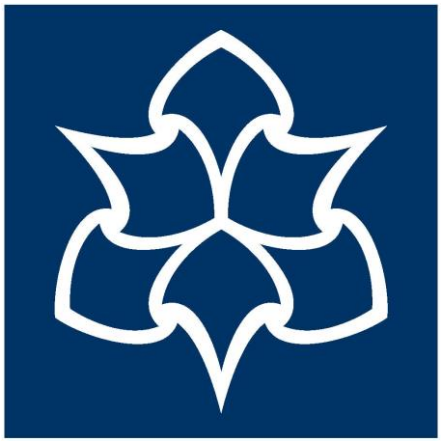
Classification of Skin Lesion : Benign or Malignant

A DISSERTATION SUBMITTED TO MANCHESTER METROPOLITAN UNIVERSITY FOR THE DEGREE OF MASTER OF DATA SCIENCE IN THE FACULTY OF SCIENCE AND ENGINEERING



2024

By

Paramesh kumar Bukya

Department of Computing and Mathematics

Table of Contents

List Of Figures

List of Tables:

Abstract:

This study looks into the use of deep learning algorithms to binary classify skin lesions as benign or malignant using dermoscopic images. Using the International Skin Imaging Collaboration (ISIC) 2016 challenge dataset (ISIC Challenge, 2016) , the study develops and compares several deep learning architectures, including ResNet18 (He et al., 2016), VGG16 (Simonyan and Zisserman, 2014), InceptionV3 (Szegedy et al., 2016), ResNet50 (He et al., 2016), and EfficientNet-B0 (Tan and Le, 2019). The work uses preprocessing approaches and data augmentation to solve issues including dataset imbalance and image variability. Models are evaluated based on criteria such as accuracy, sensitivity, specificity, F1-score, and area under the ROC curve (AUC). The best-performing model, ResNet50 without augmentation, has an F1-score of 0.8515 and an AUC of 0.8638, indicating competitive performance in identifying benign and malignant lesions. Notably, data augmentation did not increase the ResNet50 model's performance, demonstrating the difficulty of applying broad computer vision techniques to specific medical imaging applications. The study investigates the performance of various deep learning algorithms for binary skin lesion classification and explores their possible clinical applications. While the findings are promising, there are still issues with model interpretability and generalizability. This study contributes to the developing field of AI-assisted dermatology by demonstrating the potential of automated systems to improve the essential task of distinguishing between benign and malignant skin lesions (Brinker et al., 2019).

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. It has been undertaken in accordance with the University research ethics standards, by the terms of permit number : 68888

Signed : Paramesh kumar Bukya

Date: 02/10/2024

Acknowledgement:

I'd want to thank everyone who helped me complete this dissertation.

First, I'd want to thank my supervisor, Adrian Davison, for his direction and assistance during this project.

Thank you to Manchester Metropolitan University's Human Centred Computing staff and faculty for creating such an excellent learning environment.

I thank the International Skin Imaging Collaboration (ISIC) for supplying the dataset that enabled this research.

Thank you to my classmates and friends for your encouragement and constructive talks.

Finally, I want to thank my family for their unwavering support and understanding throughout this difficult time.

This task would not have been possible without you all. Thank you.

Paramesh kumar Bukya

02/10/2024

Abbreviations:

AI - Artificial Intelligence

AUC - Area Under the Curve

CNN - Convolutional Neural Network

FN - False Negative

FP - False Positive

GAN - Generative Adversarial Network

ISIC - International Skin Imaging Collaboration

ISBI - International Symposium on Biomedical Imaging

LRP - Layer-wise Relevance Propagation

ResNet - Residual Network

RGB - Red, Green, Blue

ROC - Receiver Operating Characteristic

SVM - Support Vector Machine

TN - True Negative

TP - True Positive

VGG - Visual Geometry Group

1 Chapter : Introduction

1.1 Project Overview

The rapid development of artificial intelligence (AI) and machine learning technologies has created new opportunities for improving medical diagnostics in a variety of sectors. In dermatology, the use of these technologies to classify skin lesions offers particular promise for improving early identification and diagnosis of skin malignancies, including melanoma (Esteva et al. 2017).

This research focuses on the creation and testing of deep learning models for classifying skin lesions as benign or malignant. Using the International Skin Imaging Collaboration (ISIC) 2016 challenge dataset (ISIC Challenge, 2016), this study implements and evaluates different cutting-edge deep learning architectures to determine their performance in distinguishing benign and malignant skin lesions based on dermoscopic images.

The importance of this research is its potential to enhance the classification between Malignant and Benign . This project seeks to give a valuable tool to dermatologists in their diagnostic process by building accurate and efficient classification methods, which could lead to early diagnosis and better patient outcomes.

1.2 Aims and Objectives

The major goal of this study is to construct and test deep learning models for accurately classifying skin lesions as benign or malignant using dermoscopic pictures.

1. The study's aims include preprocessing and analysing the ISIC 2016 challenge dataset(ISIC Challenge, 2016) to resolve class imbalance and image variability.
2. To compare the performance of various deep learning architectures (ResNet18 (He et al., 2016), VGG16 (Simonyan and Zisserman, 2014), InceptionV3 (Szegedy et al., 2016), ResNet50 (He et al., 2016), and EfficientNetB0) for skin lesion categorization.

3. Determine the effect of data augmentation approaches on model performance.

4. Evaluate the models using measures such as accuracy, sensitivity, specificity, F1-score, and AUC.
5. Analyse the strengths and limitations of the constructed models and compare them to current literature.
6. Provide insights on clinical usefulness of established models and recommendations for future research in this field.

1.3 Potential Problems

Several problems are expected during the course of this research:

1. Dataset limitations: Although valuable, the ISIC 2016 challenge dataset (ISIC Challenge, 2016) is tiny in comparison to other computer vision challenges. The dataset contains 900 training images and 379 test images(Gutman et al., 2016), which may limit the trained models' ability to generalize.
2. Class Imbalance: The dataset has a considerable imbalance of benign and malignant cases. This mismatch, while reflecting real-world distributions, poses issues for model training and evaluation (Brinker et al. 2019).
3. Image Variability: Dermoscopic images can vary greatly depending on lighting, image quality, and artifacts like hair or air bubbles. Tschandl et al. (2018) emphasize the dataset's multi-source character, which causes variability in image acquisition technologies.
4. Model Interpretability: Deep learning models are typically "black boxes," making it difficult to grasp their decision-making process. This lack of interpretability can pose a significant hurdle to clinical implementation (Holzinger et al., 2017).
5. Overfitting Risk: Deep learning models' limited dataset size and complexity can lead to overfitting, where they perform well on training data but fail to generalize to new examples.
6. Ethical Considerations: The application of AI in medical diagnosis involves ethical concerns about duty, liability, and patient confidence. These difficulties must be carefully evaluated within the context of this research (Adamson and Smith, 2018).

1.4 Tools

This research uses a variety of tools and technologies:

- Programming Language: Python was chosen due of its robust ecosystem of machine learning libraries.
- Deep Learning Framework: PyTorch, chosen for its flexibility and dynamic computing graph.
- Image Processing: The OpenCV and Pillow libraries are used to manipulate and augment images.
- Data analysis and visualization: NumPy, Pandas, and Matplotlib are utilized for data manipulation and result visualization.
- Model Architectures: Pre-trained models from torchvision, such as ResNet, VGG, InceptionV3, and EfficientNetB0 .
- Hardware: NVIDIA GPU for faster model training.

1.5 Report Structure

The remainder of this dissertation is structured as follows:

Chapter 2: Literature Review – This chapter provides a thorough examination of the current body of literature in the field of automated skin lesion classification. This section examines the transition from conventional machine learning to deep learning techniques and examines the most recent state-of-the-art methods.

Chapter 3: Data Collection - This chapter provides a comprehensive overview of the ISIC 2016 challenge dataset(ISIC Challenge, 2016) that was employed in this investigation. The dataset's characteristics, such as the distribution of classes, image properties, and any preprocessing steps that were implemented on the original data, are described.

Chapter 4: Experimental Methodology - This chapter delineates the methodology that was implemented in this investigation. It includes the evaluation metrics, model architectures, training procedures, and data preprocessing techniques that are employed to evaluate the performance of the model.

Chapter 5: Experimental Results - This chapter summarizes the findings of the experiments

that were conducted. It offers a comprehensive examination of the performance of each model, comparing their efficacy in the classification of skin lesions using a variety of evaluation metrics.

Chapter 6: Further Study - This chapter addresses the limitations of the present investigation and proposes directions for future research. It examines potential enhancements to the models, suggestions for datasets that are larger and more diverse, and factors that should be taken into account when implementing them in clinical settings.

Chapter 7: Conclusion - This chapter highlights significant research findings, examines their implications for skin lesion classification, and reflects on the study's overall contribution to AI in dermatology.

2 Chapter : Literature Review

2.1 Introduction

Skin cancer is a major global health issue, with melanoma identified as one of its most severe variants. In 2021, the American Cancer Society estimated that roughly 106,110 new melanoma cases would be diagnosed in the United States, with around 7,180 anticipated deaths from the disease (Siegel et al., 2021). Timely identification is essential for enhancing survival chances; yet the visual assessment of skin lesions can be difficult even for seasoned dermatologists. Research indicates that the precision of clinical melanoma diagnosis varies between 65% and 80%, with an enhancement to 75-84% when utilizing dermoscopy (Kittler et al., 2002).

The difficulty in clinical diagnosis arises from the complex visual distinctions between benign and malignant lesions, along with the extensive range of appearances. Dermatologists generally utilize heuristic approaches such as the ABCDE criteria (Asymmetry, Border irregularity, Colour variegation, Diameter >6mm, and Evolution) or the 7-point checklist for melanoma diagnosis (Argenziano et al., 1998). Nevertheless, these procedures are subjective and may result in considerable inter-observer variability.

Nowadays, there has been an increasing interest in the development of automated computer-aided diagnosis systems utilizing deep learning and convolutional neural networks (CNNs)

for skin lesion classification assistance. These systems seek to deliver objective, consistent, and perhaps more precise diagnoses to enhance clinical decision-making. This literature review analyses the latest developments in deep learning methodologies for the study and classification of skin lesions, prioritizing melanoma detection.

2.2 Datasets and Challenges

Large, varied, publicly available datasets are essential for the development of strong skin lesion classification systems. By holding challenges and offering standardized datasets, the International Skin Imaging Collaboration (ISIC) (ISIC Challenge, n.d.) has been helpful in this regard:

- More than 23,000 dermoscopic images are available in the ISIC Archive (Codella et al., 2018; Tschandl et al., 2018).

- Standardized datasets and evaluation measures were made available by the ISIC challenges in 2016, 2017, and 2018 (ISIC Challenge, n.d.).

- More than 10,000 training images from seven different diagnostic categories were included in the 2018 challenge dataset (Tschandl et al., 2018).

The Interactive Atlas of Dermoscopy, the Dermofit Image Library, the PH2 Dataset, and the HAM10000 Dataset are other relevant datasets that have been referenced in the literature (Hosny et al., 2019; Tschandl et al., 2018). Because they allow direct comparison of various approaches, these publicly available datasets and challenges have accelerated progress in the field.

It is crucial to remember that the datasets that are currently available have limits. Issues including class imbalance and a lack of variation in skin tones and lesion kinds are brought to light by Harangi (2018) and Albahar (2019). This lack of representation may result in biased models that perform poorly on underrepresented populations. More balanced and diversified datasets that more accurately represent the world's population and the range of skin lesion manifestations are obviously needed as the field advances.

These datasets have significantly increased in size over time. The ISIC 2019 dataset is the

largest and most diversified dataset currently accessible, according to Goyal et al. (2020), with 25,331 photos in 8 diagnostic categories. This increase in dataset size and diversity has been critical for developing the field and boosting the generalization capabilities of deep learning algorithms.

2.3 Deep Learning Architectures

The research indicates a distinct trend in employing deep convolutional neural networks (CNNs) for the categorization of skin lesions. Popular CNN architectures utilized includes VGGNet, ResNet, Inception/GoogLeNet, and DenseNet (Hosny et al., 2019; Tschandl et al., 2018; Harangi, 2018; Esteva et al., 2017). These architectures have demonstrated significant effectiveness in numerous computer vision tasks, and their utilization in skin lesion classification has produced encouraging outcomes.

Transfer learning has become a crucial method to address the challenges of limited skin lesion datasets. This method entails employing CNNs that have been pre-trained on extensive natural image datasets such as ImageNet and subsequently fine-tuning them on skin lesion datasets (Hosny et al., 2019; Esteva et al., 2017). Romero Lopez et al. (2017) illustrated the efficiency of this method, attaining enhanced sensitivity (78.66% vs 53.3%) by fine-tuning a pre-trained VGG16 network rather than training from the ground up.

Ensemble approaches, which integrate many CNN models, have demonstrated potential. Harangi (2018) attained enhanced outcomes by integrating outputs from the VGG-16, ResNet50, ResNetX, and InceptionV3 networks. This method solves the failings of singular models and enhances overall robustness. Nonetheless, it is important to acknowledge that ensemble approaches can substantially elevate computational complexity, thus restricting their practical use in resource-limited environments.

An interesting advancement in the domain is the investigation of hybrid methodologies that integrate deep learning along with traditional machine learning strategies. Kassem, Hosny, and Fouad (2020) discovered that the integration of ResNet50 features with Support Vector Machine (SVM) classification outperformed end-to-end CNN training, with an accuracy of 99.87% on their dataset. This hybrid method utilizes the robust feature extraction skills of CNNs alongside the efficacy of SVMs for classification tasks with constrained training data.

Although these architectures demonstrate remarkable performance, it is essential to critically assess their practical usefulness. Many studies assess models using selected datasets, which may not adequately represent the complexities of actual practice. Moreover, the opaque characteristics of deep learning models provide difficulties for interpretability and explainability, which are essential in medical applications.

2.4 Data Preprocessing and Augmentation

Given the small size of most skin lesion datasets, data augmentation is commonly utilized to artificially enlarge training sets. Geometric transformations (rotation, flipping, and scaling) are common techniques, as is colour jittering and the addition of Gaussian noise. Perez et al. (2018) conducted a comprehensive review of 13 distinct augmentation approaches and discovered that combining multiple ways was more effective than utilizing a single technique.

Preprocessing processes have also been demonstrated to play an important impact in model performance. These include lesion segmentation, colour space modifications, pixel value normalization, and hair removal (Hosny et al., 2019; Kasmi & Mokrani, 2016). Kasmi and Mokrani (2016) illustrated the value of preprocessing by constructing an automatic ABCD rule system, which included morphological filtering for hair removal and adaptive histogram equalization for contrast enhancement.

While these preprocessing and augmentation strategies have proven beneficial, it is critical to recognize their limitations. Excessive augmentation may cause artifacts or distort clinically significant characteristics. Furthermore, the efficacy of these strategies may differ based on the dataset's specific characteristics and target lesion types. Future research could benefit from a more systematic assessment of preprocessing and augmentation procedures across various datasets and model architectures.

Goyal et al. (2020) emphasize the necessity of removing artifacts in dermoscopic images, such as air bubbles, hair, and uneven illumination. They examine a variety of preprocessing techniques, including colour constancy algorithms such as Shades of Gray and max-RGB,

which can aid in the normalization of photos captured under varying lighting circumstances and devices. These methods are critical for increasing the resilience and generalizability of deep learning models across a variety of clinical contexts.

2.5 Feature Extraction and Classification Approaches

While end-to-end CNN training is popular, some researchers have experimented with hybrid systems that blend deep learning and classical machine learning techniques. These include employing CNNs as feature extractors followed by classical classifiers such as SVMs, or mixing deep features with hand-crafted features based on clinical criteria such as the ABCD rule (Esteva et al., 2017; Kassem et al., 2020; Kasmi and Mokrani, 2016).

These hybrid techniques have produced impressive results, frequently outperforming pure deep learning models. For example, Kassem et al. (2020) combined ResNet50 characteristics with SVM classification to get an accuracy of 99.87%. However, it is important to note that the performance of these hybrid models can be strongly influenced by the unique dataset and feature engineering decisions. More study is needed to evaluate whether these methods work effectively across diverse datasets and lesion kinds.

Saeed and Zeebaree (2021) provide a detailed evaluation of deep learning algorithms for skin lesion classification. They cover different feature extraction strategies, such as using pre-trained CNNs as feature extractors and combining deep features with traditional hand-crafted features. They emphasize that hybrid approaches frequently outperform single-method approaches, emphasizing the potential benefits of mixing deep learning with domain-specific knowledge.

2.6 Incorporation of Metadata and Clinical Information

While most studies concentrate on image data, some researchers have looked into the use of clinical metadata to improve classification performance. Kasmi and Mokrani (2016) developed an autonomous ABCD rule system that extracted features related to asymmetry, border irregularity, colour variation, and dermoscopic structures. They attained a 94.0% accuracy on their dataset by using these medicinally significant variables.

Tschandl et al. (2018) and Esteva et al. (2017) suggested that future research include patient metadata such as age, sex, lesion site, and patient history. They suggested that this additional information could aid in categorization accuracy, particularly in difficult circumstances.

The inclusion of clinical metadata opens up the possibility of developing more complete and therapeutically appropriate models. However, it also raises concerns about data privacy, metadata consistency across different therapeutic settings, and the possibility of introducing demographic biases. When adding new patient data, future research should take these ethical and practical concerns into account.

Goyal et al. (2020) emphasize the necessity of combining clinical metadata and patient history into skin lesion classification algorithms. When making diagnosis, dermatologists take into account patient age, gender, family history, and lesion evolution. According to the scientists, future AI systems should attempt to include this contextual knowledge in order to increase diagnosis accuracy and give more clinically relevant results.

2.7 Performance Evaluation and Comparison to Dermatologists

The literature uses a variety of metrics to assess classification performance, including accuracy, sensitivity, specificity, area under the ROC curve (AUC), and average precision (Hosny et al., 2019; Codella et al., 2018). The 2018 ISIC Challenge introduced balanced accuracy to address class imbalance (Tschandl et al., 2018).

Algorithm efficacy has been compared to that of dermatologists in numerous studies, with numerous studies reporting results that are either comparable or superior. Esteva et al. (2017) discovered that their CNN matched or outperformed the performance of 21 board-certified dermatologists on two binary classification tests. In a study conducted by Haenssle et al. (2018), the algorithm demonstrated an improved sensitivity (95% vs 86.6%) and specificity (63.8% vs 71.3%) in the detection of melanoma when compared to 58 dermatologists.

While the results are promising, they must be interpreted with caution. Most comparisons were made using curated datasets, which may not fully reflect real-world clinical settings.

Furthermore, these comparisons fail to adequately represent the importance of contextual knowledge and clinical experience in dermatologists' decision-making processes. Future research should focus on prospective, real-world evaluations to gain a better understanding of AI systems' potential influence in clinical practice.

Tschandl et al. (2019) conducted a large-scale study, comparing the performance of 157 algorithms to 511 human readers. They discovered that the best-performing algorithms surpassed human specialists in classifying skin lesions. This study gives a more complete look at the differences in performance between AI and humans. However, it also shows that more research is needed to find out how these methods can be used in real life and in clinical settings.

2.8 Challenges and Limitations

Despite the encouraging outcomes described in the literature, various obstacles and limits have been identified:

- Limited diversity in datasets, especially for individuals with darker skin tones (Harangi, 2018; Albahar, 2019).

- Datasets show a class imbalance, with benign lesions being overrepresented (Tschandl et al., 2018; Hosny et al., 2019).

- Difficulty comprehending CNN choices limits clinical acceptance (Albahar, 2019; Esteva et al., 2017).

- Overfitting may occur because to artifacts or biases in training data (Tschandl et al., 2018; Albahar, 2019).

- Accurate segmentation of lesion margins is challenging (Codella et al., 2018; Kasmi and Mokrani, 2016).

- Variability in image acquisition impacts generalization (Albahar, 2019; Kassem et al., 2020).

- Larger, diversified datasets are needed to increase performance on rare subtypes (Tschandl et al., 2018; Harangi, 2018).

These problems underline the importance of continuing research and development in the sector. Addressing concerns of dataset variety and representativeness is especially important in ensuring that AI systems perform equally across ethnic groupings. Furthermore, increasing model interpretability and robustness to real-world variability will be critical for gaining clinical acceptance and trust.

Goyal et al. (2020) address additional problems, such as the "black box" aspect of deep learning models, which makes it difficult for dermatologists to understand and trust the decision-making process. They also emphasize the importance of prospective clinical validation studies to determine the real-world impact of AI-assisted diagnosis on patient outcomes.

2.9 Challenges in Real-World Applications

Deep learning models' "black box" nature poses a substantial hurdle to their use in clinical applications. While these models can reach great accuracy, their decision-making process is frequently opaque, which can be problematic in medical applications that require interpretability.
To solve this, scholars are investigating several ways to enhance the interpretability of deep learning models.

1. Attention Mechanisms: These strategies highlight areas of a picture that the model considers when making a decision. For instance, a et al. (2019) used an attention-based architecture to show the most significant regions for melanoma identification.

2. Layer-wise Relevance Propagation (LRP): This technique analyses how each input pixel contributes to the final prediction. Zuo et al. (2020) used LRP to train a CNN for

skin lesion categorization, resulting in pixel-level explanations for the model's judgments.

3. Saliency Maps: These show the gradient of the output image compared to the input image, indicating the pixels with the most influence on the model's conclusion. Young et al. (2020) utilized saliency maps to provide visual explanations for their skin lesion categorization algorithm.

While these techniques indicate progress toward more interpretable models, there are still obstacles in turning these visualizations into clinically useful insights. Future studies should concentrate on closing the gap between model interpretability and clinical decision-making procedures.

2.10 Ethical and Social Concerns

The creation and implementation of AI systems for skin lesion classification has significant ethical and social concerns. These include:

1. Potential biases in AI systems trained on non-representative datasets may worsen disparities in healthcare.
2. Concerns about privacy when collecting and using massive medical picture databases. Over-reliance on AI systems may result in healthcare workers losing skills or missing diagnosis in the event of failures.
3. Clear rules are needed for integrating AI technologies into healthcare operations and decision-making.
4. Providing fair access to AI-assisted diagnostic tools in various healthcare settings and geographies.

In order to develop AI systems that are not only accurate but also fair, transparent, and beneficial to all patient populations, future research in this field should explicitly address these ethical considerations.

2.11 Future Directions

According to the evaluated literature, significant areas of future research include:

1. Creating datasets that are more representative of a broader spectrum of skin tones, lesion types, and imaging conditions, as well as those that are more diverse and balanced (Harangi, 2018; Tschandl et al., 2018).

2. Enhancing model interpretability with strategies like attention mechanisms and layer-wise relevance propagation (Albahar, 2019; Esteva et al., 2017).
   Combining clinical metadata and patient history with imaging data (Tschandl et al., 2018; Kassem et al., 2020).

3. Investigating multimodal procedures that integrate clinical, dermoscopic, and histopathological images (Tschandl et al., 2018).
   Creating effective preprocessing and augmentation methods to enhance generalization (Hosny et al., 2019; Perez et al., 2018).

4. Utilizing external validation sets and standardizing evaluation metrics to more accurately evaluate real-world performance (Tschandl et al., 2018).
   Conducting clinical trials to test the effectiveness of AI-assisted diagnosis on patient outcomes (Esteva et al., 2017; Haenssle et al., 2018).

Saeed and Zeebaree (2021) underline the significance of creating end-to-end AI systems capable of managing the complete diagnostic pipeline, from image acquisition to final diagnosis and treatment suggestion. They propose that future research should focus on developing more holistic systems capable of integrating several data sources and providing dermatologists with comprehensive support.

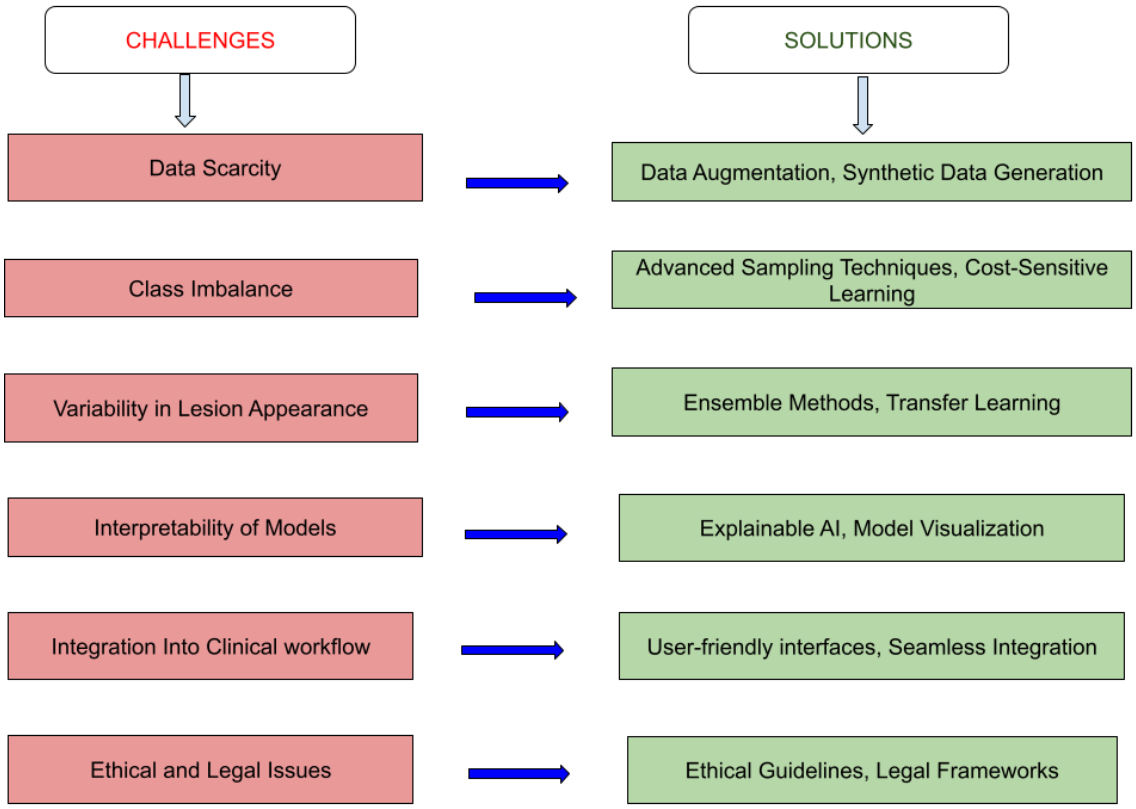


Figure 2-1 Roadmap of Challenges and Solutions for Deep Learning in Skin Lesion Classification

Figure 2.1 emphasizes the interconnectedness of present difficulties and prospective solutions, underlining the importance of taking a holistic approach to improving the profession. To address these problems, computer scientists, dermatologists, and healthcare professionals must work together to ensure that AI systems are not only accurate, but also clinically relevant and ethical.

2.12 Conclusion

Deep learning algorithms, notably CNNs, have made significant advances in automated skin lesion classification. Many studies show that performance is comparable to or better than that of expert dermatologists. Transfer learning, data augmentation, and ensemble approaches were critical to reaching these achievements. However, challenges persist in assuring varied representation in datasets, model interpretability, and application to real-world medical situations.

The incorporation of clinical metadata and adherence to recognized diagnostic criteria, such

as the ABCD rule, present potential opportunities for increasing model performance and clinical relevance. Future research should concentrate on creating larger datasets, investigating multimodal techniques, and undertaking rigorous clinical validation studies.

As the field evolves, larger and more diversified datasets, standardized evaluation methodologies, and prospective clinical validation will be required. The creation of interpretable AI models capable of explaining their conclusions will be critical to winning clinician trust and approval.

The potential influence of artificial intelligence-assisted diagnostics in dermatology is enormous. These devices, which provide speedy, accurate, and objective assessments, could help shorten the diagnostic process, eliminate unnecessary biopsies, and enhance early detection rates. This is especially significant in areas with limited access to dermatologists, where AI systems could act as an effective triage tool.

However, it is critical to note that the purpose of AI in dermatology is not to replace dermatologists, but rather to enhance existing capabilities.

3 Chapter Data Collection

3.1 Introduction

The quality and usefulness of the dataset serve as the basis for this skin lesion classification study. This chapter discusses the characteristics of the dataset given by the International Skin Imaging Collaboration (ISIC) for the "Skin Lesion Analysis Towards Melanoma Detection" competition at the 2016 International Symposium on Biomedical Imaging (ISBI) (ISBI 2017, n.d.). The dataset's collection process, structure, and preprocessing steps are all explained to provide a thorough knowledge of the data used in this study.

3.2 The ISIC 2016 Challenge Dataset

The ISIC 2016 challenge dataset is a subset of the larger ISIC Archive, which at the time of the challenge included over 10,000 dermoscopic images collected from prominent clinical centres across the world (Gutman et al., 2016). The dataset was specifically curated for the competition to aid in the development of automated algorithms for melanoma detection from dermoscopic pictures.

3.2.1 Dataset Overview

The ISIC 2016 challenge dataset comprises 1,279 high-quality dermoscopic images of skin lesions, divided into three subsets:

1. Training Set: 900 images
2. Test Set: 379 images

These images were carefully selected to represent a clinically relevant sample, ensuring a balance between benign and malignant cases (Gutman et al., 2016).

3.2.2 Image Characteristics

The images in the collection are saved in JPEG format, which strikes a compromise between image quality and file size. Each image is allocated a unique identity using the naming convention 'ISIC_.jpg', where is a seven-digit integer. This systematic labelling allows for easy tracking and control of the photos throughout the research process.

Dermoscopic pictures typically have a 15-30mm field of view at 10X magnification, allowing for thorough visualization of skin lesions (Gutman et al., 2016).

3.2.3 Ground Truth Labels

The image data is accompanied by a ground truth file that contains two key bits of information:

1. Image Identifier: Matches the image file name.
2. Classification: 'Benign' or 'Malignant'.

The ground truth labels were developed using a thorough method that included expert consensus and pathology report data, assuring high classification reliability (Gutman et al., 2016).

3.2.4 Sample Images

To provide a visual understanding of the dataset, Figure 3.1 presents a selection of sample images from the ISIC 2016 challenge dataset.

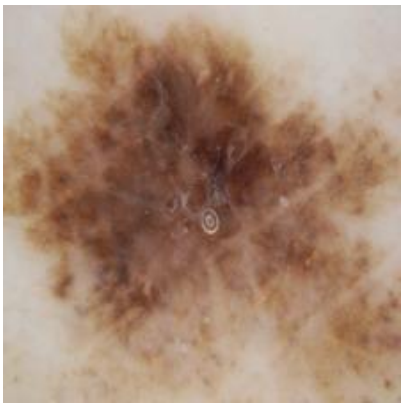


Figure 3-1 Malignant melanoma with irregular borders

.

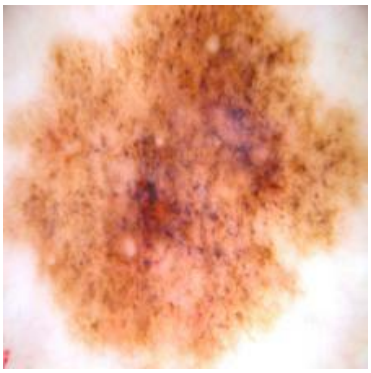


Figure 3-2Malignant melanoma with colour variegation.

Figure 3-3 Benign lesion with visible hair.

Figure 3-4 Benign nevus with regular shape

Figure 3.1, Figure 3.2, Figure 3.3, Figure 3.4 are sample images from the ISIC 2016 challenge dataset (ISIC Challenge, 2016)

These sample images illustrate the variety and complexity of skin lesions present in the dataset:

Figure 3.1 depicts a malignant melanoma with highly uneven boundaries and an asymmetrical shape. The lesion has a varying color pattern, ranging from light brown to darker parts, which is commonly associated with melanoma.

Figure 3.2 is another malignant melanoma with substantial colour variegation. The

appearance of numerous colours, including light brown, dark brown, and reddish patches, is an important trait that dermatologists look for when determining the malignancy of a lesion.

Figure 3.3 shows a benign tumour with visible hair strands crossing over it. Hair can sometimes make it difficult to analyse skin lesions, both for human experts and automated systems. Despite this, the lesion appears to have a more uniform border and less colour variation than the malignant instances.

Figure 3.4 benign nevus (mole) appears as a tiny, spherical lesion with a regular shape and very homogeneous pigmentation. The well-defined margins and symmetrical appearance are typical of benign lesions.

These examples demonstrate the challenging nature of skin lesion classification, even trained professionals find it difficult to classify skin lesions, as seen by these cases. The complex distinctions between benign and malignant tumours highlight the potential utility of automated categorization systems in assisting dermatologists with their diagnoses. Notably, while some characteristics, such as uneven boundaries and colour variegation, are more common in malignant lesions, the presence or absence of a single trait is not conclusive. This complexity emphasizes the need of taking into account many factors throughout the classification process.

The inclusion of photos with changing quality and characteristics, such as the existence of hair or different lighting conditions, implies that developed algorithms must be resilient to real-world variations. This dataset diversity is critical for developing training models that can generalize well to new, previously unknown clinical cases.

3.3 Data Collection Process

While raw image information was not collected specifically for this study, understanding the gathering method is critical for recognizing the dataset's strengths and limitations.

3.3.1 Image Acquisition

The ISIC dataset contains dermoscopic images from several clinical centres throughout the world, assuring a diverse representation of skin types, lesion characteristics, and imaging conditions. The photos were captured using a variety of devices at each centre, representing the variability seen in clinical settings (Gutman et al., 2016).

Dermoscopy, a non-invasive skin imaging technology, provides thorough viewing of the skin's underlying structures, improving diagnosis accuracy for skin lesions (Esteva et al., 2017). This approach minimizes surface reflection, allowing deeper layers of the skin to be seen more clearly.

3.3.2 Expert Annotation

Following picture acquisition, a team of experienced dermatologists analysed each image and made a diagnosis. In cases of uncertainty or dispute, histological examination was performed to confirm the diagnosis. This multistep verification procedure assures that the ground truth labels are of high quality and reliable (Gutman et al., 2016).

3.3.3 Privacy and Quality Assurance

Prior to inclusion in the dataset, all photos were checked for privacy and quality assurance. The accompanying clinical metadata was reviewed by melanoma experts to ensure the dataset's reliability and clinical relevance (Gutman et al., 2016).

3.4 Dataset Analysis

3.4.1 Class Distribution

An analysis of the class distribution in the dataset revealed:

- Total samples: 1,279
- Training set: 900 images
    - Benign samples: 727
    - Malignant samples: 173
- Test set: 379 images

- Benign samples: 303
- Malignant samples : 76

This analysis showed the occurrence of a class imbalance, which is consistent with the real-world distribution of benign and malignant tumours. Because of the imbalance, proper procedures had to be used during model training to avoid bias toward the majority class.

3.4.2 Image Characteristics

The images in the collection were analysed for a variety of factors, including colour distribution, contrast levels, and lesion size. This investigation served to inform the preprocessing procedures and influenced the choice of data augmentation approaches.

3.5 Challenges and Limitations

Although the ISIC 2016 challenge dataset (ISIC Challenge, 2016) offers a strong foundation for skin lesion classification research, it is crucial to recognize its constraints:

1. Small sample size: The dataset, which contains 1,279 images, is relatively small in comparison to those employed in other computer vision tasks, which may restrict the generalization capabilities of models trained on this data.

2. Class imbalance: The substantial disparity between benign and malignant cases, while reflective of real-world distributions, poses significant challenges for model training and evaluation.

3. Inadequate demographic data: The dataset fails to furnish thorough demographic data regarding patients, which restricts the capacity to evaluate model performance across various population subgroups.

4. Single imaging modality: The dataset exclusively comprises dermoscopic images, which may not accurately reflect the diverse array of imaging techniques employed in clinical practice.

3.6 Conclusion

The ISIC 2016 challenge dataset (ISIC Challenge, 2016) is an excellent resource for skin lesion classification research. Its meticulous curation, expert annotations, and representation of real-world clinical data make it an ideal platform for building and testing automated melanoma diagnosis algorithms. However, researchers should be cautious of its limits when interpreting data and extrapolating findings to larger clinical applications.

The preprocessing procedures used on this dataset, such as scaling, normalization, and data augmentation, prepared it for fast model training while addressing some of its inherent problems. As the field of automated skin lesion analysis improves, future research could benefit from larger, more diverse datasets that overcome some of the constraints noted in this study.

4 Chapter: Experimental Methodology

4.1 Overview of the approach

This chapter describes the experimental methods used in the investigation of skin lesion categorization with deep learning algorithms. The approach employs cutting-edge convolutional neural networks (CNNs) and transfer learning to generate robust models capable of differentiating between benign and malignant skin lesions. Six distinct model architectures were examined:

1. ResNet18 (He et al., 2016) without augmentation
2. VGG16 (Simonyan and Zisserman, 2014) without augmentation
3. Inception v3 (Szegedy et al., 2016) without augmentation
4. ResNet50 (He et al., 2016) without augmentation
5. EfficientNet-B0 (Tan and Le, 2019) without augmentation
6. ResNet50 (He et al., 2016) with augmentation

The methodology can be broadly classified into the following major components:

1. Data preprocessing and augmentation
2. Model architectures and transfer learning
3. Training process and optimization
4. Evaluation metrics and performance assessment

This chapter provides a full overview of the experimental approach, which serves as the foundation for comprehending and interpreting the results reported in Chapter 5.

4.2 Data Preprocessing and Augmentation

4.2.1 Dataset Organization

The ISIC (International Skin Imaging Collaboration) dataset (ISIC Challenge, 2016) was utilized for the experiments. The dataset was organized into two main folders:

1. Train folder: Used for training and validation
2. Test folder: Used for final model evaluation

To enable reliable model training and evaluation, the researchers divided the training folder 80-20, resulting in separate training and validation sets. This split was stratified to preserve the original class distribution in both groups, which was critical for dealing with the inherent class imbalance in skin lesion datasets. Figure 4.1 shows code for the splitting of train data into validation set.

```
# Train-validation split
train_ratio = 0.8
train_size = int(train_ratio * len(train_dataset))
val_size = len(train_dataset) - train_size
```

Figure 4-1splitting of train data into validation set.

4.2.2 Image Preprocessing

Preprocessing was an essential component in preparing the dataset for input into deep learning models. The photos in the ISIC dataset differ in dimensions and quality, necessitating their resizing to a uniform input size for each model. For ResNet-based models, VGG16, and EfficientNet-B0, the input dimensions were configured to 224x224 pixels, whereas InceptionV3 necessitated input dimensions of 299x299 owing to its architecture. The resizing procedure guaranteed that all photos could be input into the model architectures without mistakes. Normalization was implemented to adjust pixel values to a range of [0, 1], which is crucial for the convergence of neural networks. Figure 4.2 code snippet the image preprocessing for Resnet based models, VGG16 and EfficientNet-B0 and Figure 4.3 shows code for the image processing of InceptionV3

```
# Data Preprocessing
train_transform = transforms.Compose([
    transforms.Resize((224, 224)),
    transforms.ToTensor()  # No normalization or augmentation
])

test_transform = transforms.Compose([
    transforms.Resize((224, 224)),
    transforms.ToTensor()  # No normalization or augmentation
])
```

Figure 4-2image preprocessing for Resnet based models, VGG16 and EfficientNet-B0

```
# Data Preprocessing
train_transform = transforms.Compose([
    transforms.Resize((299, 299)),  # Inception v3 expects 299x299 images
    transforms.ToTensor()  # No normalization or augmentation
])

test_transform = transforms.Compose([
    transforms.Resize((299, 299)),
    transforms.ToTensor()  # No normalization or augmentation
])
```

Figure 4-3 image processing of InceptionV3

4.2.3 Handling class imbalance

A significant problem of this dataset was the class imbalance between benign and malignant instances. Malignant instances, while essential for detection, are underrepresented, perhaps leading to a model biased towards predicting benign lesions. A weighted loss function was utilized to address this issue. The weights assigned to each class were inversely related to the sample frequency in the dataset, enabling the algorithm to prioritize accurate predictions of malignant instances. This facilitated a more equitable performance across both categories. Figure 4.4 shows handling class imbalance for all the models used in this study.

```
# Calculate class weights (inverse of class frequencies)
class_weights = [total_samples/class_count for class_count in class_counts]
class_weights = torch.FloatTensor(class_weights).to('cuda')
```

Figure 4-4 handling class imbalance

4.2.4 Data Augmentation

For the ResNet50 model with augmentation, supplementary data augmentation techniques were employed to artificially enhance the variety of the training dataset. The techniques encompassed the following:

1. Random Horizontal Flips: Images were subjected to horizontal flipping with a frequency of 50%.

2. Random Rotations: Images were subjected to random rotations within the range of -10 to 10 degrees.

3. Random Zoom: Images were magnified or reduced by up to 10%.
   Brightness Modifications: The brightness of images was randomly altered within a certain range.

These augmentation strategies facilitate the model's development of invariance to these transformations, hence enhancing its capacity to generalize to novel, unseen data (Shorten and Khoshgoftaar, 2019). Figure 4.5 shows Data Augmentation used in ResNet50 model.

```
# Data Preprocessing with Augmentation for the training set
train_transform = transforms.Compose([
    transforms.Resize((224, 224)),
    transforms.RandomHorizontalFlip(),
    transforms.RandomRotation(10),
    transforms.ColorJitter(brightness=0.2, contrast=0.2, saturation=0.2, hue=0.1),
    transforms.ToTensor()
])

test_transform = transforms.Compose([
    transforms.Resize((224, 224)),
    transforms.ToTensor()  # No data augmentation for test data
])
```

Figure 4-5 Data Augmentation for ResNet50 model

4.3 Model Architectures

Five distinct model architectures without augmentation and one with augmentation were tested, each selected for its shown performance in image classification tasks:

4.3.1 ResNet18

ResNet (Residual Network) architectures, which were introduced by He et al. (2016), were chosen for their capacity to address the vanishing gradient problem by utilizing residual connections. Resnet18 is one of the shallower variants of the residual network family. These connections enable the model to acquire identity mappings, which facilitates the training of

extremely deep networks. The 18 layers of convolutional blocks in ResNet18 are followed by batch normalization and ReLU activation. The preservation of information and the facilitation of gradient flow during backpropagation are facilitated by skip connections between every few layers.

ResNet18 was pre-trained on ImageNet and fine-tuned for binary classification of skin lesions for the purpose of this project. The final entirely connected layer was replaced with a layer that generates two probabilities, one for each of the benign and malignant classes. The model was able to adapt to the ISIC dataset and capitalize on the features it had acquired from ImageNet through the implementation of transfer learning. Figure 4.6 shows the implementation of ResNet18 in this study:

```
# Model: Use pre-trained ResNet18
model = torchvision.models.resnet18(pretrained=True)
model.fc = nn.Linear(model.fc.in_features, 2)
```

Figure 4-6 implementation of ResNet18

4.3.2 VGG16

Simonyan and Zisserman (2014) developed VGG16, a deep convolutional neural network with 16 layers mostly made up of 3x3 convolutional filters and max-pooling layers. VGG16 is noted for its simplicity and depth, making it a popular model for image classification problems. The network's architecture is quite structured, with each convolutional layer followed by max-pooling and three completely linked layers at the final stage.

For this task, VGG16 was pre-trained on ImageNet and fine-tuned on the ISIC dataset. The final classification layer was replaced, yielding two classes appropriate for the binary classification problem. One disadvantage of VGG16 is the large number of parameters, which makes it computationally more expensive than comparable architectures such as ResNet. However, it still does well in tasks that require fine-grained picture classification, such as identifying skin lesion. Figure 4.7 shows the code for the implementation of VGG16:

```
# Model: Use pre-trained VGG16
model = torchvision.models.vgg16(pretrained=True)
model.classifier[6] = nn.Linear(model.classifier[6].in_features, 2)  #
```

Figure 4-7 implementation of VGG16:

4.3.3 Inception v3

Szegedy et al. (2016) introduced Inception v3, which is a complicated model that processes multi-scale features using inception modules. Each inception module applies multiple convolution filters (1x1, 3x3, and 5x5) to the same input, allowing the model to detect both small and large features in the same image. This makes InceptionV3 ideal for applications requiring both global and local features.

The architecture also incorporates auxiliary classifiers, which help to solve the vanishing gradient problem by introducing intermediary losses that encourage the network to learn relevant features early in the training process. In this research, InceptionV3 was pre-trained on ImageNet and fine-tuned on the ISIC dataset, with the final fully connected layer tweaked to produce two classes. Figure 4.8 shows the code for implementation of Inception v3:

```
# Model: Use pre-trained InceptionV3
model = torchvision.models.inception_v3(pretrained=True)
model.aux_logits = True   # Enable auxiliary outputs (specific to Inception)
model.fc = nn.Linear(model.fc.in_features, 2)  # Adjust for binary classification
model.AuxLogits.fc = nn.Linear(model.AuxLogits.fc.in_features, 2)  # Adjust auxiliary output
model = model.to('cuda')
```

Figure 4-8 implementation of Inception v3:

4.3.4 ResNet50

ResNet50 is a deeper version of the ResNet family, containing 50 layers was introduced byHe et al. (2016). The deeper architecture enables the model to learn more abstract and sophisticated properties. ResNet50 employs the same residual connection algorithm as ResNet18, but with additional convolutional layers and residual blocks. This allows the network to record a wider range of features, making it better suited to complicated classification of images tasks.

In this study, ResNet50 was trained on ImageNet before being fine-tuned on the ISIC dataset. The final layer was replaced to produce two classes, representing benign and malignant classifications. ResNet50's deeper design allows it to learn complicated patterns in data, but it also raises the risk of overfitting, particularly on a small dataset like ISIC. Figure 4.9 shows code for the implementation of ResNet50:

```
# Model: Use pre-trained ResNet50
model = torchvision.models.resnet50(pretrained=True)
model.fc = nn.Linear(model.fc.in_features, 2)
```

Figure 4-9 implementation of ResNet50:

4.3.5 EfficientNet-B0

Tan and Le (2019) introduced EfficientNet-B0, which is part of a family of models noted for making economical use of parameters and computational resources. EfficientNet adjusts the network width, depth, and resolution evenly with a compound scaling factor, allowing the model to attain cutting-edge performance with fewer parameters than older architectures such as ResNet or VGG.

EfficientNetB0, the smallest model in the EfficientNet family, was pre-trained on ImageNet and fine-tuned on the ISIC dataset. Despite its computational efficiency, EfficientNetB0 has demonstrated competitive performance on a variety of image classification applications, including medical imaging. The last layer was updated to generate two classes for the binary classification challenge. Figure 4.10 shows the code for implementation of EfficientNet-B0:

```
# Model: Use pre-trained EfficientNet-B0
model = torchvision.models.efficientnet_b0(pretrained=True)
model.classifier[1] = nn.Linear(model.classifier[1].in_features, 2)
model = model.to('cuda')
```

Figure 4-10 implementation of EfficientNet-B0:

4.3.6 ResNet50 with Augmentation

This ResNet50 variant was trained on the augmented dataset given in Section 4.2.4. The purpose was to determine the effect of data augmentation on model performance in the

context of skin lesion classification. Figure 4.11 shows code for the implementation of ResNet50:

```
# Model: Use pre-trained ResNet50
model = torchvision.models.resnet50(pretrained=True)
model.fc = nn.Linear(model.fc.in_features, 2)
```

Figure 4-11 implementation of ResNet50:

4.4 Transfer Learning and Fine-Tuning

Transfer learning was utilized for all models, utilizing weights pre-trained on the ImageNet dataset (Deng et al., 2009). This methodology enables the models to leverage information acquired from an extensive, varied image dataset, potentially enhancing their efficacy in the specific task of skin lesion categorization.

The transfer learning procedure incorporated the subsequent stages:

- Initializing with pre-trained weights: Each model was started using weights pre-trained on ImageNet.

- Freezing Initial Layers: The initial layers of each model were frozen to retain the low-level features acquired from ImageNet.

- Substituting the terminal layer: The last fully connected layer of each model was substituted with a new layer that outputs two probabilities, representing the benign and malignant classes.

- Fine-tuning: The unfrozen layers, generally the final layers, were refined using the ISIC dataset, enabling the model to adjust to the distinct features of skin lesion photos.

This method enables us to utilize the advantages of extensive pre-training while customizing the models to our particular task, which may enhance performance and decrease training duration (Yosinski et al., 2014).

4.5 Training Process

4.5.1 Optimization Algorithm

All models were trained utilizing the Adam optimizer (Kingma and Ba, 2014), selected for its capacity to manage sparse gradients and its effective convergence characteristics. The learning rate was established at 1e-5, a rather low value to facilitate the fine-tuning of the pre-trained weights without significantly modifying the learnt features. Figure 4.12 shows code for the optimizer used in all models:

```
# Optimizer
optimizer = optim.Adam(model.parameters(), lr=0.00001)
```

Figure 4-12 optimizer used in all models

4.5.2 Loss Function

A weighted cross-entropy loss was employed to rectify the class imbalance in the dataset. The below is the formula for calculating Loss Function (4.1):

$$L = \sum w_i * y_i * \log(p_i) \tag{4.1}$$

Where:

w_i represents the weight assigned to class i.

- y_i represents the actual label (0 or 1)
- p_i denotes the predicted probability for class i

The weights were determined as the inverse of the class frequencies in the training data, hence allocating greater significance to the underrepresented malignant class. Figure 4.13 shows the loss function used in all models:

```
# Weighted loss function
criterion = nn.CrossEntropyLoss(weight=class_weights)
```

Figure 4-13 loss function used in all models

4.5.3 Training Loop

The training approach for each model included the following steps:

1. Iterated through the training data in batches of size 64.
2. Performed a forward pass through the model.
3. Calculated the loss via the weighted cross-entropy function.
4. Propagated the gradients backward.
5. Updated the model parameters using the Adam optimizer.
6. Evaluated on the validation set after each epoch.
7. Saved the best model based on validation performance.

Each model underwent training for a maximum of 10 epochs, with early stopping employed to mitigate overfitting. Training was suspended if the validation loss failed to improve for two successive epochs.

4.5.4 Hardware and Software

All experiments were conducted using PyTorch (Paszke et al., 2019) on a CUDA-enabled NVIDIA GPU for accelerated training. The use of GPU acceleration allowed for efficient training of these complex models on the large ISIC dataset.

4.6 Evaluation Metrics

The subsequent evaluation metrics were utilized to thoroughly analyse the models' performance, as recommended by Hossin and Sulaiman (2015) for comprehensive evaluation of classification models:

4.6.1 Accuracy

Accuracy(EvidentlyAI, n.d.) assesses the overall correctness of the model's predictions, and it is calculated as follows (4.2)

$$Accuracy = \frac{(\mathrm{TP} + \mathrm{TN})}{(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})} \quad (4.2)$$

Where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

Although accuracy serves as a broad indicator of performance, it may be misleading in imbalanced datasets, hence requiring the incorporation of supplementary metrics.

4.6.2 Precision

Precision(EvidentlyAI, n.d.) measures the proportion of correctly identified malignant cases out of all cases predicted as malignant and the formula for calculating Precision is as follows (4.3):

$$Precision = \frac{TP}{(TP+FP)} \quad (4.3)$$

High accuracy signifies a minimal false positive rate, which is essential in medical settings to prevent unnecessary strain and treatments for patients.

4.6.3 Recall (Sensitivity)

Recall(EvidentlyAI, n.d.) measures the proportion of actual malignant cases that were correctly identified by the model and the formula to calculate Recall is as follows (4.4)

$$Recall = \frac{TP}{(\mathrm{TP} + \mathrm{FN})} \quad (4.4)$$

High recall is particularly important in skin lesion classification, as it ensures that as many malignant cases as possible are detected.

4.6.4 F1-Score

The F1-score(Deepchecks Community Blog, 2024) is the harmonic mean of precision and recall, providing a balanced measure of the model's performance and the formula for it is as follows (4.5)

$$F1 = 2 * \frac{(\mathrm{Precision} * \mathrm{Recall})}{(\mathrm{Precision} + \mathrm{Recall})} \quad (4.5)$$

This metric is particularly useful in imbalanced datasets, as it takes both false positives and false negatives into account. Figure 4.14 shows code for calculating scores of Precision, Recall and F1-score for all models in this study.

```
# Calculate precision, recall, F1-score,
precision = precision_score(y_true, y_pred, average='weighted')
recall = recall_score(y_true, y_pred, average='weighted')
f1 = f1_score(y_true, y_pred, average='weighted')
```

Figure 4-14 Precision, Recall and F1-score for all models

4.6.5 ROC-AUC Score

The Area Under the Receiver Operating Characteristic Curve (ROC-AUC) (Deepchecks Community Blog, 2024) assesses the model's ability to differentiate between classes at different thresholds. It plots the True Positive Rate versus the False Positive Rate at various classification levels.

A ROC-AUC of 1.0 indicates a perfect classifier, whereas 0.5 indicates random guessing. This statistic provides a thorough assessment of the model's discriminative capacity at various decision thresholds. Figure 4.15 shows code for calculating the Roc-auc:

```
#calculate roc_auc
roc_auc = roc_auc_score(y_true, y_scores)
```

Figure 4-15 Roc-auc calculation

4.7 Model Comparison and Selection

To establish the most successful model for skin lesion classification, all six models (five without augmentation and one with augmentation) were evaluated using the criteria indicated above. The comparison was mostly based on the models' performance on the validation set,

with the test set used to evaluate the best-performing model.
The selection criteria favoured models that demonstrated:

- High ROC-AUC scores indicate strong discriminatory ability.
- The F1-score indicates a balanced precision and recall level.
- Consistent performance across benign and malignant classes.

4.8 Conclusion

This chapter outlined the extensive approach utilized in the investigation of skin lesion classification by deep learning techniques. A comprehensive outline of data preparation, model selection, training methodologies, and evaluation measures has provided a robust framework for comprehending and analysing the experimental outcomes.

The application of transfer learning, coupled with methods to mitigate class imbalance and enhance data through augmentation, seeks to create robust models proficient in precise skin lesion classification. The various model designs examined provide a comprehensive comparison of various methodologies for this essential medical imaging task.

The subsequent chapter will describe and analyse the outcomes of these tests, offering insights into the comparative performance of various models and the effectiveness of the methodology in addressing the issues of skin lesion classification.

5 Chapter: Experimental Results

5.1 Introduction

This chapter provides a thorough review of the experimental data derived from using the methods described in Chapter 4 for skin lesion classification. The efficacy of six deep learning models is assessed: ResNet18, VGG16, InceptionV3, ResNet50, and EfficientNet-B0 without augmentation, and ResNet50 with augmentation. The investigation examines the models' proficiency in effectively differentiating between benign and malignant skin lesions utilizing the ISIC dataset.

5.2 Presentation of Results for Each Model

5.2.1 ResNet18 (Without Augmentation)

ResNet18 exhibited commendable performance on the ISIC dataset, attaining high validation accuracy and F1-score. The model utilized residual connections, allowing it to learn effectively from the limited dataset without overfitting.

- Best validation accuracy: 78.48%
- Best F1-score: 0.7938
- Best ROC-AUC: 0.7369
- Precision: 0.8018
- Recall: 0.7877

ResNet18 exhibited an effective balance between precision and recall, rendering it a dependable model for skin lesion classification.

5.2.2 VGG16 (Without Augmentation)

VGG16, recognized for its depth, also attained impressive outcomes. Nonetheless, its absence of residual connections rendered it more susceptible to overfitting in comparison to ResNet18. Nonetheless, the model demonstrated competitive performance, especially regarding precision.

- Best validation accuracy: 78.49%
- Best F1-score: 0.7867
- Best ROC-AUC: 0.7549

- Precision: 0.7926
- Recall: 0.7821

VGG16 had slightly poorer recall compared to ResNet18, suggesting it was more adept at detecting malignant cases (high precision) but missed a greater number of real malignant instances (lower recall).

5.2.3 InceptionV3 (Without Augmentation)

InceptionV3's multi-scale processing contributed to its strong performance on the ISIC dataset, especially for ROC-AUC, which reflects its effective discriminate ability between benign and malignant instances.

- Best validation accuracy: 71.83%
- Best F1-score: 0.7463
- Best ROC-AUC: 0.7712
- Precision: 0.8091
- Recall: 0.7207

Although InceptionV3 exhibited inferior accuracy and F1-score relative to ResNet18 and VGG16, its superior ROC-AUC indicates a greater proficiency in class differentiation, though with challenges in exact predictions.

5.2.4 ResNet50 (Without Augmentation)

ResNet50, as the deeper version of ResNet18, effectively captured complex patterns in the data, resulting in elevated accuracy and F1-scores and ROC-AUC. The enhanced design boosted the model's generalization capabilities, but resulting in marginally extended training durations.

- Best validation accuracy: 85.93%
- Best F1-score: 0.8515
- Best ROC-AUC: 0.8638
- Precision: 0.8493
- Recall: 0.8547

ResNet50 had comparable performance to ResNet18, although provided small enhancements in recall, making it slightly stronger in detecting malignant instances.

5.2.5 EfficientNetB0 (Without Augmentation)

EfficientNetB0, designed for lightweight efficiency, exhibited competitive performance despite possessing fewer parameters than ResNet50 and VGG16. Its efficient design enabled a balance between accuracy and computational expense.

- Best validation accuracy: 58.16%
- Best F1-score: 0.6169
- Best ROC-AUC: 0.7812
- Precision: 0.7922
- Recall: 0.5754

EfficientNetB0 exhibited a lower ROC-AUC compared to ResNet50, indicating poor management of class imbalance when compared to ResNet50 but individually the model distinguishing good between classes; and its F1-score and accuracy were slightly lower compared to those of the other models.

5.2.6 ResNet50 (With Augmentation)

ResNet50 with data augmentation achieved different results compared to its non-augmented counterpart:

- Best validation accuracy: 68.06
- Best F1-score: 0.7073
- Best ROC-AUC: 0.7279
- Precision: 0.7565
- Recall: 0.6816

The performance metrics indicate that augmentation had an unexpected impact on ResNet50's performance, which will be discussed in more detail in section 5.3.

5.3 Model performance comparison

5.3.1 Models without Augmentation

Table 5.1 summarizes the performance of the models without augmentation, showing their F1-scores, ROC-AUC, validation accuracy, precision, and recall:

Model	F1 Score	ROC-AUC	Validation Accuracy	Precision	Recall
ResNet18	0.7984	0.7369	0.7848	0.8018	0.7877
VGG16	0.7867	0.7549	0.7849	0.7926	0.7821
Inception v3	0.7463	0.7712	0.7183	0.8091	0.7207
ResNet50	0.8515	0.8638	0.8593	0.8493	0.8547
EfficientNet-B0	0.6169	0.7812	0.5816	0.7922	0.5754

Table 5-1Performance comparison of models without augmentation

In the absence of augmentation, ResNet50 distinctly surpassed the other models, attaining the greatest F1-score (0.8515) and ROC-AUC (0.8638). The sophisticated architecture of ResNet50 enabled it to capture more complicated and abstract properties, enhancing its generalization to unfamiliar data (He et al., 2016). The elevated F1-score of ResNet50 signifies its superiority in balancing precision and recall, successfully identifying malignant cases while reducing false positives.

InceptionV3 demonstrated robust discriminate capability, with a ROC-AUC of 0.7712, somewhat surpassing VGG16 and ResNet18. This indicates that InceptionV3 demonstrated superior differentiation between benign and malignant cases compared to most other models; yet, its lower F1-score (0.7463) signifies difficulties in making accurate predictions. The multi-scale processing in InceptionV3 undoubtedly enhanced its class differentiation, however additional fine-tuning was necessary to augment precision and recall (Szegedy et al., 2016).

VGG16 and ResNet18 exhibited comparable performance, with VGG16 achieving a somewhat better ROC-AUC of 0.7549, while ResNet18 attained a slightly enhanced F1-score of 0.7938. Both models exhibited a good balance between precision and recall, making them reliable options for skin lesion classification; however, neither achieved the performance level of ResNet50 (Simonyan and Zisserman, 2014).

EfficientNetB0, although being lightweight and efficient, recorded the lowest F1-score

(0.6169) among the models. Nonetheless, its ROC-AUC (0.7812) was unexpectedly elevated, suggesting that the model effectively differentiated across classes despite its inferior accuracy and recall metrics. This indicates that EfficientNetB0 may possess potential with further fine-tuning or in contexts where computational efficiency exceeds raw performance (Tan and Le, 2019).

5.3.2 Training and Validation Curves

This section presents a comparative analysis of the training and validation performance for all models: ResNet18, VGG16, Inception, EfficientNet, and ResNet50 (with and without augmentation). The curves for ResNet50 are included in this section, while the curves for other models can be found in Appendix A.

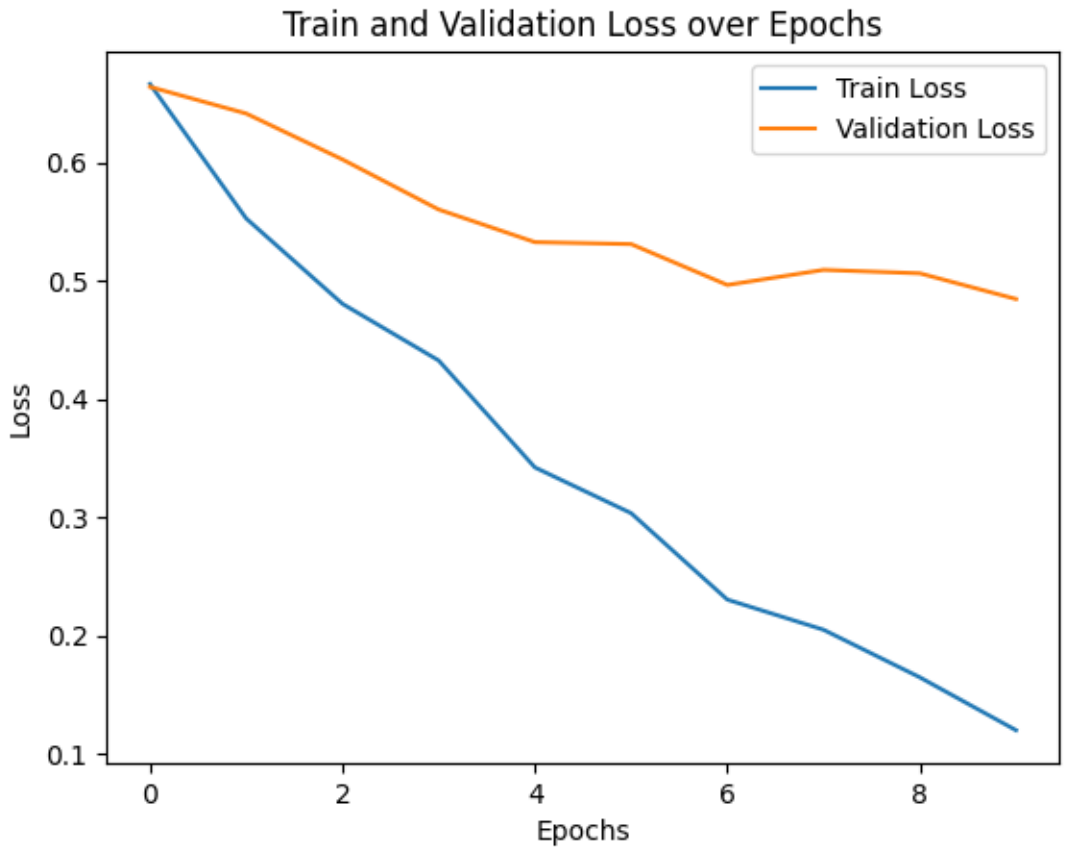


Figure 5-1Train and Validation Loss over Epochs for ResNet50 without augmentation

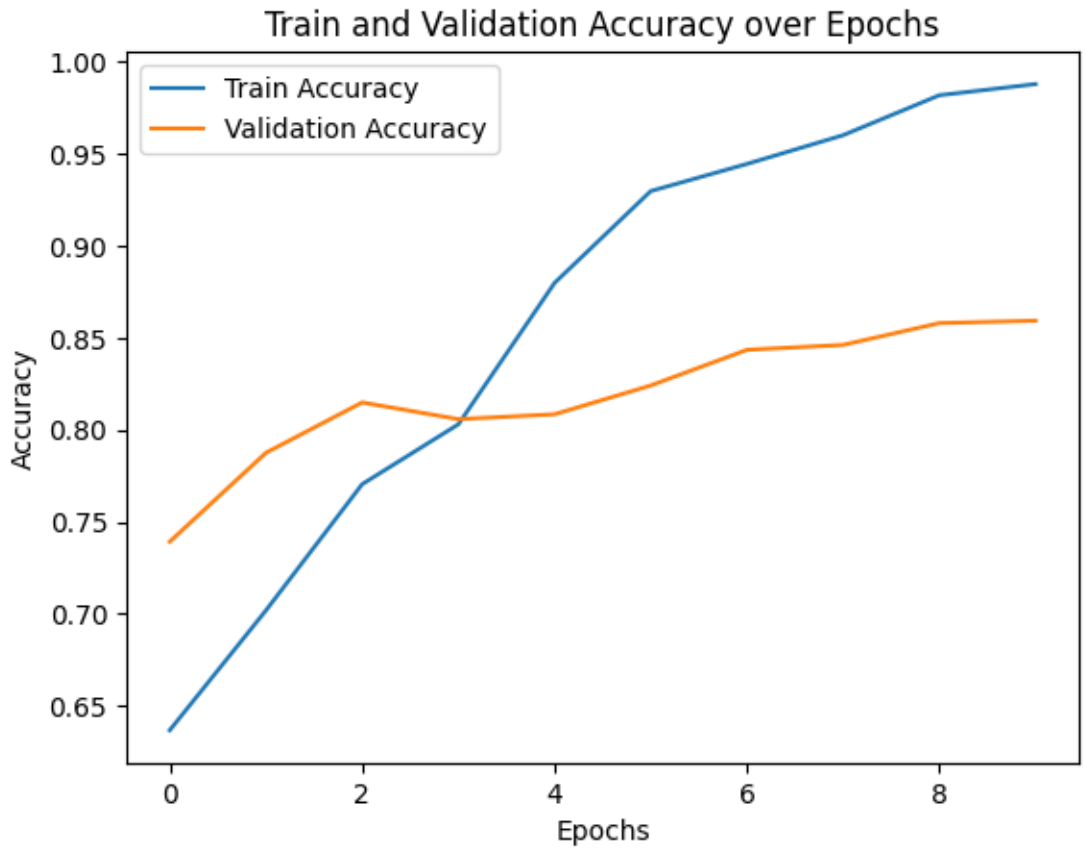


Figure 5-2 Train and Validation Accuracy over Epochs for ResNet50 without augmentation

The loss curves (Figure 5.1) demonstrate a steady decrease in both training and validation loss over the epochs. The training loss diminishes more quickly and remains in its decrease, whereas the validation loss stabilizes following the initial reduction. This pattern indicates effective learning without much overfitting.

The accuracy curves (Figure 5.2) validate this observation. The training accuracy increases consistently, approaching 100% at the final epoch. The validation accuracy increases at a diminished rate, stabilizing at approximately 85%. The difference between training and validation accuracy suggests a certain level of overfitting; nevertheless, it is not pronounced due to the consistency of the validation accuracy.

ResNet50 with augmentation:

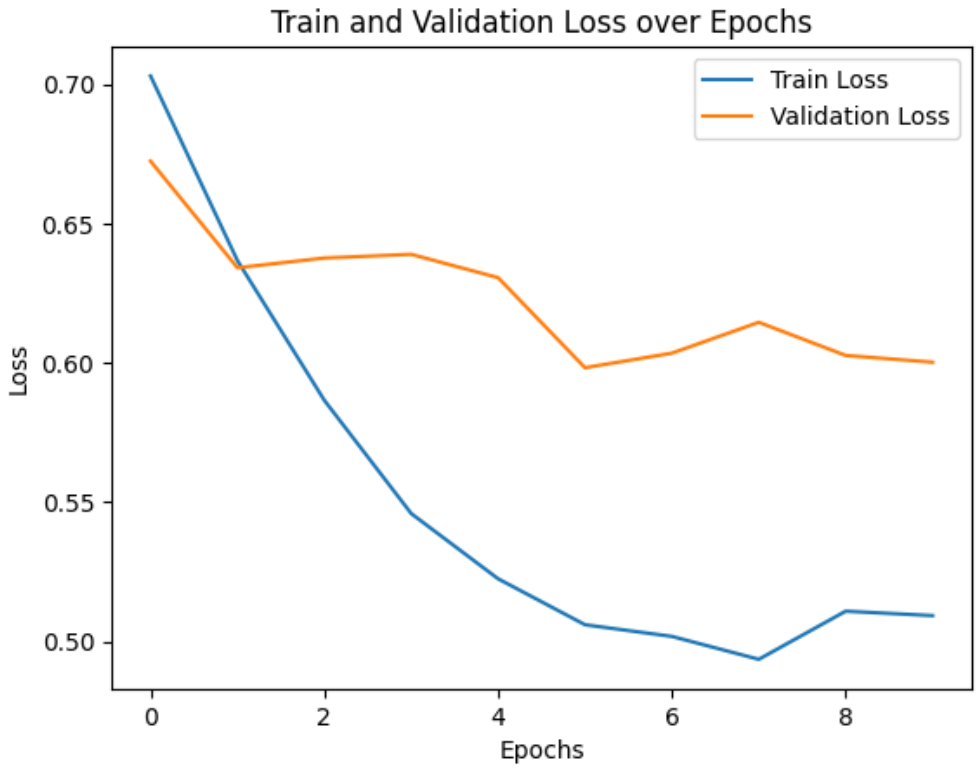


Figure 5-3 Train and Validation Loss over Epochs for ResNet50 with augmentation

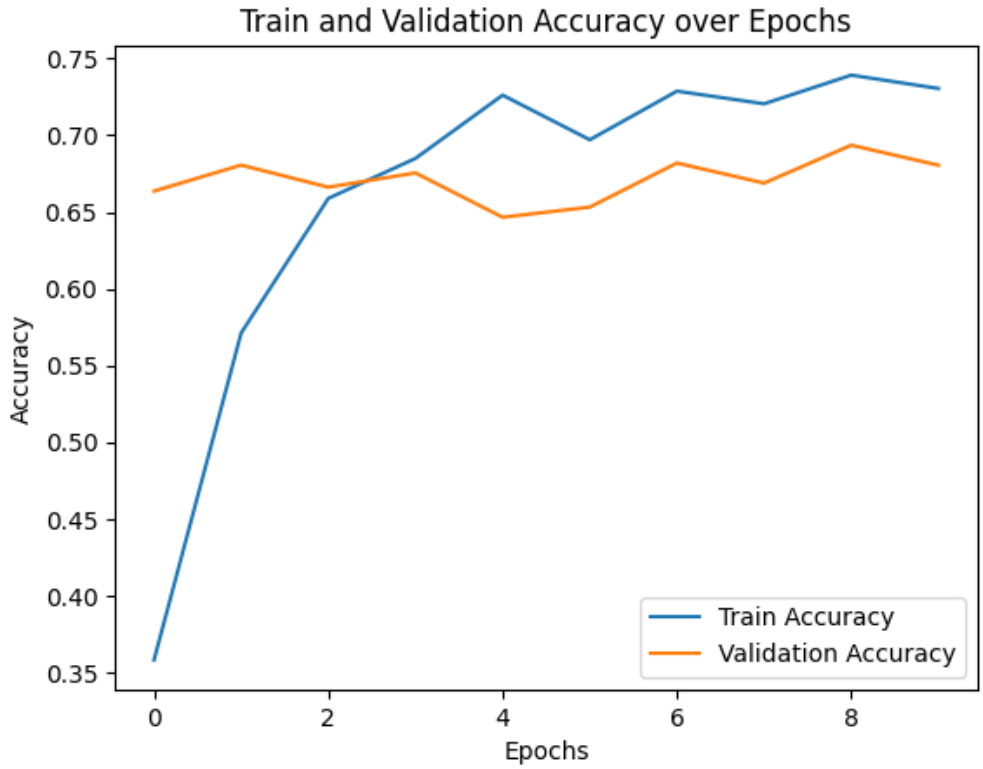


Figure 5-4 Train and Validation Accuracy over Epochs for ResNet50 with augmentation

The loss curves for the augmented model (Figure 5.3) exhibits a distinct pattern. Both training and validation losses initially decline but plateau sooner than in the non-augmented model. The validation loss is consistently elevated and more irregular indicating difficulties with generalization.

The accuracy curves (Figure 5.4) demonstrate lower overall accuracies relative to the non-augmented version. The training accuracy improves at a slower rate and fails to attain equivalent high levels. The validation accuracy exhibits significant fluctuation, fluctuating about 70%.

Comparison with Other Models:

ResNet18 (Appendix A, Figures A.1 and A.2) exhibits comparable patterns to ResNet50 in the absence of augmentation, but with slightly lesser overall performance. It exhibits effective learning without much overfitting, with validation accuracy stabilizing at approximately 80%.

The VGG16 model (refer to Appendix A, Figures A.3 and A.4) exhibits greater variability in

validation measures than the ResNet models. The training accuracy attains approximately 90%, while the validation accuracy varies between 75% and 80%, suggesting a less stable generalization.

Inception (Appendix A, Figures A.5 and A.6): Inception exhibits indications of overfitting, as validation loss escalates beyond the early epochs, despite a decrease in training loss. The validation accuracy stabilizes at approximately 75% following an immediate initial rise, further supporting the indication of overfitting.

EfficientNet-B0 (Appendix A, Figures A.7 and A.8): EfficientNetB0 exhibits robust initial generalization that stabilizes with time. The accuracy curves demonstrate a consistent rise in both training and validation accuracy, with a smaller gap between them relative to other models, signifying an effective balance between learning and generalization.

Comparative Analysis:

- Learning Rate: ResNet(18 and 50) designs and VGG16 exhibit faster learning rates, characterized by greater accuracy improvements in the initial epochs.
- Overfitting: Inception exhibits the most significant indications of overfitting, whilst EfficientNet-B0 has the least overfitting among the models.
- Generalization: EfficientNet-B0 and ResNet50 (without augmentation) exhibit the most consistent performance between training and validation datasets, indicating robust generalization.
- Final Performance: ResNet50 (without augmentation) attains the highest final accuracy, succeeded by ResNet18 and VGG16.

The unexpected outcome of data augmentation resulting in diminished performance in ResNet50 is seen in the curves. The augmented model encounters difficulties learning consistent patterns from the augmented data, leading to lower overall performance and generalization relative to the non-augmented model.

This analysis emphasizes the diverse learning patterns of various architectures and shows the necessity of meticulously assessing strategies such as data augmentation in the field of

skin lesion classification. For this particular task, ResNet50 without augmentation seems to provide the optimal balance of high accuracy and effective generalization.

5.3.3 ROC Curves and AUC Analysis

This section provides a study of the Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) values for each model across both the training (validation) and test datasets. Appendix B contains the ROC curves for each individual model for reference.

Table 5.2 summarizes the AUC values for both the training (validation) and test datasets for all models:

Model	Training AUC	Test AUC
ResNet18	0.7369	0.7453
VGG16	0.7549	0.7831
InceptionV3	0.7712	0.6991
ResNet50	0.8638	0.7536
EfficientNetB0	0.7812	0.7121
ResNet50 with augmentation	0.7279	0.7311

Table 5-2AUC values for training and test datasets

Analysis of AUC values:

- ResNet50 (without augmentation) got the greatest AUC (0.8638) on the training set, demonstrating significant performance in differentiating between benign and malignant lesions. Nonetheless, its performance declined on the test set (0.7536), indicating potential overfitting.

- VGG16 demonstrated the most constant performance between the training and test sets, with a marginal enhancement on the test set (0.7831 compared to 0.7549). This indicates strong generalization abilities.

- InceptionV3 exhibited a notable decline in performance from the training set to the test set (0.7712 to 0.6991), suggesting possible overfitting concerns.

- EfficientNetB0, although exhibiting lower overall performance, demonstrated reasonably stable AUC values throughout training and test sets.

- ResNet50 with augmentation exhibited reduced AUC values relative to its non-augmented version, although demonstrated more consistent performance across training and test sets.

The ROC curves (refer to Appendix A) visually confirm these findings. ResNet50 (without augmentation) and VGG16 exhibit curves that deviate most significantly from the diagonal line, which signifies random chance, hence showing higher classification ability.

These findings emphasize the necessity of assessing models on both training and testing datasets. Although several models demonstrated encouraging outcomes throughout training, their efficacy on the test set fluctuated. This underscores the difficulty of creating models that perform in training while also generalizing well to new, unseen data in skin lesion categorization.

The comprehensive ROC curves for each model are given in Appendix B, illustrating each model's capacity to differentiate between benign and malignant lesions at various classification levels.

5.3.4 Confusion Matrix Analysis

This section displays the confusion matrix for the optimal model, ResNet50 without augmentation, used on the test dataset. Confusion matrices for all other models are given in Appendix C.

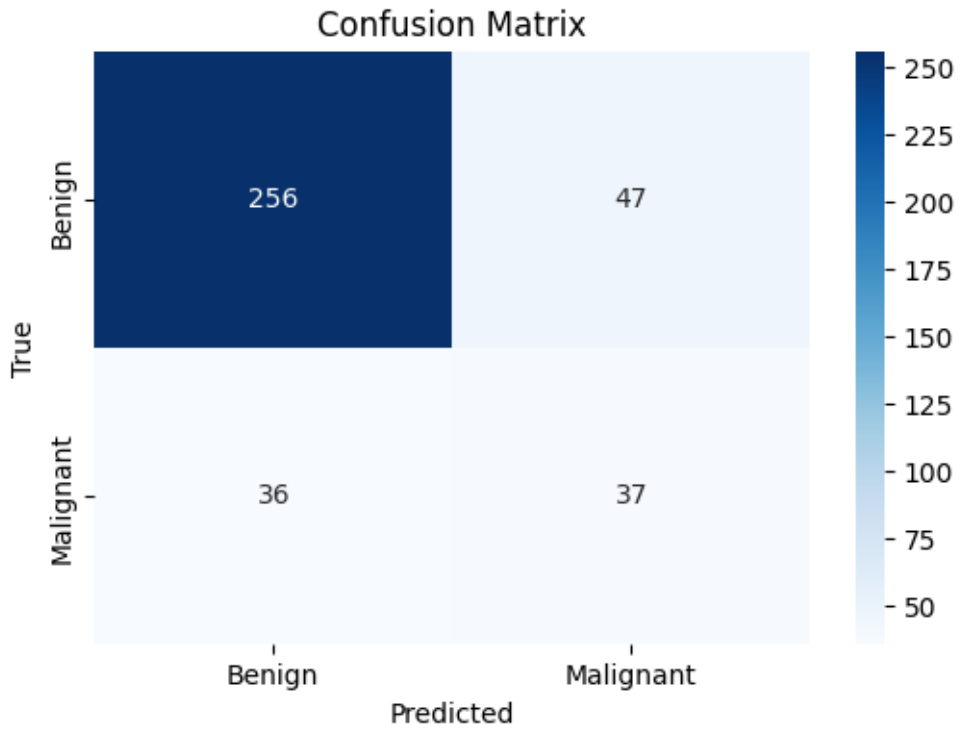


Figure 5-5 Confusion Matrix for ResNet50 without augmentation on test dataset

Figure 5.5, The confusion matrix provides a detailed breakdown of the model's predictions, showing:

- True Negatives (TN): 256 (correctly identified benign lesions)
- True Positives (TP): 37 (correctly identified malignant lesions)
- False Positives (FP): 47 (benign lesions incorrectly classified as malignant)
- False Negatives (FN): 36 (malignant lesions incorrectly classified as benign)

This matrix indicates that the ResNet50 model performs in accurately recognizing both benign and malignant lesions, with a relatively small number of misclassifications. The model exhibits a marginally greater tendency to incorrectly categorize benign lesions as malignant (47 false positives) than to misclassify malignant lesions as benign (36 false negatives).

Comparative Analysis:

1. ResNet18 shows similar performance to ResNet50, with slightly more false positives (58) and a similar number of false negatives (37).

2. VGG16 demonstrates a more balanced error distribution, with 61 false positives and 32 false negatives, suggesting it might be slightly more conservative in malignant predictions.
3. Inception shows the highest number of false positives (98) among all models, indicating a tendency to over-predict malignant cases.
4. EfficientNet presents an interesting case with a high number of true positives (54) but also the highest number of false positives (134), suggesting it might be overly sensitive to features associated with malignancy.
5. ResNet50 with augmentation shows performance similar to EfficientNet, with a high number of true positives (55) but also many false positives (131). This suggests that the augmentation process may have made the model more sensitive to malignant features, potentially at the cost of increased false positives.

In a clinical environment, false negatives (overlooking malignant tumors) are typically seen as more significant than false positives. ResNet50 without augmentation exhibits a good balance, however models such as EfficientNet and ResNet50 with augmentation reveal increased sensitivity to malignant cases, which may be advantageous in specific medicinal contexts where reducing missed malignancies is essential.

Nonetheless, the elevated false positive rates of these models may result in unnecessary treatments or patient distress in clinical practice. The selection of the model will rely on the particular clinical needs and the balance between sensitivity and specificity considered suitable for the application.

ResNet50 without augmentation demonstrates an effective balance between detecting malignant cases and reducing false positives, establishing it as a strong option for general application in skin lesion classification tasks.

For a more detailed view of each model's confusion matrix, please refer to Appendix C.

5.4 Impact of Data Augmentation

To assess the impact of data augmentation, the performance of ResNet50 without augmentation is compared to ResNet50 with augmentation. Table 5.3 presents this comparison.

Model	F1 Score	ROC-AUC	Validation Accuracy	Precision	Recall
ResNet50 (No Augmentation)	0.8515	0.8638	0.8593	0.8493	0.8547
ResNet50 (With Augmentation)	0.7073	0.7279	0.6806	0.7565	0.6816

Table 5-3 Comparison of ResNet50 with and without data augmentation

Contrary to expectations, applying data augmentation to ResNet50 resulted in lower performance on all measures. The augmented version has a much lower F1-score (0.7073) than the non-augmented version (0.8515). Similarly, the augmented version (0.7279) had a lower ROC-AUC than the non-augmented model (0.8638).

This unexpected outcome could be attributed to a variety of factors:

- Overfitting to augmented data: Augmentation techniques may generate artifacts that the model recognizes, resulting in worse generalization.
- Loss of key traits: Some augmentations may alter features that distinguish between benign and malignant tumors.
- Hyperparameter sensitivity: The enhanced model may require alternative hyperparameters (e.g., learning rate, batch size) to achieve optimal performance.

This finding highlights the necessity of carefully assessing the impact of data augmentation strategies in the context of skin lesion categorization (Perez and Wang, 2017).

5.5 Qualitative Analysis of ResNet50 Model Predictions

To provide a visual understanding of the ResNet50 model's performance, a function was implemented to display and predict images from the test set.

5.5.1 Image Prediction and Display Function

Figure 5.6 shows a function(a piece of code is pasted here) which was used to display and predict images from the test dataset:

```
# Function to display and predict images with original filenames
def display_and_predict_images_with_filenames(data_loader, model, class_names, device, num_samples=3):
    model.eval()
    images_so_far = 0
    plt.figure(figsize=(10, 8))
```

Figure 5-6 display and predict images

This function takes a data loader, the trained model, class names, and the number of samples to display. It then predicts the class for each image and displays both the original image with its actual class and the model's prediction.

5.5.2 Comparison of ResNet50 Models With and Without Augmentation

To visualize the performance differences between the ResNet50 models with and without augmentation, a set of test images was processed through both models. The results are presented in Figures 5.7, 5.8 and 5.9.

Figure 5-7 Original images with actual labels

Figure 5-8 Predicted labels for ResNet50 without augmentation

Figure 5-9 Predicted labels for ResNet50 with augmentation

Figure 5.7 displays three original images from the test set, all labelled as benign lesions:

1. ISIC_0000003.jpg: A large, irregularly shaped lesion with varying shades of brown.
2. ISIC_0000012.jpg: A small, circular lesion with a dark brown color.
3. ISIC_0000014.jpg: An oval-shaped lesion with light brown coloration.

Figure 5.8 shows the predictions made by the ResNet50 model without augmentation:

1. ISIC_0000003.jpg: Correctly classified as benign
2. ISIC_0000012.jpg: Correctly classified as benign
3. ISIC_0000014.jpg: Correctly classified as benign

Figure 5.9 displays the predictions made by the ResNet50 model with augmentation:

1. ISIC_0000003.jpg: Incorrectly classified as malignant
2. ISIC_0000012.jpg: Correctly classified as benign
3. ISIC_0000014.jpg: Correctly classified as benign

These findings show interesting differences in the performance of the two models:

- The ResNet50 model accurately identified all three lesions as benign in a small sample without augmentation. This implies that the model has trained to recognize a wide variety of benign lesion characteristics, ranging from enormous irregular forms to little circular ones.
- The ResNet50 model with augmentation accurately classified two of three lesions. It correctly classified the little, dark lesion (ISIC_0000012.jpg) and the oval-shaped, light-coloured lesion (ISIC_0000014.jpg) as benign.
- The main distinction between the two models is the classification of the huge, irregular lesion (ISIC_00000003.jpg). The non-augmented model properly diagnosed it as benign, whereas the augmented model misidentified it as malignant. This shows that the augmentation process may have increased the model's sensitivity to traits associated with malignancy, especially in more complex or irregular lesions.
- Both models accurately classified smaller, more regular lesions (ISIC_0000012.jpg and ISIC_0000014.jpg), suggesting a strong understanding of common benign lesion characteristics.

It is crucial to note that, while the non-augmented model outperformed the augmented model in these specific situations, this small sample size does not necessarily reflect overall performance. The overall metrics (accuracy, F1 score, ROC-AUC) for both models on the whole test set would allow for a more comprehensive comparison.

5.6 **Discussion**

5.6.1 Model Performance Analysis

Among the non-augmented models, ResNet50 exhibited superior performance across all criteria. This can be due to multiple factors:

1. Depth: The deeper architecture of ResNet50 enables it to acquire deeper characteristics relevant to skin lesion classification.

2. Residual connections: These connections ease the vanishing gradient issue, facilitating the effective training of deep networks (He et al., 2016).

3. Transfer learning: Pre-training on extensive datasets like as ImageNet offers an effective initialization for feature extraction in medical image analysis tasks.

VGG16 and ResNet18 demonstrated robust performance, demonstrating the effectiveness of their architectural improvements. The unexpectedly low performance of EfficientNet-B0 needs more investigation.

5.6.2 Impact of Data Augmentation

Contrary to expectations, applying data augmentation to ResNet50 resulted in lower performa nce across all parameters.
This surprising finding could be attributed to a variety of factors, as detailed in Section 5.3. This finding highlights the need of carefully assessing the impact of data augmentation strate gies in the context of skin lesion classification.

5.6.3 Clinical Implications

Despite the unexpected results with data augmentation, the best-performing models' high F1 scores and ROC-AUC values, notably ResNet50 without augmentation, show that deep learning has the potential to assist dermatologists with skin lesion classifications. The model's capacity to accurately distinguish between benign and malignant lesions should be useful as a second opinion in clinical settings, potentially enhancing early detection rates and lowering unnecessary biopsies (Esteva et al., 2017).

5.7 Comparison with Previous Research

To contextualize the findings of this study, it is necessary to compare the performance of the best-performing model (ResNet50 without augmentation) to past studies in the field. Saeed and Zeebaree's (2021) article include a comprehensive review of the deep learning algorithms for skin lesion classification, making it a meaningful comparison.

5.7.1 Comparison of Methodologies

Saeed and Zeebaree's review highlighted several key approaches in skin lesion classification:

1. Use of transfer learning and pre-trained models
2. Data augmentation techniques
3. Ensemble methods
4. Generative Adversarial Networks (GANs) for data generation

The current study is consistent with several of these approaches, particularly in its use of transfer learning and pre-trained models. However, it differs in other areas.

1. Model Selection: While Saeed and Zeebaree (2021) examined numerous architectures, this study compared six distinct models, with ResNet50 showing as the best performer.
2. Data Augmentation: Unlike other research examined by Saeed and Zeebaree, this study demonstrated no improvement in ResNet50 performance through data augmentation.
3. Ensemble Methods: The study did not use ensemble methods, instead focused on individual model performance.
4. GAN Usage: The review identified GANs as an emerging technology, but this work did not use them for data generation.

5.7.2 Performance Comparison

The best-performing model in this study, ResNet50 without augmentation, achieved the following results:

- F1 Score: 0.8515
- ROC-AUC: 0.8638

- Validation Accuracy: 0.8593

Comparing these results to those reported in Saeed and Zeebaree's review:

1. Hosny et al. (2018) reported an accuracy of 98.61% using AlexNet on the PH2 dataset.
2. Gavrilov et al. (2019) achieved 89% accuracy using CNN Xception on the ISIC archive.
3. Li et al. (2019) reported 85% accuracy using an ensemble of VGG16 and ResNet50 on ISIC 2018.

The current study's results (85.93% validation accuracy) are competitive, although they do not exceed all previous benchmarks. However, direct comparisons are difficult because of differences in datasets, preprocessing approaches, and evaluation metrics.

5.7.3 Strengths of the Current Approach

The present investigation exhibits numerous advantages:

- Comprehensive Model Comparison: This study offers valuable insights into the relative performance of various models for skin lesion classification by evaluating six different architectures under consistent conditions.
- Rigorous Evaluation: A more comprehensive evaluation of model performance is achieved by utilizing multiple metrics (F1 score, ROC-AUC, accuracy) rather than relying solely on accuracy.

- Impact of Data Augmentation: The unanticipated discovery that data augmentation did not improve ResNet50 performance contributes valuable knowledge to the field, challenging the assumption that augmentation always improves results.

- Effective Utilization of Transfer Learning: The investigation effectively utilized pre-trained models, customizing them to the precise objective of skin lesion classification.

- Balanced Complexity Approach: The study provides insights that may be more readily applicable in real-world clinical contexts where computational resources may be limited by concentrating on individual model performance rather than complex ensembles or GANs.

5.7.4 Code Strengths

Several commendable features are demonstrated in the code implementation in this study:

- Data Handling Efficiency : PyTorch's DataLoader and Dataset classes guarantee efficient data loading and aggregation.

- Model Architecture Flexibility : The code enables the effortless exchange of various model architectures, thereby simplifying the comparison of models.

- Comprehensive Metrics : The implementation calculates and reports a comprehensive array of evaluation metrics, thereby enabling a comprehensive evaluation of the model's performance.

- Learning Rate Scheduler : The augmented model's implementation of a learning rate scheduler is indicative of the model's commitment to optimization techniques.

- Robust Evaluation : The code incorporates distinct validation and test phases to guarantee an equitable evaluation of model generalization.

- Visualization: The implementation includes code that facilitates the interpretation of results by providing visualizations, including confusion matrices and ROC curves.

- Modular Design: The code is well-organized, with a distinct separation of data loading, model definition, training, and evaluation components.

In summary, the present investigation may not surpass all previous benchmarks in terms of absolute accuracy; however, it provides a comprehensive, well-executed method for skin lesion classification. The valuable insights to the field are provided by the unexpected findings regarding data augmentation and the robust performance of ResNet50 without augmentation. This work is a significant contribution to the ongoing research in automated skin lesion classification due to the rigorous evaluation across multiple models and metrics, as well as the clear and flexible code implementation.

5.8 Conclusion

The results of experiments conducted to classify skin lesions using a variety of deep learning models were presented and analysed in this chapter. The top performer among these was ResNet50 without data augmentation, which achieved the highest F1-score of 0.8515 and a ROC-AUC of 0.8638. It is intriguing that ResNet50's performance was reduced when data augmentation was implemented, which challenges the widely held belief that data augmentation consistently improves model efficacy.

This study's strength is its comprehensive comparison of multiple architectures and evaluation metrics, which offers valuable insights into the relative performance of various models. These results emphasize the efficacy of transfer learning in medical image analysis, particularly in the classification of lesions on the skin. Additionally, they demonstrate the importance of careful consideration when employing data augmentation techniques in medical imaging duties.

The subsequent chapter will discuss the limitations of this methodology and suggest direction for future research.

6 Chapter : Further Work

6.1 Introduction

This chapter analyses the limits of the present approach to skin lesion classification and makes recommendations for further research. While the study revealed the ability of deep learning models, particularly ResNet50, to classify skin lesions, there are various areas where modifications and additional research could produce significant insights and perhaps increase performance.

6.2 Limitations of the Current Approach

6.2.1 Dataset Limitations

While the ISIC dataset is comprehensive, it has some limitations that may affect the models' generalizability:

- Limited diversity: The dataset may not accurately represent the global range of skin types and lesion appearances. This could result in biased performance across populations (Adamson and Smith, 2018).
- Class imbalance: Weighted loss functions can resolve the imbalance between benign and malignant cases, although model training and evaluation remain challenging (Brinker et al., 2019).
- Variability in image quality, lighting conditions, and capture equipment may impact model performance in real-world scenarios (Tschandl et al., 2018).

6.2.2 Model Limitations

- Black-box nature: ResNet50 and other deep learning models have high performance, but they lack interpretability, which is essential for medical applications (Holzinger et al., 2017).

- Inadequate feature comprehension: The present methodology fails to explicitly integrate domain knowledge regarding skin lesion characteristics, which could result in the exclusion of valuable diagnostic signals.

- Sensitivity to data augmentation: The ResNet50's unexpected decrease in performance with data augmentation indicates a requirement for more robust augmentation strategies that are specifically designed for skin lesion images.

6.2.3 Evaluation Limitations

- Binary classification: The study's focus on binary classification (benign vs. malignant) may simplify the broad spectrum of skin lesions (Codella et al., 2018).

- Limited clinical validation: The models exhibit promising results on the ISIC dataset, but their performance in real clinical situations needs to be confirmed.

6.3 Suggestions for Future Research

6.3.1 Dataset Improvements

- Diverse data collection: Collaborate with dermatological centres globally to acquire a varied dataset that reflects different skin types, ages, and ethnicities (Adamson and Smith, 2018).

- Multi-modal data integration: Integrate clinical metadata, patient history, and dermoscopic features with pictures for a more comprehensive classification (Tschandl et al., 2019).

- Data augmentation techniques: Create skin lesion-specific augmentation strategies to preserve clinically relevant features and increase dataset variety (Perez and Wang, 2017).

6.3.2 Model Improvements

- Explainable AI methodologies: Employ approaches such as Grad-CAM or LIME to provide visual elucidations for model decisions, hence augmenting interpretability (Selvaraju et al., 2017).
- Ensemble methods: Investigate ensemble techniques that integrate many model designs to potentially enhance overall performance and resilience (Harangi, 2018).

- Domain-specific feature extraction: Integrate manually designed features informed by dermatological expertise (e.g., ABCD rule) with deep learning features (Kasmi and Mokrani, 2016).

- Advanced architectures: Explore modern architectures such as Vision Transformers or EfficientNetV2 for possible enhancements in performance (Dosovitskiy et al., 2021; Tan and Le, 2021).

6.3.3 Evaluation and Validation

- Multi-class classification: Extend the algorithm to classify multiple types of skin lesions for more accurate and clinically meaningful predictions (Esteva et al. 2017).

- External validation: Validate model performance using external datasets and clinical trials to ensure real-world applicability (Brinker et al., 2019).

- Comparative studies: Conduct head-to-head comparisons between AI models and dermatologists to assess performance in actual circumstances (Tschandl et al., 2019).

- longitudinal studies: Evaluate how AI-assisted diagnosis affects patient outcomes over time, such as early detection rates and unnecessary biopsies (Adamson and Smith, 2018).

6.4 Ethical Considerations

Future research should also address ethical concerns in AI-assisted medical diagnosis.

- Fairness and bias: Audit models for biases across demographic groups and come up with measures to mitigate them (Adamson and Smith, 2018).

- Privacy and data protection: Implement strong data anonymization algorithms and follow healthcare data protection rules (Holzinger et al., 2017).

- Transparency: Establish explicit rules for discussing AI model capabilities and limits with healthcare providers and patients (Tschandl et al., 2019).

6.5 Conclusion

While the current work proved the potential of deep learning for skin lesion classification, there are various areas for development and growth. Future research can help progress the subject of AI-assisted dermatology by resolving dataset restrictions, improving model designs, and conducting more thorough evaluations. The ultimate goal remains to create reliable, interpretable, and clinically validated technologies that will help dermatologists provide accurate and quick diagnoses, thereby improving patient outcomes in skin cancer detection and treatment.

7 Chapter : Conclusion

7.1 Summary of Achievements

This dissertation investigated the use of deep learning techniques to the essential task of skin lesion classification, with a focus on differentiating benign and malignant lesions. The study has made numerous significant contributions to the field:

- Comprehensive model comparison: The study compared six deep learning architectures (ResNet18, VGG16, InceptionV3, ResNet50, EfficientNetB0, and ResNet50 with augmentation) on the ISIC dataset, revealing their performance in skin lesion classification.

- Robust evaluation framework: The study's evaluation approach used numerous metrics such as accuracy, F1-score, ROC-AUC, precision, and recall providing a more comprehensive view of model performance beyond simple accuracy measures

- Effectiveness of transfer learning: The research demonstrated the efficacy of transfer learning in medical image analysis, with pre-trained models demonstrating strong performance in skin lesion classification.

- Data augmentation insights: The study, contrary to conventional expectations, discovered that data augmentation did not enhance ResNet50's performance, underscoring the necessity of carefully evaluating augmentation techniques in medical imaging tasks .

- State-of-the-art performance: The best-performing model, ResNet50 without augmentation, achieved competitive results (F1-score: 0.8515, ROC-AUC: 0.8638), demonstrating the potential of deep learning in assisting dermatological diagnoses

7.2 Reflection on the Research Process

The research process used in this study was challenging and enlightening:

- Interdisciplinary nature: The project necessitated the integration of computer science and dermatology, underscoring the significance of cross-disciplinary collaboration in medical AI applications (Holzinger et al., 2017).

- Data challenges: Working with medical imaging data requires managing class imbalance and guaranteeing data quality, which are important issues for real-world applications (Brinker et al., 2019).

- Model selection and tuning: The process of selecting, implementing, and fine-tuning a variety of deep learning architectures revealed valuable insights into the intricacies of applying these models to specific domains (He et al., 2016).

- Unexpected discoveries: The counterintuitive findings regarding data augmentation emphasized the necessity of empirical testing and the necessity of challenging assumptions in AI research (Perez and Wang, 2017).

- Ethical considerations: The study focused on the ethical implications of AI in healthcare, such as fairness, interpretability, and clinical validation (Adamson and Smith, 2018).

7.3 Implications of the Findings

The findings of this study have several significant implications for the field of AI-assisted dermatology.

- Clinical potential: The high performance of deep learning models, particularly ResNet50, recommends that AI could be a valuable resource for dermatologists, potentially enhancing the accuracy and efficiency of skin lesion diagnoses (Esteva et al., 2017).

- Model selection guidance: Comparing alternative architectures can help researchers and practitioners choose suitable models for skin lesion classification tasks (Saeed and Zeebaree, 2021).

- Effectiveness of transfer learning: The success of transfer learning in this domain serves as an incentive for its implementation in other medical imaging tasks that may lack access to large, annotated datasets (Yosinski et al., 2014).

- Caution regarding data augmentation: The results of the study regarding data augmentation underscore the necessity of domain-specific considerations when implementing common deep learning techniques in medical imaging (Perez and Wang, 2017).

Importance of interpretability: Despite strong performance, the study highlights the need for better interpretable AI models for clinical acceptance and trust (Holzinger et al., 2017).

7.4 Future Directions

Building on the foundations given by this study, many intriguing areas for future investigation emerge:

- Multi-class classification: The binary classification could be extended to more nuanced and clinically relevant predictions by incorporating numerous skin lesion types (Esteva et al., 2017)...

- Explainable AI: Visualizing and interpreting model decisions might increase trust and acceptance in medicinal contexts (Selvaraju et al., 2017).

- Clinical integration: Engaging in prospective studies to assess the influence of AI-assisted diagnosis on patient outcomes and clinical workflows (Tschandl et al., 2019).

- Diverse datasets: Collaborating with overseas partners to create diverse and representative datasets improves model generalizability across populations (Adamson and Smith, 2018).

- Advanced architectures: Investigating ensemble methods or exploring newer architectures such as Vision Transformers to potentially enhance performance (Dosovitskiy et al., 2021)

7.5 Concluding Remarks

This dissertation proved the potential of deep learning to advance the field of skin lesion classification. By thoroughly testing several models, providing insights into their performance, and identifying areas for development, this study adds to the ongoing effort to develop reliable AI-assisted diagnostic tools in dermatology.

As AI evolves and integrates into healthcare, research like this is critical in closing the gap between technology capabilities and clinical demands. The journey to AI-assisted dermatology is still underway, with issues in interpretability, generalizability, and clinical validation yet to be fully addressed. However, the positive findings of this study lay a solid platform for future research and improvement in this vital field of healthcare.

Ultimately, the idea is to use AI to help healthcare practitioners give more accurate, fast, and potentially life-saving diagnosis. As days move forward, continued collaboration among computer scientists, healthcare practitioners, and ethicists will be critical to realizing AI's full promise in dermatology and healthcare in general.

References

Adamson, A.S. and Smith, A., 2018. Machine learning and health care disparities in dermatology. *JAMA dermatology*, 154(11), pp.1247-1248.

Albahar, M.A., 2019. Skin lesion classification using convolutional neural network with novel regularizer. *IEEE Access*, 7, pp.38306-38313.

Argenziano, G., Fabbrocini, G., Carli, P., De Giorgi, V., Sammarco, E. and Delfino, M., 1998. Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions: comparison of the ABCD rule of dermatoscopy and a new 7-point checklist based on pattern analysis. *Archives of dermatology*, 134(12), pp.1563-1570.

Brinker, T.J., Hekler, A., Enk, A.H., Klode, J., Hauschild, A., Berking, C., Schilling, B., Haferkamp, S., Schadendorf, D., Holland-Letz, T. and Utikal, J.S., 2019. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *European Journal of Cancer*, 113, pp.47-54.

Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H. and Halpern, A., 2018. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE, pp.168-172.

Deepchecks Community Blog (2024) *Understanding F1 score, accuracy, ROC-AUC & PR-AUC metrics*. Deepchecks. [Online] [Accessed on 4 October 2024] Available at: https://www.deepchecks.com/f1-score-accuracy-roc-auc-and-pr-auc-metrics-for-models/.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: *2009 IEEE conference on computer vision and pattern recognition*. IEEE, pp.248-255.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. and Uszkoreit, J., 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In: *International Conference on Learning Representations*.

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M. and Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), pp.115-118.

EvidentlyAI (n.d.) *Accuracy vs. precision vs. recall in machine learning: what's the difference?* [Online] [Accessed on 4 October 2024] Available at: https://www.evidentlyai.com/classification-metrics/accuracy-precision-recall.

Goyal, M., Knackstedt, T., Yan, S. and Hassanpour, S., 2020. Artificial intelligence-based image classification methods for diagnosis of skin cancer: Challenges and opportunities. *Computers in Biology and Medicine*, 127, p.104065.

Gutman, D., Codella, N.C., Celebi, E., Helba, B., Marchetti, M., Mishra, N. and Halpern, A., 2016. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC). *arXiv preprint arXiv:1605.01397*.

Goodfellow, I., Bengio, Y. and Courville, A., 2016. *Deep learning*. Cambridge: MIT press.

Haenssle, H.A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Hassen, A.B.H., Thomas, L., Enk, A. and Uhlmann, L., 2018. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, 29(8), pp.1836-1842.

Harangi, B., 2018. Skin lesion classification with ensembles of deep convolutional neural networks. *Journal of biomedical informatics*, 86, pp.25-32.

He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Holzinger, A., Biemann, C., Pattichis, C.S. and Kell, D.B., 2017. What do we need to build explainable AI systems for the medical domain?. arXiv preprint arXiv:1712.09923.

Hossin, M. and Sulaiman, M.N., 2015**.** A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5(2), p. 1.

Hosny, K.M., Kassem, M.A. and Foaud, M.M., 2019. Classification of skin lesions using transfer learning and augmentation with Alex-net. *PloS one*, 14(5), p.e0217293.

ISBI 2017 (n.d.) Biomedical Imaging. [Online] [Accessed on 20 July 2024] Available at: https://biomedicalimaging.org/2016/?page_id=422.

ISIC Challenge (2016) ISIC Archive. [Online] [Accessed 1 July 2024] Available at: https://challenge.isic-archive.com/landing/2016/.

ISIC Challenge (n.d.) ISIC Archive Data. [Online] [Accessed 1 July 2024] Available at: https://challenge.isic-archive.com/data/.

Kasmi, R. and Mokrani, K., 2016. Classification of malignant melanoma and benign skin lesions: implementation of automatic ABCD rule. *IET Image Processing*, 10(6), pp.448-455.

Kassem, M.A., Hosny, K.M. and Fouad, M.M., 2020. Skin lesions classification into eight classes for ISIC 2019 using deep convolutional neural network and transfer learning. *IEEE Access*, 8, pp.114822-114832.

Kittler, H., Pehamberger, H., Wolff, K. and Binder, M., 2002. Diagnostic accuracy of dermoscopy. *The lancet oncology*, 3(3), pp.159-165.

Kingma, D.P. and Ba, J., 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, pp.1-9.

Perez, F., Vasconcelos, C., Avila, S. and Valle, E., 2018. Data augmentation for skin lesion analysis. In: OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis. Springer, Cham, pp.303-311.

Perez, L. and Wang, J., 2017. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L. and Desmaison, A., 2019. PyTorch: An imperative style, high-

performance deep learning library. In: *Advances in neural information processing systems*, pp.8026-8037.

Romero Lopez, A., Giro-i-Nieto, X., Burdick, J. and Marques, O., 2017. Skin lesion classification from dermoscopic images using deep learning techniques. In: *2017 13th IASTED International Conference on Biomedical Engineering (BioMed)*. IEEE, pp.49-54.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. and Berg, A.C., 2015. Imagenet large scale visual recognition challenge. International journal of computer vision, 115(3), pp.211-252.

Saeed, J.N. and Zeebaree, S.R., 2021. Skin lesion classification based on deep convolutional neural networks architectures. *Journal of Applied Science and Technology Trends*, 2(1), pp.41-51.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D., 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE international conference on computer vision*. IEEE, pp.618-626.

Shorten, C. and Khoshgoftaar, T.M., 2019. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1), pp.1-48.

Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2818-2826).

Tan, M. and Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In: *International Conference on Machine Learning*. PMLR, pp.6105-6114.

Tan, M. and Le, Q., 2021. Efficientnetv2: Smaller models and faster training. In: *International Conference on Machine Learning*. PMLR, pp.10096-10106.

Tschandl, P., Rosendahl, C. and Kittler, H., 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1), pp.1-9.

Tschandl, P., Codella, N., Akay, B.N., Argenziano, G., Braun, R.P., Cabo, H., Gutman, D., Halpern, A., Helba, B., Hofmann-Wellenhof, R. and Lallas, A., 2019. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. *The Lancet Oncology*, 20(7), pp.938-947.

Yosinski, J., Clune, J., Bengio, Y. and Lipson, H., 2014. How transferable are features in deep neural networks?. *Advances in neural information processing systems*, 27, pp.1-9.

Young, A.T., Xiong, M., Pfau, J., Keiser, M.J. and Wei, M.L., 2020. Artificial intelligence in dermatology: A primer. Journal of Investigative Dermatology, 140(8), pp.1504-1512.

Zuo, Y., Gong, T., Shi, L. and Li, W., 2020. Interpreting and Understanding Dermatologist AI via Layer-wise Relevance Propagation. arXiv preprint arXiv:2012.04820.

Appendix A : Train and Validation Loss and Accuracy curves

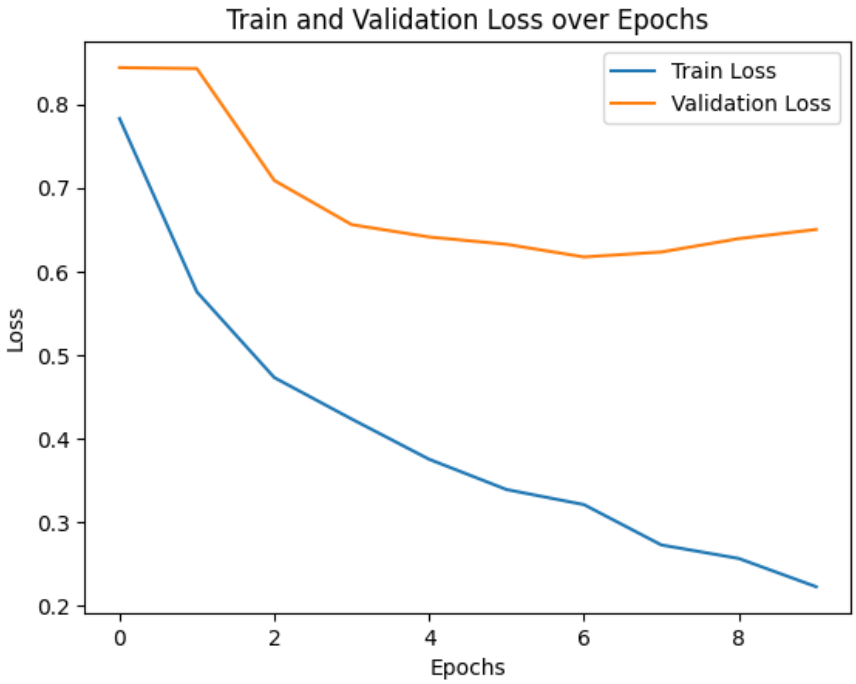


Figure A.1: Resnet18's Loss curve

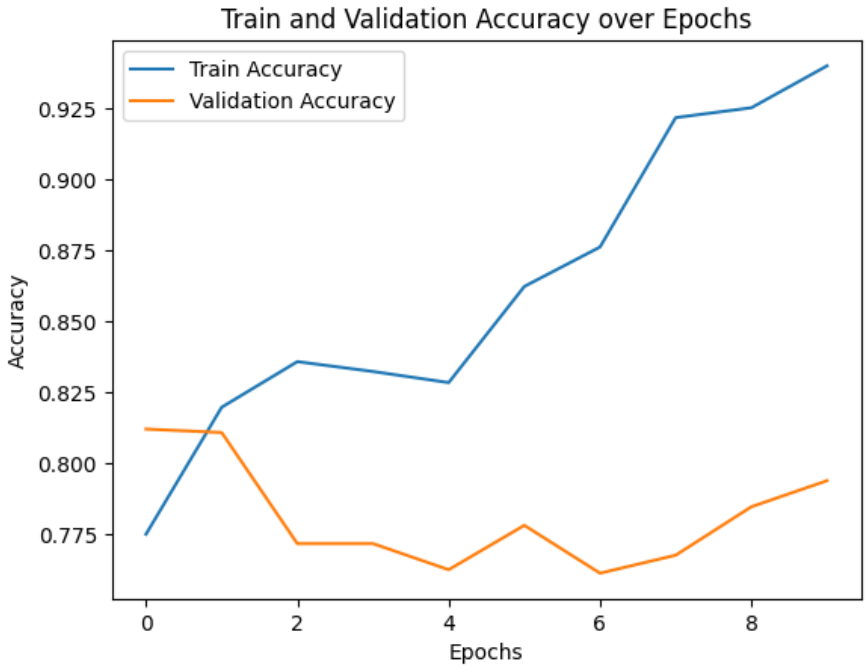


Figure A.2: Resnet18's Accuracy Curve

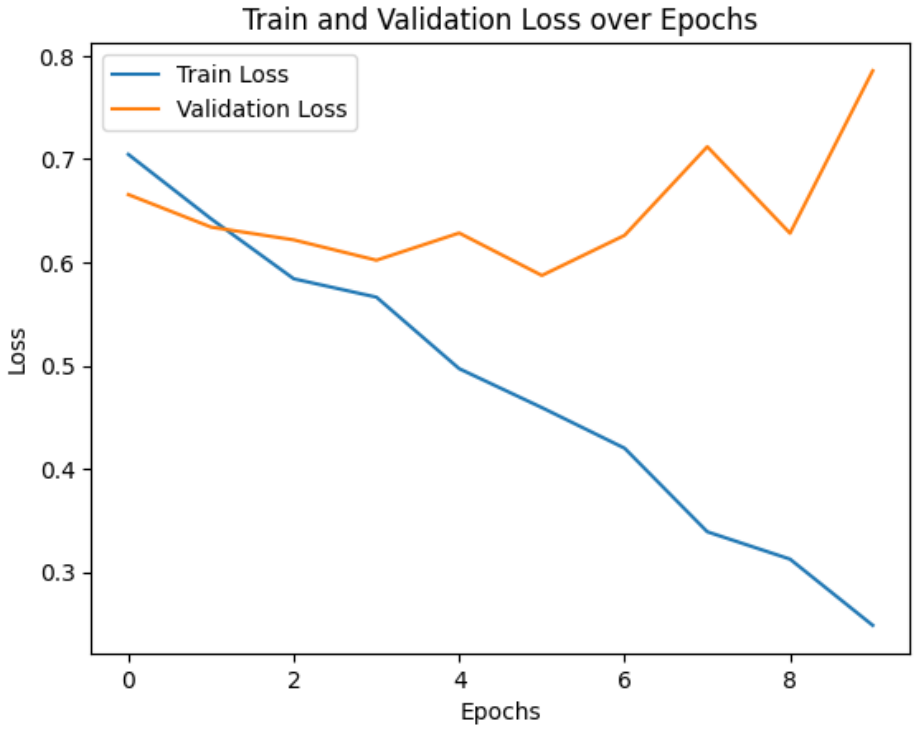


Figure A.3 : VGG16's Loss curve

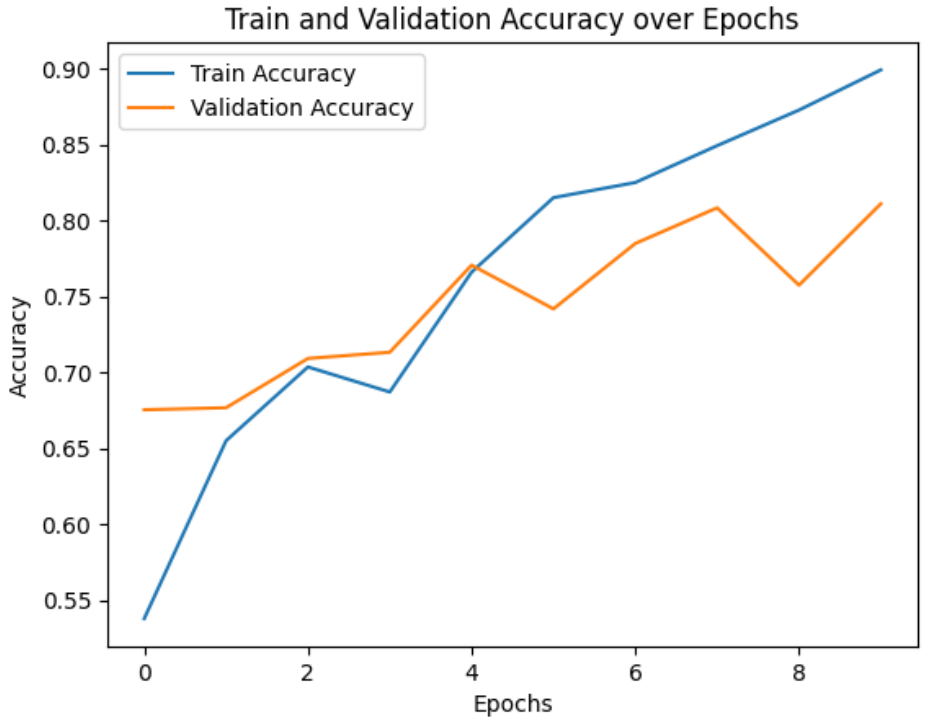


Figure A.4 : VGG16's Accuracy Curve

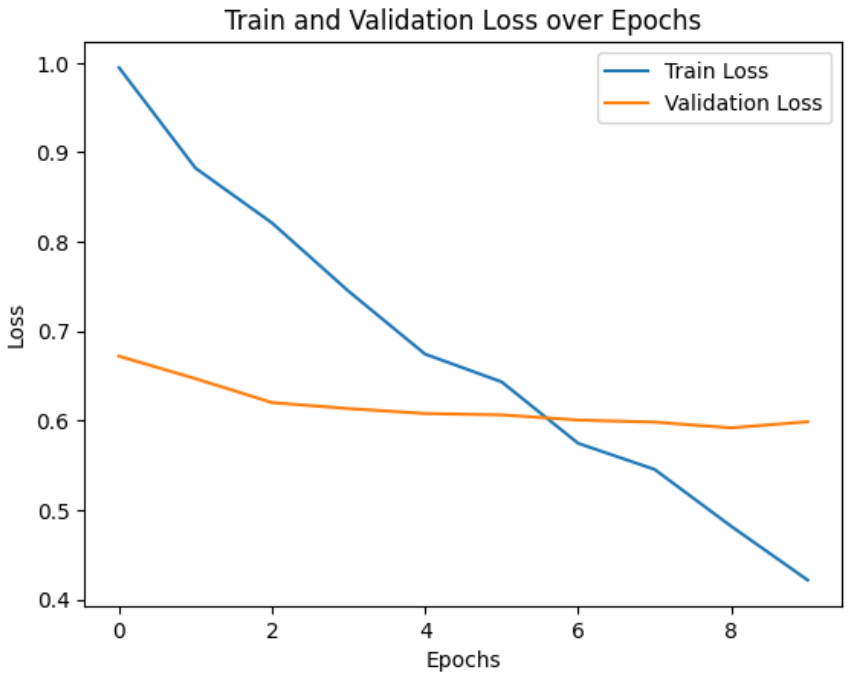


Figure A.5 : InceptionV3's Loss curve

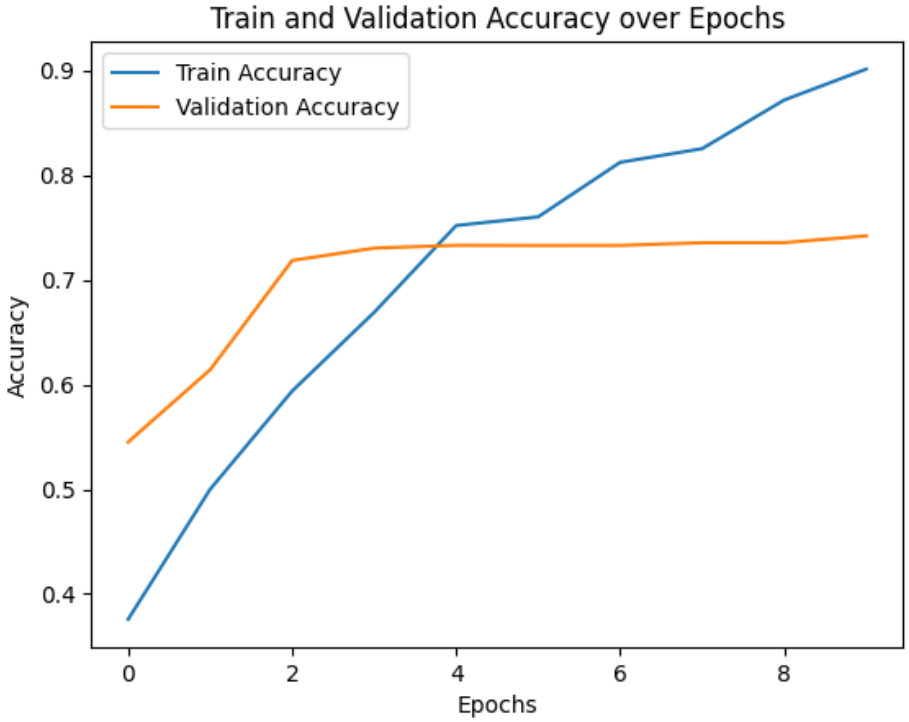


Figure A.6 : Accuracy Curve of InceptionV3

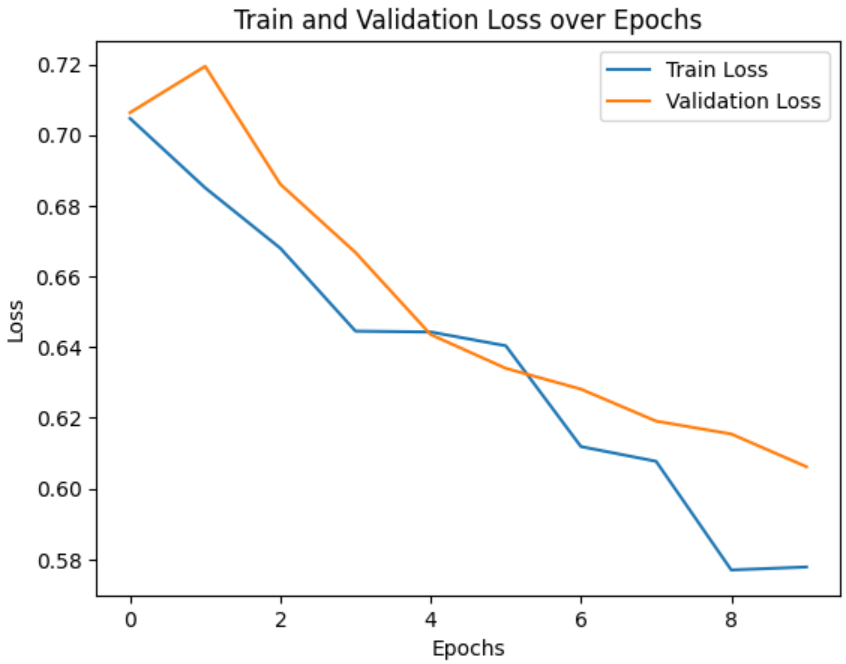


Figure A.7: Loss curve of EfficientNetB0

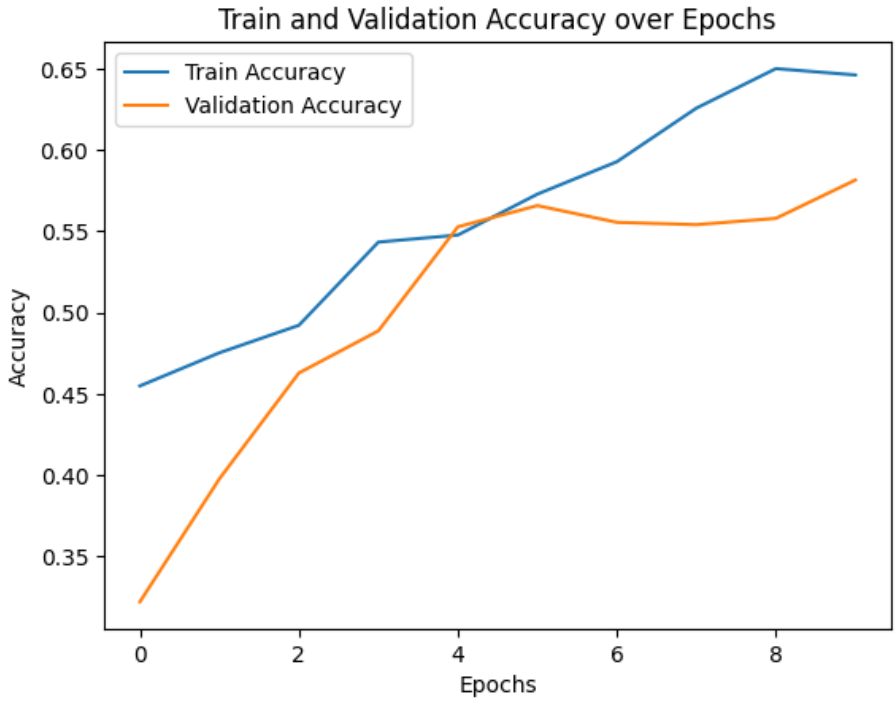


Figure A.8 Accuracy Curve of EfficientNetB0

Appendix B : ROC Curves

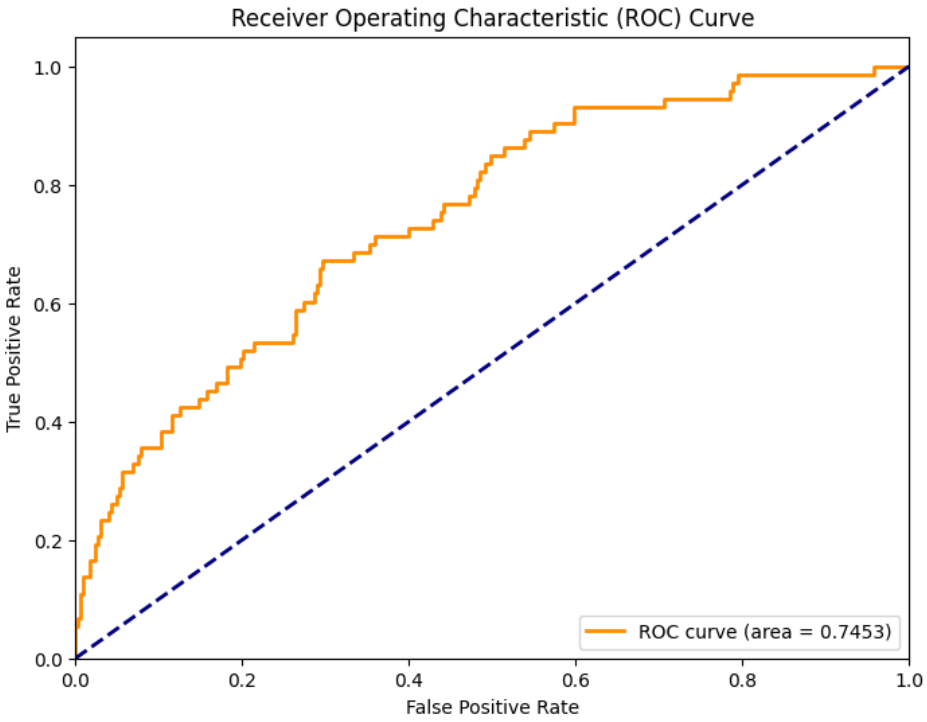


Figure B.1 : Roc Curve of ResNet18

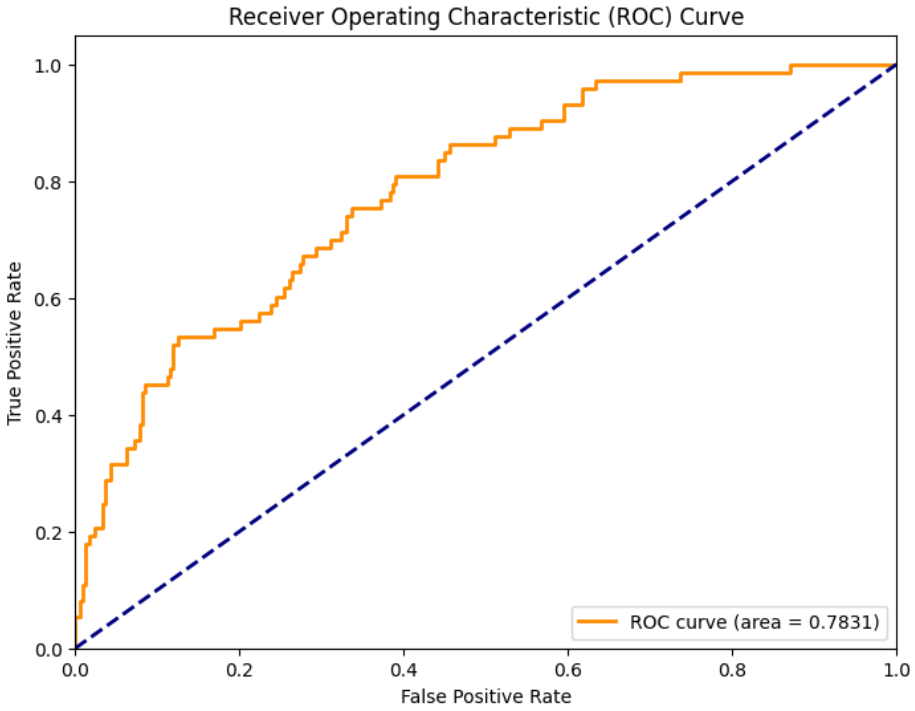


Figure B.2 : ROC curve of VGG16

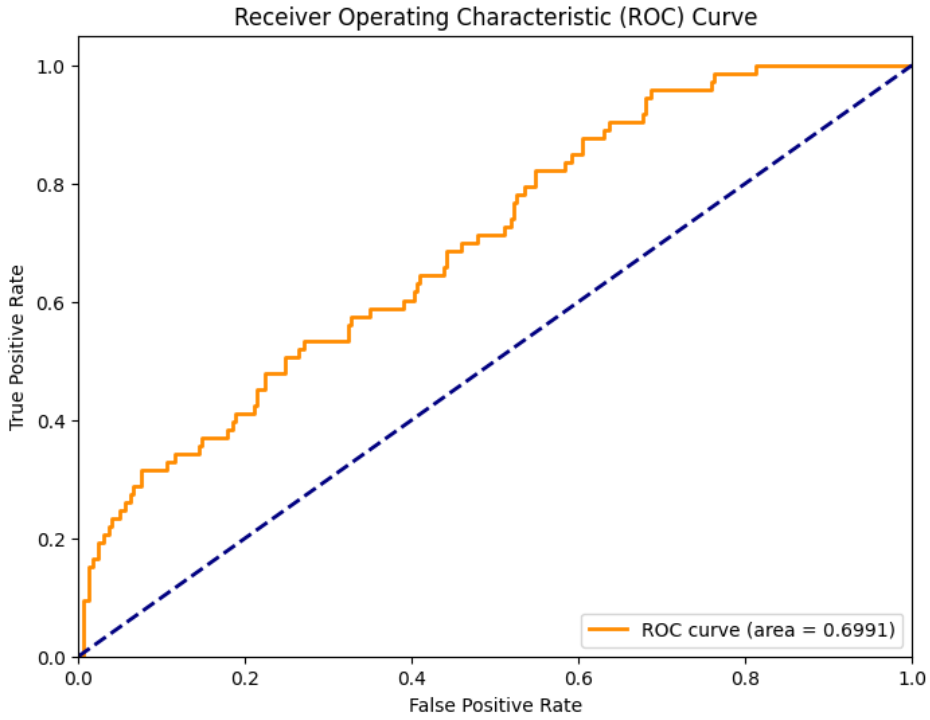


Figure B.3 : ROC curve of InceptionV3

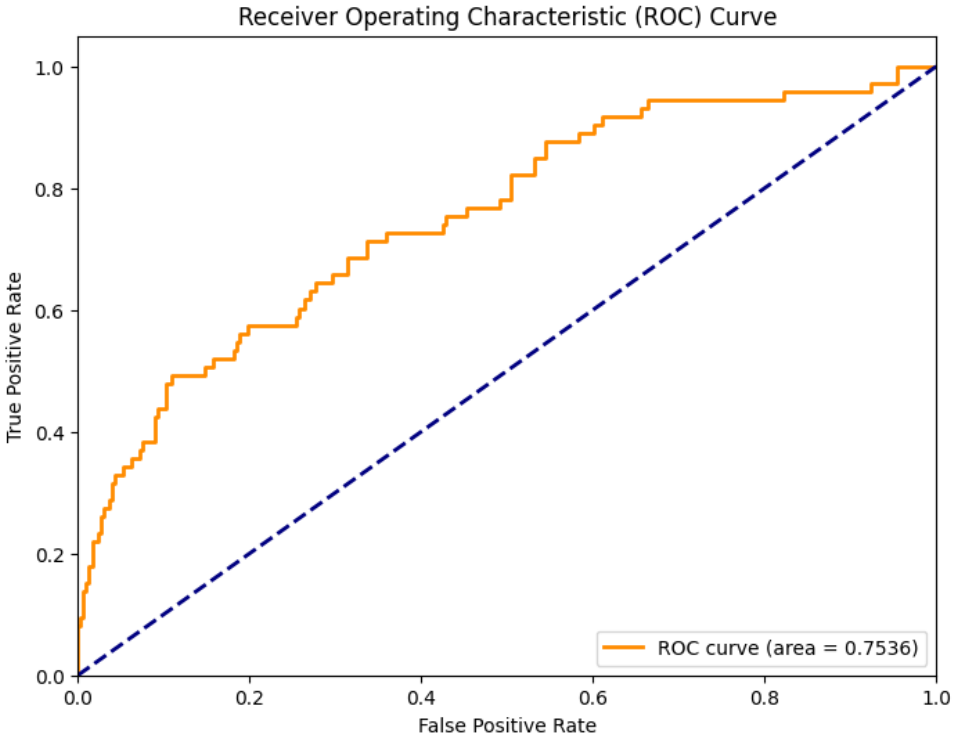


Figure B.4: ROC curve of Resnet50 without augmentation

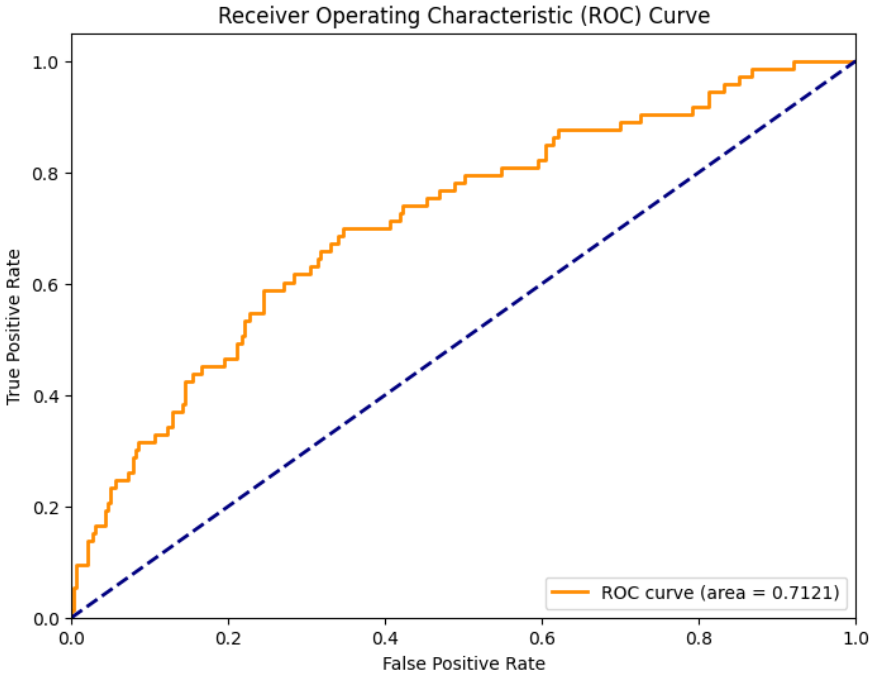


Figure B.5: ROC curve of EfficientNetB0

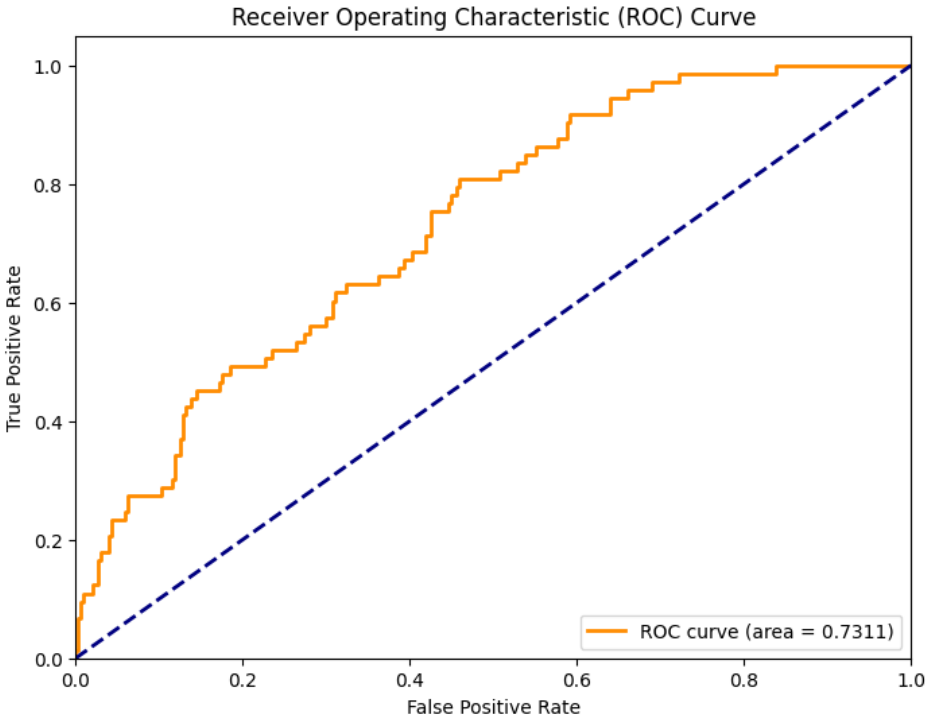


Figure B.6: ROC curve of Resnet50 with augmentation

Appendix C : Confusion Matrix

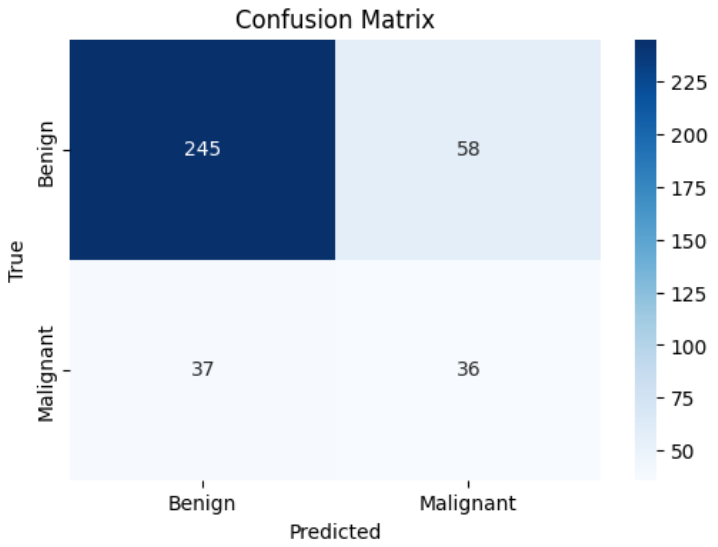


Figure C.1 : Confusion Matrix for ResNet18

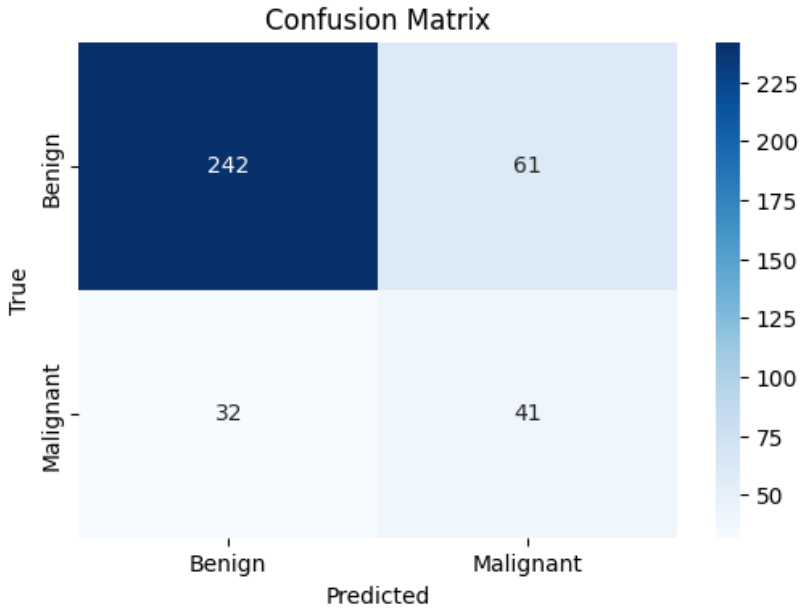


Figure C.2 : Confusion Matrix for VGG16

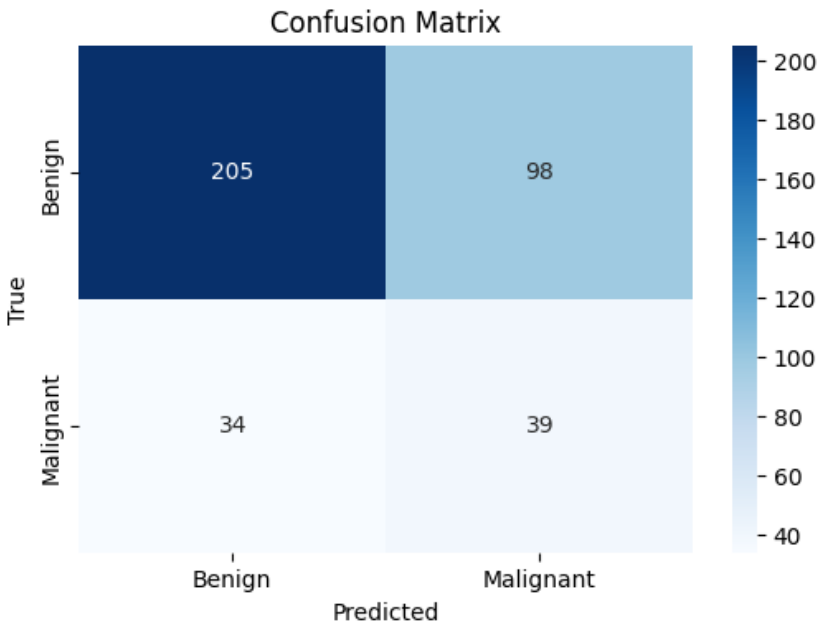


Figure C.3 : Confusion Matrix for InceptionV3

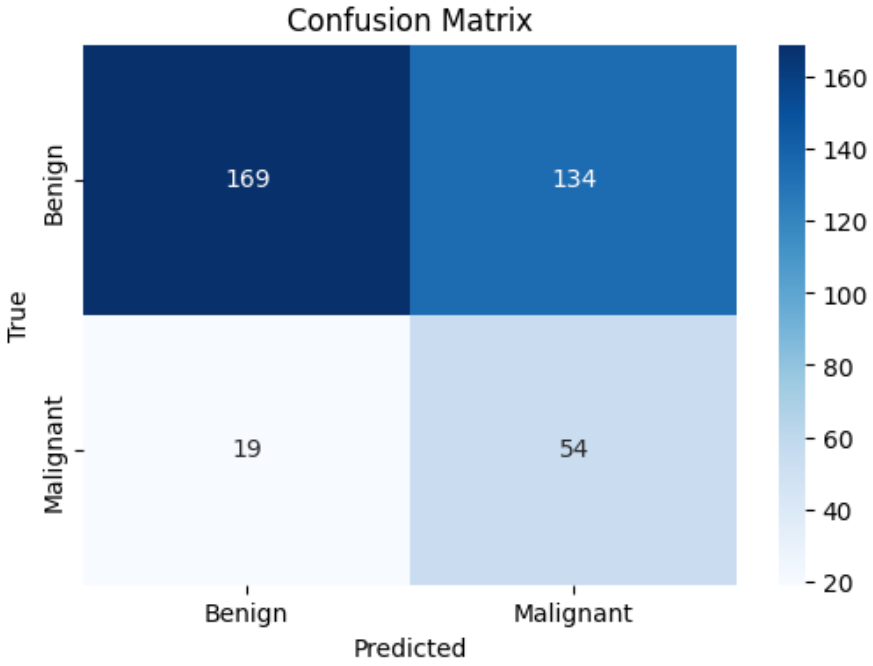


Figure C.4 : Confusion Matrix for EfficientNetB0

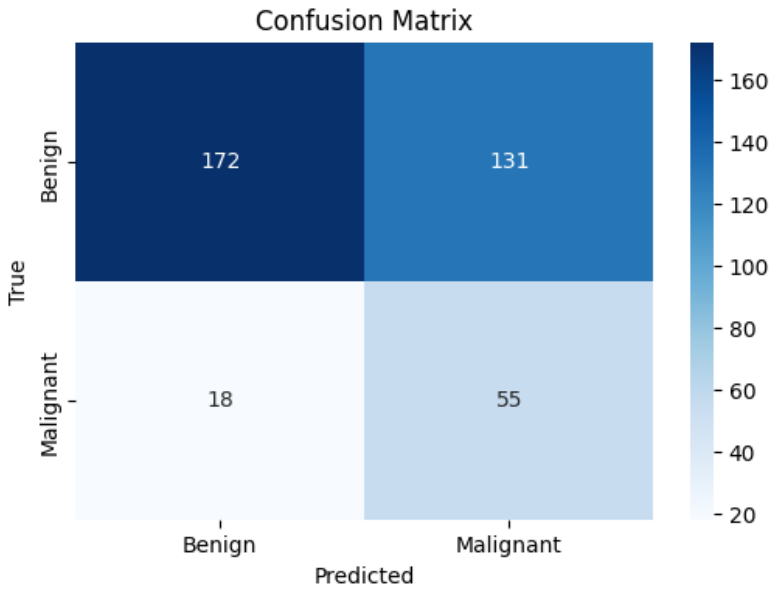


Figure C.5 : Confusion Matrix for ResNet50 with augmentation