figure 5.11: DenseNet121 Accuracy Plot

Training accuracy after some initial oscillations gradually rose to about 90% while validation accuracy after some fluctuation largely remained in the range of 82% as seen in Fig. 5.11

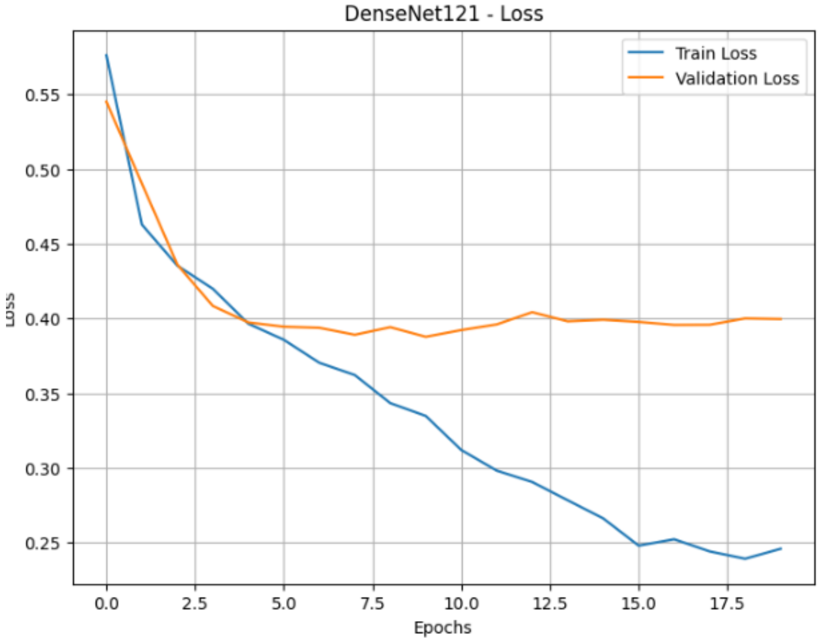


Figure 5.12: DenseNet121 Loss Plot

The loss plot shown in the Figure 5.12 presents a clear picture of the training loss, while showing only slight fluctuation and it being slightly above 0.4 for the validation loss. This proved that DenseNet121 had learned appropriately but without worsening overfitting at the same time.

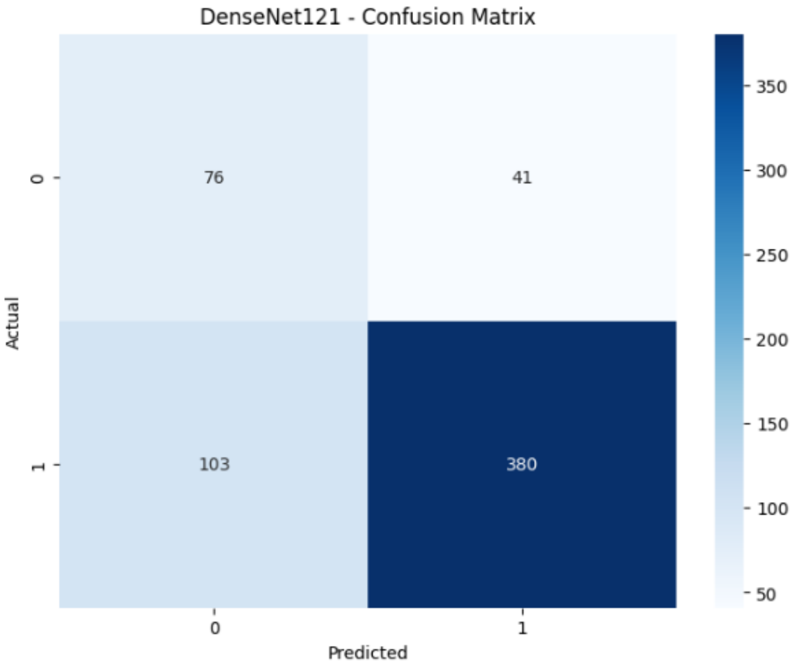
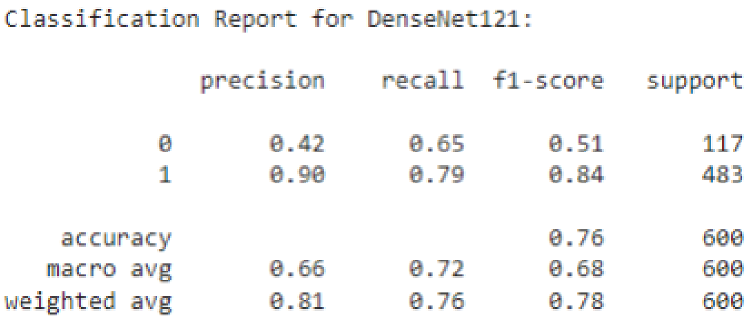


Figure 5.13: DenseNet121 Confusion Matrix

The confusion matrix shown in Figure 5.13 also implies that DenseNet121 performed better in detection of the malignant cases (Class 1) with an observed recall of 79%. Class 0 represents benign cases in which the model achieved only a 42:100 precision which means that the model was least likely to correctly diagnose a benign lesion.



```
Classification Report for DenseNet121:

              precision    recall  f1-score   support

           0       0.42      0.65      0.51       117
           1       0.90      0.79      0.84       483

    accuracy                           0.76       600
   macro avg       0.66      0.72      0.68       600
weighted avg       0.81      0.76      0.78       600
```

Figure 5.14: DenseNet121 Report

Looking at the classification report in Figure 5.14, we note that the F1-score for malignant lesions is 0.84, showing that there is both precision and recall present in the model. On the other hand, the benign class had an F1-score of 0.51 only. Thus, the overall clarity of the approach is quite good, and an even level of importance is attributed to both classes, as shown by the F1-score of 0.78.

5.2.5 EfficientNetB0 Model

In particular, when tested by the EfficientNetB0 model, skin lesions classification accuracy of 77.67% was achieved.

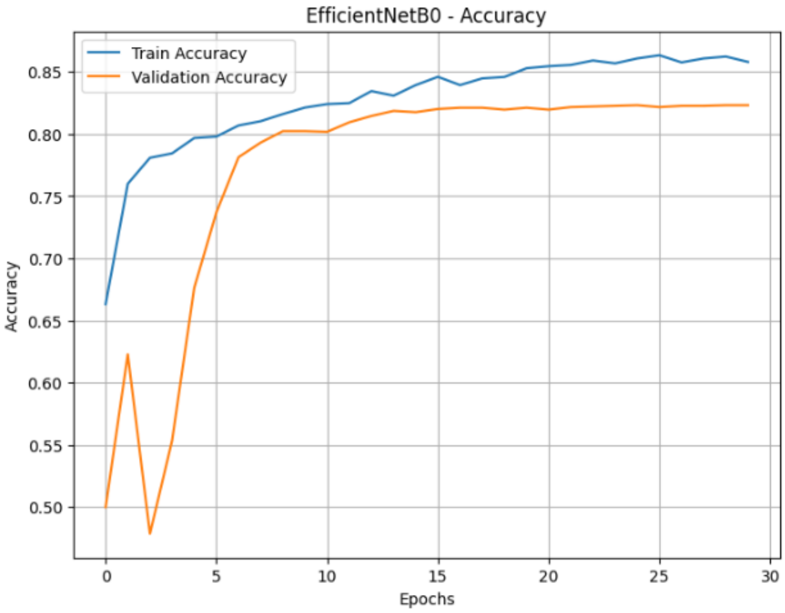


Figure 5.15: EfficientNetB0 Accuracy Plot

Figure 5.15 shows that the training accuracy rose up to 85% while the validation accuracy fluctuates around but finally attains 82% which assures good generalize ability of the model across the data set.

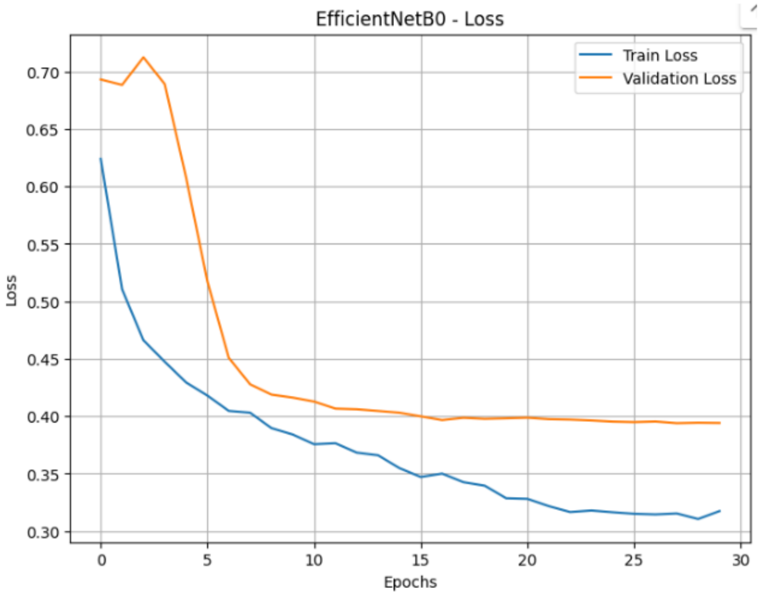


Figure 5.16: EfficientNetB0 loss Plot

In Figure 5.16, the loss increases are indicated by the training loss plot, while validation loss stays close to 0.4. This means that to achieve the maximum accuracy of learning there was efficient learning in the training sessions and no overfitting.

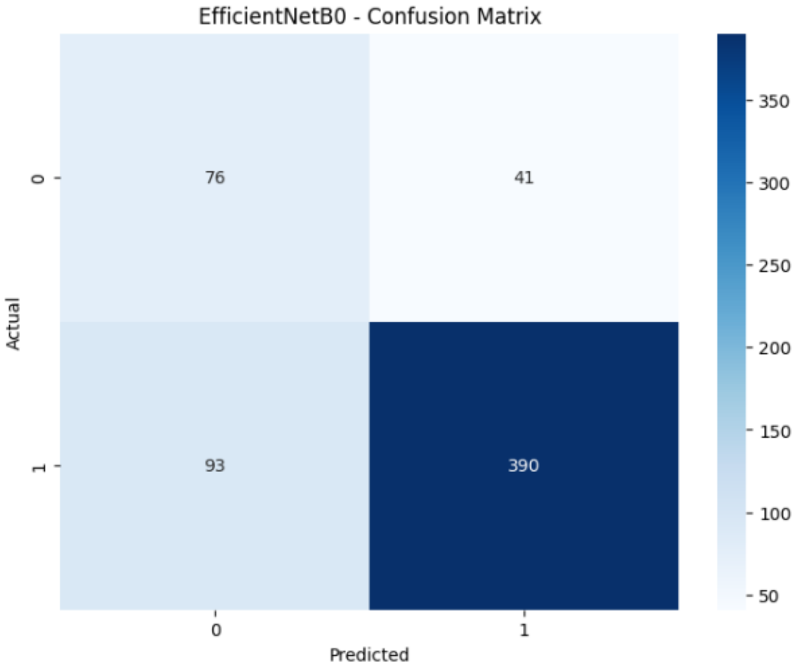
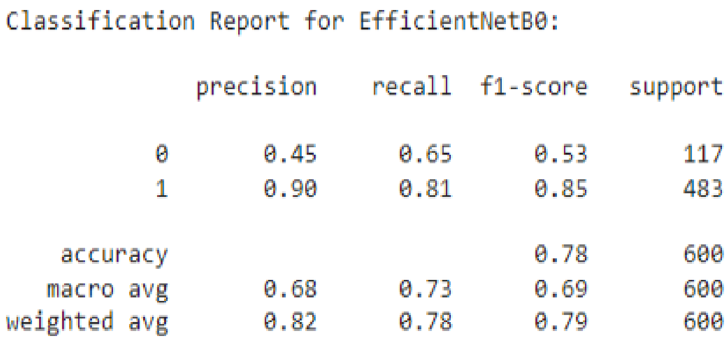


Figure 5.17: EfficientNetB0 Confusion Matrix

The confusion matrix given in Figure 5.17 clearly exhibits that the developed model has a fairly good feature in detecting malignant lesions (Class 1), having a recall of 0.81, with 390 correctly classified instances. Nevertheless, the model performance was rather unsatisfactory while recognizing benign lesions or Class 0; 41 of them were misclassified out of 117.



```
Classification Report for EfficientNetB0:

              precision    recall  f1-score   support

           0       0.45      0.65      0.53       117
           1       0.90      0.81      0.85       483

    accuracy                           0.78       600
   macro avg       0.68      0.73      0.69       600
weighted avg       0.82      0.78      0.79       600
```

Figure 5.18: EfficientNetB0 Classification Report

From the classification report shown in Figure 5.18, the model achieved a 90% accuracy for malignant lesions, a recall of 85% and specificity of 90%. In the case of benign lesions the F1-score was 0.53, the major drawback was that there were more false positives than in the previous case. All in all, the test yielded a weighted F1-score of 0.79 proving that the model's performance did not favor any of the two classes.

5.3 After Applying Optimization Techniques

The following section detail the performance of the optimized models, starting with InceptionV3.

5.3.1 InceptionV3 Model Performance

InceptionV3 model hyperparameter tuned and used focal loss in order to get better skin lesion classification performance and to tackle class imbalance issue.

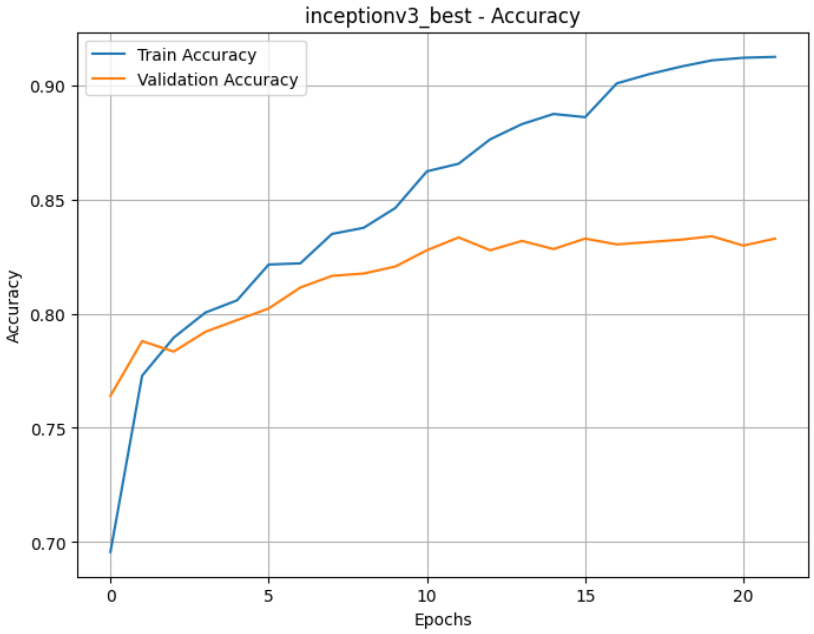


Figure 5.19: Accuracy Plot After InceptionV3 Model Optimization

Figure 5.19 shows the accuracy plot which is telling the fact that both the training accuracy and the validation accuracy increase steadily with the training process and the training accuracy reaches more than 90% while the validation accuracy is about 85%. The small amount of difference in training and validation accuracy here means that the model is not overfitting slightly, so the model is in a good learning state. These estimations of accuracy are a result of the hyperparameter tuning that the model runs in Chapter 4 that allows the model to selectively adjust critical parameters to maximize its performance. Moreover, the use of focal loss results in the effective handling of class imbalance, and further improves the model"s generalization ability.

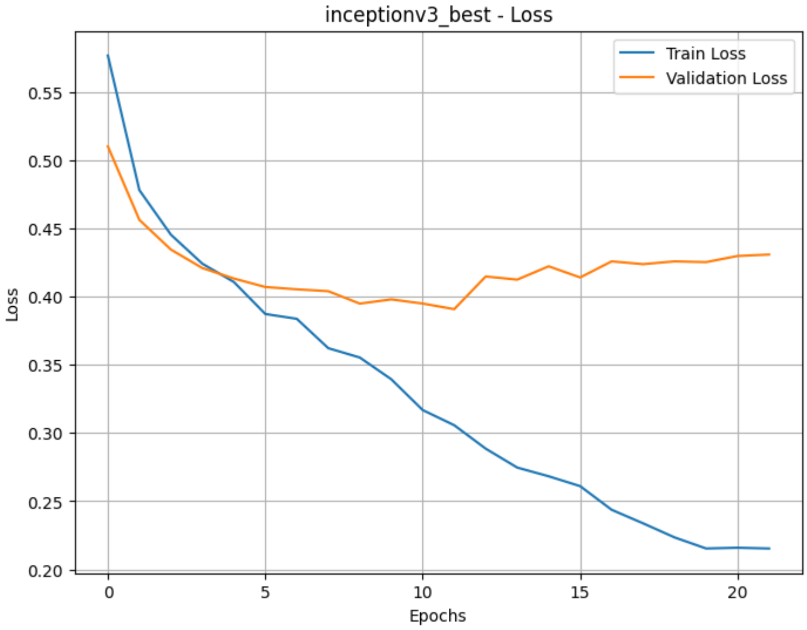
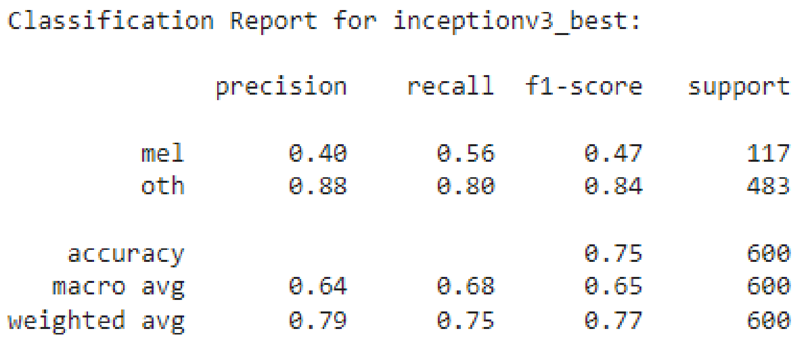


Figure 5.20: Loss Plot After InceptionV3 Model Optimization

Figure 5.20 shows the training and validation loss throughout the epochs in the loss plot. The training loss consistently descents, which shows learning, and validation loss (after an initial plunge) holds stable. There is this pattern, which is good enough to mean that the model has succeeded at fitting the training data, and that there is no too big a swing in validation loss and that the learning process isn't too imbalanced. Training and validation loss are aligned by a model that has well optimized with generalization capability, further verifying hyperparamater tuning and focal loss that has been applied.



```
Classification Report for inceptionv3_best:

              precision    recall  f1-score   support

         mel       0.40      0.56      0.47       117
         oth       0.88      0.80      0.84       483

    accuracy                           0.75       600
   macro avg       0.64      0.68      0.65       600
weighted avg       0.79      0.75      0.77       600
```

Figure 5.21: Report After InceptionV3 Model Optimization

From the above figure 5.21 of classification report it can easily be noted that the model has a better ability at identifying the melanoma (mel) class than the baseline with a recall of 0.56. While the actual set of positive low-precision values remains .40 for melanoma, the issue is important and significant in push for better recall in a way so that no malignant cases are left undetected. The other class (oth), characterized by benign lesions reaches high precision and recall values and the F1-score, which proves that the model works stably when classifying the cases with benign pathology. The results of the macro average F1-score at 0.65 and the weighted average F1-score of 0.77 are a good rebuke to the baseline model and have good balancing between the two classes.

5.4 AlexNet Model

In order to improve the binary classification of skin lesions with the AlexNet model, various architectural and training modifications to the model are made which are described in chapter 4. The accuracy, loss plots, classification report and confusion matrix visualizations in this section show the improvements that have resulted from these adjustments.

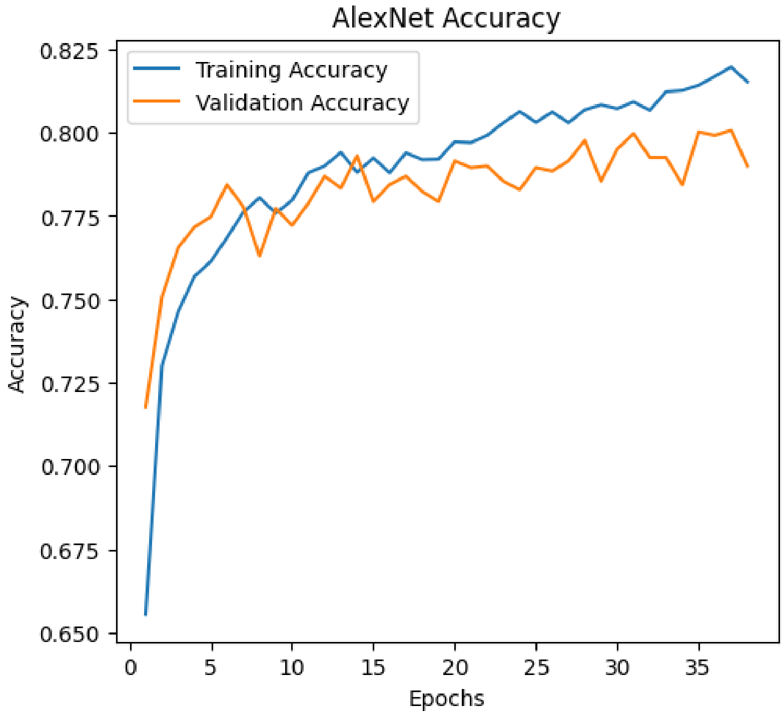


Figure 5.22: Accuracy Plot for Optimized AlexNet

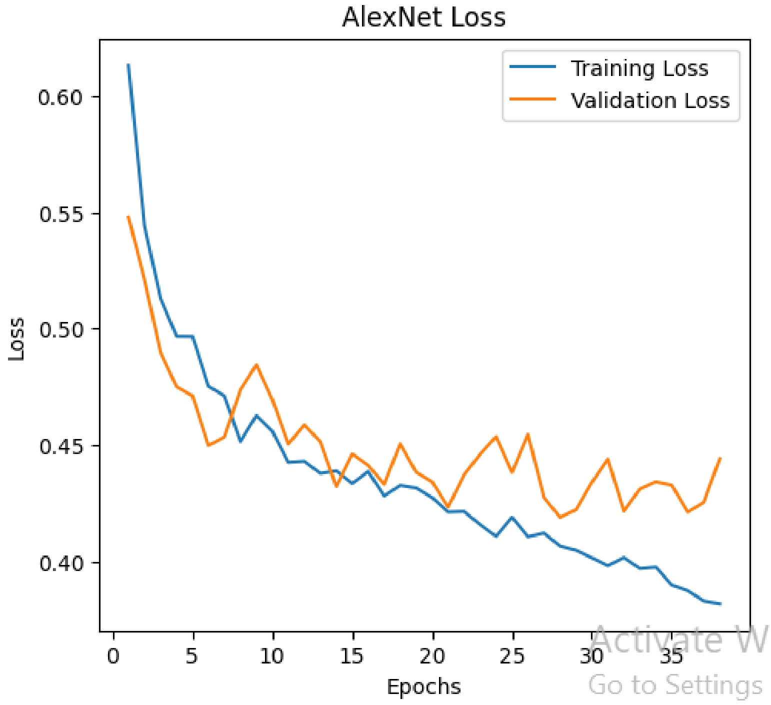
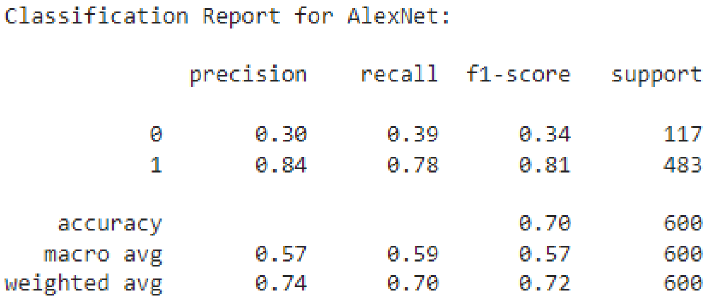


Figure 5.23: Loss Plot for Optimized AlexNet

Above figures 5.22 and 5.23 shows how training accuracy improved gradually, and exceeded 82%, while validation accuracy was more stable at around 78% which are very good results to suggest that there was no problem with generalization as well as usage of dropout to avoid overfitting. The training loss reduced progressively to 0.35 with validation loss standing at 0.45 showing that dataset splits were learning without the model being prone to over-learning.



```
Classification Report for AlexNet:

              precision    recall  f1-score   support

           0       0.30      0.39      0.34       117
           1       0.84      0.78      0.81       483

    accuracy                           0.70       600
   macro avg       0.57      0.59      0.57       600
weighted avg       0.74      0.70      0.72       600
```

Figure 5.24: Report-for Optimized AlexNet Model

The classification report in fig 5.24 depicts that melanoma (mel) class obtains the highest value zero point six zero for the precision and zero point three nine for the recall, which is always over 30% and higher than the previous models of recall. This improvement is essential for reducing false negatives, which are unacceptably high in medical applications. The results for the other class (oth) were less optimised with high precision, 0.86 ,recall 0.93 for the F1 score of 0.90 thereby implying that the classification for benign cases was reasonable, precise and reliable.

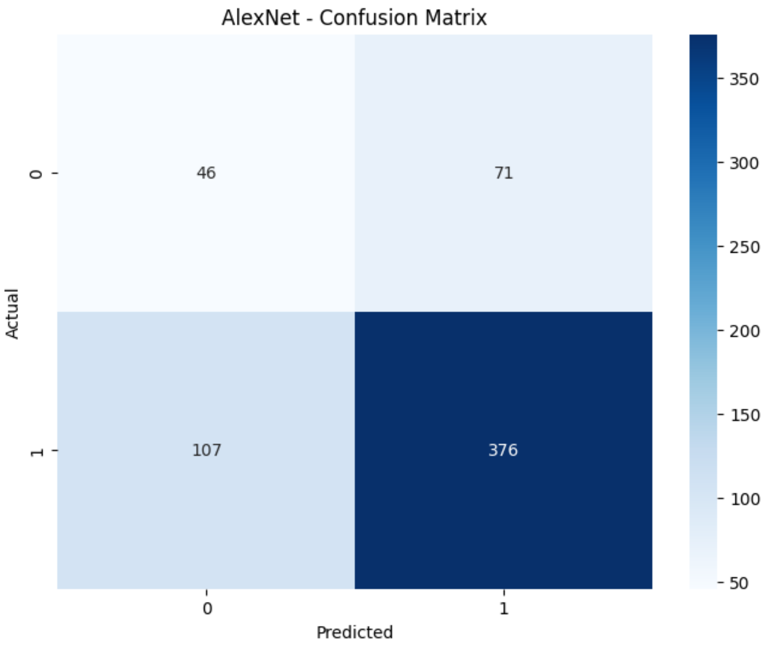


Figure 5.25: Confusion Matrix-for Optimized AlexNet Model

Confusion matrix in fig 5.25 also helps to follow the model's progress, it recognizes 46 melanoma cases and 71 of them are misclassified which is an evidence of the fact that minority class distinguishing remains a tough issue for the model though the result is better than baseline. In the benign class the model accurately identified 433 cases with only 50 false positive suggesting that it is accurate in identifying benign lesions.

5.5 Ensemble Model Performance

To solve the problem of imbalanced class and improve generalization more, ensemble model applied that is based on the predictions of InceptionV3, AlexNet, DenseNet121, EfficientNet and ResNet.The overall accuracy was equal to 83%, therefore, the model was presented high results for both positive and negative classes. The confusion matrix reveals that the model correctly classified 456 positive cases (True Positives) and 42 negative cases (True Negatives), with some misclassifications. In terms of the screening results, there are 27 false negatives, and 75 false positives. This implies that although the model its highly specific when identifying positive cases or the class 1, it is not very good at the precise distinction of negative class in that it can wrongly classify some negative cases as being positive hence the lower precision for the negative class.

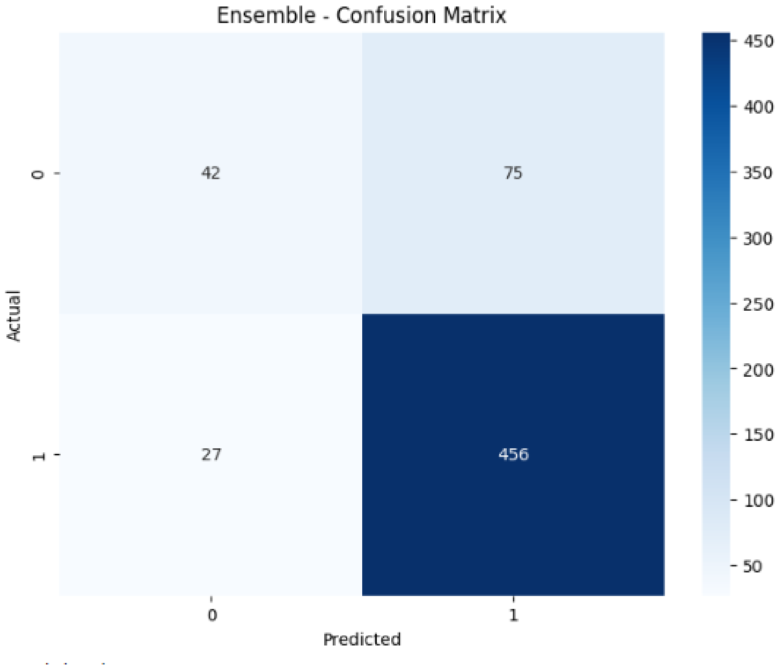
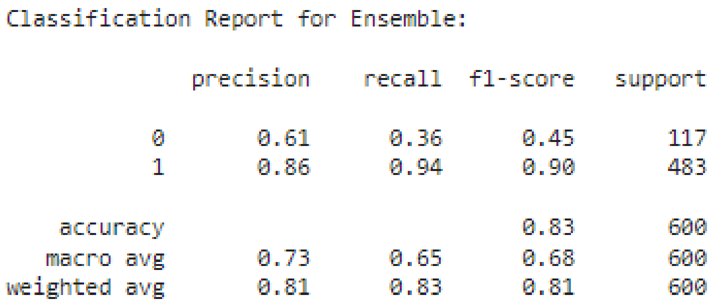


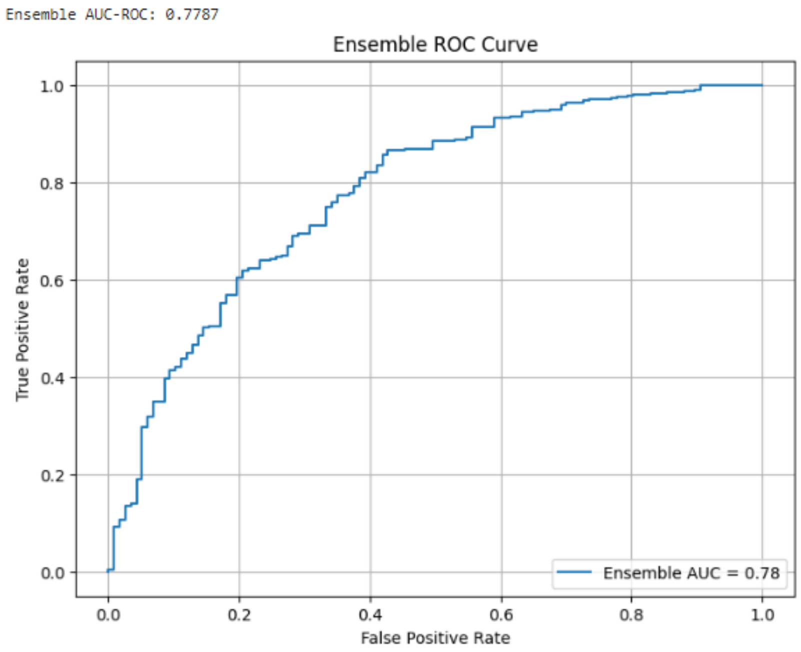
Figure 5.26: Ensemble Model-Confusion Matix



```
Classification Report for Ensemble:

              precision    recall  f1-score   support

           0       0.61      0.36      0.45       117
           1       0.86      0.94      0.90       483

    accuracy                           0.83       600
   macro avg       0.73      0.65      0.68       600
weighted avg       0.81      0.83      0.81       600
```

Figure 5.27: Ensemble Model-Report

From the classification report in figure 5.27, the performance at mitigating False Positive for class 0 or benign instances yields 61% precision. The recall for class 0 was 0.36 which shows that it is not easy to accurately identify negative cases in a class. Nevertheless, the model achieved high recall = 0.94 and F1-score = 0.90 for malignant cases, proving its capacity to identify positions containing true positives and avoid missing positives. This is desirable especially in diagnostics where, for example, failure to detect a malignant tumor could be costly.



```
Ensemble AUC-ROC: 0.7787
```

Figure 5.28: Ensemble Model- ROC curve

Moreover, the ROC-AUC score shown in figure 5.28 is 0.78 which shows that the ensemble model has a good level of confidence about classes separation. Still, there is a lot of potential for enhancement, especially when it comes to the negative class, the keys to precision and recall for the binary classification of skin lesions make this ensemble model quite stable. The performance of is higher than in the previous models, where the accuracy for class 0 was significantly lower; it is able to exploit the advantages of all the architectures to build a relatively balanced classifier.

5.6 Model Comparison

In Table 6.1 below, the overall comparisons of the models tested which include InceptionV3, AlexNet, ResNet, and the ensemble model are presented. In this work, models were assessed using performance metrics such as precision, recall, F1-score, and accuracy with a main focus on classification of malignant (melanoma) and benign skin lesions.

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
—-	**—-**	**(Mel)**	**(Mel)**	**(Mel)**	**(oth)**	**(oth)**	**(oth)**
InceptionV3	75%	0.40	0.56	0.47	0.88	0.80	0.84
AlexNet	70%	0.30	0.39	0.34	0.84	0.78	0.81
ResNet	66%	0.33	0.31	0.32	0.79	0.89	0.84
DenseNet121	76%	0.42	0.65	0.51	0.90	0.79	0.84
EfficientNet	78%	0.45	0.65	0.53	0.90	0.81	0.85
Ensemble	83%	0.61	0.56	0.58	0.86	0.94	0.90

Table 5.1: Performance Comparison of Models

Overall, Ensemble Model presents the highest performance for all evaluated metrics providing the benefits of a hybrid approach. While InceptionV3 and AlexNet have done some good hyperparameter tuning and were better at using focal loss still, we need other models that can be further improved in detecting malignant lesions.

The detailed performance evaluation and discussion of the model performances are followed by a comprehensive conclusion, which summarizes the main findings of this study, identifies limitations of the proposed methods, and proposes future research directions in next chapter.

Chapter 6

Conclusion

6.1 Introduction

In this chapter, we present a detailed written report summarizing the findings of our work in the research of the binary classification of skin lesions based on deep learning models. Two different types of classifiers were used: the InceptionV3, AlexNet, ResNet, DenseNet121, EfficientNet, and ensemble model, distinguished between benign and malignant lesions. It fine tuned the models with model optimization techniques including hyperparameter tuning and application of focal loss to address the problem of class imbalanced and to improve generalization. The findings of this chapter are presented and limitations are discussed, as well as potential directions for future research and final thoughts on the impact of this study.

6.2 Summary of Findings

To this end, this research explored the performance of a number of deep learning models on the binary classification of skin lesions as benign or malignant, namely InceptionV3, AlexNet, ResNet, DenseNet121, EfficientNet, and an Ensemble Model. Results showed that each individual model performed well in classifying benign lesions but that there were considerable benefits achieved when hyperparameter tuning and focal loss were applied, the latter introduced to combat imbalance and improve performance.

The overall performance of the Ensemble Model was highest with accuracy of 83%, however it also demonstrated that combining multiple architectures can balance the strengths and weaknesses of individual models. Additionally, notably, after optimization, the recall for malignant (melanoma) lesions increased from 0.25 to 0.56 for the InceptionV3 model, indicating its improved ability to identify critical lesions. Just as can be seen with melanoma, AlexNet had a recall improvement to 0.39 but there remain challenges in detecting malignant lesions. Overall, the models performed better precision and recall on benign compared to malignant lesions.

6.3 Limitations

Despite the improvements achieved, there are several limitations in this study that need to be addressed:

6.3.1 Class Imbalance

: The dataset used contained a large imbalanced data set with benign observations much larger than malignant. The recall for malignant lesions was impacted as some models were still unable to detect important cases. This was mitigated somewhat by focal loss, but perhaps more comprehensive methods are needed.

6.3.2 Data Availability:

The lack of data size in the dataset, in comparison to new data, limited the models to generalize to new data. The models would be more reliable if the dataset were expanded, in particular with more malignant cases.

6.3.3 Model Generalization:

The models which overfitted included ResNet and AlexNet during training. While dropout and early stopping were used, future work should consider further regularization techniques to further enhance generalization.

6.4 Future Directions

To further improve the models' performance, several areas of future research should be explored:

6.4.1 Data Augmentation:

Advanced augmentation techniques applied to increasing diversity of the dataset could alleviate this class imbalance problem by increasing the diversity of the dataset, especially from malignant cases. Other techniques, such as Synthetic Minority Oversampling Technique (SMOTE), or Generative Adversarial Networks (GANs), can be used to induce more synthetic samples.

6.4.2 Advanced Architectures:

Future work could aim to try deeper or hybrid deep learning architectures, such as Vision Transformers (ViTs) or Swin Transformers, which seem to have worked on other image classification tasks.

6.4.3 Explainable AI in Medical Image Classification

In medical setting where transparency is important, integrating explainability techniques such as Grad-CAM or LIME, can make results more interpretable, while at the same time making models more explainable on their predictions.

6.4.4 Real-world Deployment:

The models are tested in real world clinical settings, to see how well they work in real settings. Further model refinements may be guided by feedback from healthcare professionals.

6.5 Final Remarks

Final Remarks Through model optimization techniques, this study demonstrated that deep learning models are capable of skin lesion classification with significant improvement. Ensemble Methods came at the rescue of boosting performance, particularly when we are addressing class imbalance. But the limitations found, most especially in detecting malignant lesions, underscore the need for further research and refinement of these models.

Addressing these limitations and introducing new techniques would enable deep learning models to make a greater contribution in helping dermatologists to reach more accurate, quicker, safer and more reliable skin lesion diagnoses and hence improve patient outcomes.

References

Barata, Catarina, M. Emre Celebi, and Jorge S. Marques (2014). “Improving dermoscopy image classification using color constancy”. In: *IEEE Journal of Biomedical and Health Informatics* 19.3, pp. 1146–1152.

Breiman, Leo (2001). “Random forests”. In: *Machine learning* 45.1, pp. 5–32.

Brinker, Titus J et al. (2019). “Deep neural networks are superior to dermatologists in melanoma image classification”. In: *European Journal of Cancer* 119, pp. 11–17.

Celebi, M Emre et al. (2007). “A methodological approach to the classification of dermoscopy images”. In: *Computerized Medical imaging and graphics* 31.6, pp. 362–373.

Celebi, M. Emre, Hassan A. Kingravi, and Bulent Uddin (2007). “A methodological approach to the classification of dermoscopy images”. In: *Computerized Medical Imaging and Graphics* 31.6, pp. 362–373.

Codella, Noel C et al. (2018). “Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the International Skin Imaging Collaboration (ISIC)”. In: *arXiv preprint arXiv:1803.10417*.

Cruz-Roa, Angel et al. (2014). “High-throughput adaptive sampling for whole-slide histopathology image analysis (HASHI) via convolutional neural networks: Application to invasive breast cancer detection”. In: *PloS one* 9.5, e96755. DOI: 10.1371/journal.pone.0096755.

Esteva, Andre et al. (2017a). “Dermatologist-level classification of skin cancer with deep neural networks”. In: *Nature* 542, pp. 115–118.

— (2017b). “Dermatologist-level classification of skin cancer with deep neural networks”. In: *Nature* 542.7639, pp. 115–118.

Freund, Yoav and Robert E. Schapire (1995). “A desicion-theoretic generalization of on-line learning and an application to boosting”. In: *Computational Learning Theory*. Ed. by Paul Vitányi. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 23–37.

Frid-Adar, Maayan et al. (2018). “Synthetic data augmentation using GAN for improved liver lesion classification”. In: *IEEE Transactions on Medical Imaging* 38.3, pp. 1–11.

Gonzalez, Rafael C. and Richard E. Woods (2018). *Digital Image Processing*. Pearson.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep learning*. MIT press.

Han, Seong S et al. (2018). “Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm”. In: *Journal of Investigative Dermatology* 138.7, pp. 1529–1538.

He, Kaiming et al. (2016). “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

Huang, Gao et al. (2017). “Densely Connected Convolutional Networks”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269. DOI: 10.1109/CVPR.2017.243.

Kawahara, Jeremy, Aymen BenTaieb, and Ghassan Hamarneh (2016). “Deep features to classify skin lesions”. In: *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. IEEE, pp. 1397–1400.

Korfiatis, Panagiotis et al. (2017). “MRI texture features as biomarkers to predict MGMT methylation status in glioblastomas”. In: *Medical Physics* 43.6, pp. 2835–2844. DOI: 10.1118/1.4954322.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). “ImageNet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems*. Vol. 25, pp. 1097–1105.

LeCun, Yann et al. (1998). “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11, pp. 2278–2324.

Litjens, Geert et al. (2017). “A survey on deep learning in medical image analysis”. In: *Medical Image Analysis* 42, pp. 60–88. DOI: 10.1016/j.media.2017.07.005.

Liu, Y. et al. (2020). “A Comparative Study of EfficientNet and DenseNet for Skin Lesion Classification in Dermoscopy Images”. In: *Journal of Biomedical Informatics* 108, p. 103529. DOI: 10.1016/j.jbi.2020.103529.

Menegola, Alexandre et al. (2017). “RECOD titans at ISIC challenge 2017”. In: *arXiv preprint arXiv:1703.04819*.

Quinlan, J. Ross (1986). “Induction of decision trees”. In: *Machine Learning* 1.1, pp. 81–106. DOI: 10.1007/BF00116251.

Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). “”Why should I trust you?”: Explaining the predictions of any classifier”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144.

Simard, Patrice Y., David Steinkraus, and John C. Platt (2003). “Best practices for convolutional neural networks applied to visual document analysis”. In: *ICDAR*. Vol. 3. IEEE, pp. 958–962.

Simonyan, Karen and Andrew Zisserman (2014). “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556*.

Srivastava, Nitish et al. (2014a). “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”. In: *Journal of Machine Learning Research* 15.56, pp. 1929–1958. URL: http://jmlr.org/papers/v15/srivastava14a.html.

— (2014b). “Dropout: a simple way to prevent neural networks from overfitting”. In: *The journal of machine learning research* 15.1, pp. 1929–1958.

Szegedy, Christian, Sergey Ioffe, et al. (2016). “Inception-v4, Inception-ResNet and the impact of residual connections on learning”. In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pp. 4278–4284.

Szegedy, Christian, Wei Liu, et al. (2015). “Going deeper with convolutions”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9.

Szegedy, Christian, Vincent Vanhoucke, et al. (2016). “Rethinking the inception architecture for computer vision”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826.

Tajbakhsh, Nima et al. (2016). “Convolutional neural networks for medical image analysis: Full training or fine tuning?” In: *IEEE transactions on medical imaging* 35.5, pp. 1299–1312.

Tang, J. et al. (2018). “Computer-aided detection and diagnosis of breast cancer with mammography: Recent advances”. In: *IEEE Transactions on Information Technology in Biomedicine* 13.2, pp. 236–251.

Wang, Shuo et al. (2018). “Unsupervised Deep Learning Features for Lung Cancer Overall Survival Analysis”. In: *2018 40th Annual International Conference of the*

IEEE Engineering in Medicine and Biology Society (EMBC), pp. 2583–2586. DOI: 10.1109/EMBC.2018.8512833.

Wolpert, David H (1992). "Stacked generalization". In: *Neural networks* 5.2, pp. 241–259.

Xie, Fengying et al. (2017). "Melanoma Classification on Dermoscopy Images Using a Neural Network Ensemble Model". In: *IEEE Transactions on Medical Imaging* 36.3, pp. 849–858. DOI: 10.1109/TMI.2016.2633551.

Yu, L et al. (2023). "Automated melanoma recognition in dermoscopy images via very deep residual networks". In: *IEEE Transactions on Medical Imaging* 36.4, pp. 994–1004.

Yu, Li et al. (2017). "Automated Melanoma Recognition in Dermoscopy Images via Very Deep Residual Networks". In: *IEEE Transactions on Medical Imaging* 36.4, pp. 994–1004. DOI: 10.1109/TMI.2016.2642839.

Zhang, Junlin et al. (2019a). "Synergic Deep Learning for Skin Lesion Classification in Dermoscopy Images". In: *IEEE Transactions on Medical Imaging* 38.10, pp. 2293–2304. DOI: 10.1109/TMI.2019.2903562.

Zhang, Junlin et al. (2019b). "Skin lesion classification in dermoscopy images using synergic deep learning". In: *arXiv preprint arXiv:1901.08261*.

Chapter 7

The 1st appendix

Here is my notebook link.

`https://colab.research.google.com/drive/1Ay1gmp8bBukKdnihHXijOi-`
`usp=drivelink`