Universität des Saarlandes

Department of Computer Science

Master thesis

Enhanced Skin Lesion Detection from Multiple Lesions with Deep Learning Models using DICOM Standard Dataset

submitted by

Hafeez Ullah

on Jan 05, 2025

Reviewers

Prof. Dr. Alexey Gurevich

Prof. Dr.-Ing. Paul Motzki

Eidesstattliche Erklärung

Ich erkläre hiermit an Eides statt, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Statement in Lieu of an Oath

I hereby confirm that I have written this thesis on my own and that I have not used any other media or materials than the ones referred to in this thesis.

Saarbrücken, Jan 05, 2025,

(Hafeez Ullah)

Einverständniserklärung

Ich bin damit einverstanden, dass meine (bestandene) Arbeit in beiden Versionen in die Bibliothek der Informatik aufgenommen und damit veröffentlicht wird.

Declaration of Consent

I agree to make both versions of my thesis (with a passing grade) accessible to the public by having them added to the library of the Computer Science/ Bio-informatik Department.

Saarbrücken, Jan 05, 2025,

(Hafeez Ullah)

I dedicate my work to my family; Dad, Mom, and my beautiful sister. You all are blessings for me. Thank you for your prayers, sacrifices, encouragement and the love you have given me throughout my life.

Abstract

Melanoma is one of the deadliest forms of skin cancer, making early detection critical for improving patient outcomes. However, accurately distinguishing between malignant and benign skin lesions remains a significant challenge for clinicians. This thesis introduces an automated skin lesion classification system designed to enhance the accuracy of melanoma detection using dermatoscopic images. A unique approach is proposed, which considers multi-lesion images from the same patient and incorporates metadata information to classify lesions as benign or malignant.

To achieve this, this thesis leverages state-of-the-art deep learning models, including EfficientNet V2, EfficientNet B7, and ResNet-50, selected for their balance of efficiency and feature extraction capability. The project employs advanced data preprocessing techniques, such as malignant-specific augmentation and stratified batching, to address class imbalance and improve model generalizability. Training utilizes Focal Loss to focus on challenging samples, further helping to mitigate class imbalance.

The proposed system is evaluated using metrics such as ROC AUC, precision-recall, and F1 score. Notably, model confidence analysis provides insights into the reliability of classifications, highlighting the model's strengths in distinguishing hard-to-classify samples. The automated skin lesion classification tool developed in this thesis achieves an accuracy of above 95% and an AUC of 0.95, demonstrating its potential as an effective support tool for dermatologists.

These results show promising advancements in melanoma detection, and the system could serve as a foundational tool for future researchers to build upon.

Acknowledgements

I pay special gratitude and thanks to Almighty ALLAH for what I have achieved and consider each and everything as a gift from him.

Now, I would like to express my sincere respect and gratitude towards my supervisor. Prof. Dr. Alexey Gurevich and Prof. Dr.-Ing. Paul Motzki for their guidance and continuous help during all this time. Without your support, this work would not have been possible. You have been a constant source of knowledge and inspiration. I hope to be, one day, as inspirational to others as you have been to me. I would also thank Prof. Dr.-Ing. Paul Motzki for reviewing my work.

Furthermore, I am also thankful to my thesis advisor, Tajbeed Ahmed Chowdhury for his valuable time and feedback throughout my thesis. My acknowledgement also goes to Tajbeed Ahmed Chowdhury in arranging valuable meetings with supervisor professors. I would also like to thank and express my appreciation to all the people who were involved directly or indirectly and guided me with their efforts and skills. Special thanks to Aftab Alam, thank you for helping with my conversational writing style and for your feedback in that regard.

I would also like to thank my friend Taha Jamil for his moral support and encouragement. Finally, my deepest gratitude goes to my parents, especially (late) dad– you are my real hero, you sacrificed everything for me, your wealth, health, and time for seeing me in this place; and mom I would express my love for you with the Khalil Gibran quote: "Mother: the most beautiful word on the lips of mankind". To my siblings for their love and support during my study and personal life.

Abbreviations and Acronyms

AI Artificial Intelligence

AUC Area Under Curve

CAD Computer-Aided Diagnosis

CNN Convolutional Neural Network

CT Computed Tomography

CUDA Compute Unified Device Architecture

DICOM Digital Imaging and Communications in Medicine

DL Deep Learning

DNNs Deep Neural Networks

FCNN Fully Convolutional Neural Network

fMRI Functional Magnetic Resonance Imaging

GPU Graphics Processing Unit

GUI Graphical User Interface

ISIC International Skin Imaging Collaboration

k-NN k-nearest neighbors

ML Machine Learning

MRI Magnetic Resonance Imaging

MBConv Mobile Inverted Bottleneck Convolution

NIfTI Neuroimaging Informatics Technology Initiative

PCA Principal Component Analysis

ReLU Rectified Linear Unit

ResNet Residual Neural Network

RF Random Forest

ROI Region of Interest

ROC Receiver Operating Characteristic

R-CNN Region Based Convolutional Neural Networks

RNN Recurrent Neural Network

ROS Random Over-Sampling

RUS Random Under-Sampling

SMOTE Synthetic Minority Oversampling Technique

SVM Support Vector Machine

TL Transfer Learning

VGGNet Visual Geometry Group Neural Network

WHO World Health Organization

YOLO You Only Look Once

Contents

Abstract **v**

Acknowledgements **vi**

1 Introduction **1**
1.1 Motivation 5
1.2 Machine Learning techniques for detecting skin lesion 8
1.3 Mathematical formulations 11
1.4 Structure of the thesis 16

2 Related Work **17**
2.1 A review of Deep Learning in medical image analysis 17
2.2 A review of Deep Learning network architectures for medical image analysis 19
2.3 Segmentation techniques for skin lesion analysis 20
2.4 ResNet and EfficientNet in biomedical imaging 24
2.5 Transfer Learning in biomedical image classification 26
2.6 Skin lesion classification on imbalanced dataset 28

3 Methodology **31**
3.1 Introduction to methodology 31
3.2 Data collection and preprocessing 35
3.2.1 Data source description 35
3.2.2 Data preprocessing steps 36
3.2.3 Focal loss for addressing class imbalance 45
3.3 Model architecture 47
3.3.1 Overview of EfficientNet architecture 48
3.4 EfficientNet-B7: Approach for skin lesion classification 51
3.4.1 Transfer Learning with EfficientNet-B7 51
3.4.2 Metadata integration for enhanced predictions 52
3.4.3 Training and validation configurations 52
3.4.4 Evaluation metrics 54
3.5 Semi-Supervised prediction mechanism in the GUI tool 56
3.5.1 Role of the `.pth` file 57
3.5.2 Prediction mechanism 59

4 Results and Discussion **62**
4.1 Chapter overview . . . 62
4.2 Experimental recap . . . 62
4.3 Results: ResNet-50 . . . 63
4.3.1 Training and validation performance . . . 63
4.3.2 Observations and challenges . . . 65
4.4 Results: EfficientNet-B7 . . . 65
4.4.1 Training and validation performance . . . 65
4.4.2 Observations and potential reasons . . . 67
4.5 Results: EfficientNet-V2 . . . 68
4.5.1 Training and validation performance . . . 68
4.5.2 Observations and explanation . . . 70
4.5.3 Confusion Matrix analysis . . . 70
4.6 Comparative evaluation and critical discussion . . . 79
4.6.1 Quantitative comparison . . . 79
4.6.2 Qualitative insights with related study . . . 80
4.7 GUI tool-based predictions . . . 81
4.8 Revisiting model performance: Why EfficientNet-B7 may outperform ResNet-50 . . . 83
4.9 Reasoning for using evaluation metrics . . . 84
4.9.1 Model confidence analysis . . . 85
4.9.2 Gaps and limitations . . . 85
4.9.3 Future work . . . 85
4.10 Conclusion . . . 86

5 Conclusion **87**
5.1 Summary of the thesis . . . 87
5.2 Contributions to the field . . . 88
5.2.1 Two-pass training for class imbalance . . . 88
5.2.2 Metadata-enhanced architecture . . . 88
5.2.3 Comprehensive evaluation of state-of-the-art models . . . 89
5.3 Limitations and future directions . . . 89
5.4 Concluding remarks . . . 90

List of Figures **91**

List of Tables **92**

Bibliography **93**

Chapter 1

Introduction

According to past research, melanoma causes a lot of the deaths which are linked to skin cancer worldwide. Although it is among the most deadly forms of skin cancer, timely identification and treatment help greatly to lessen its effects [1]. Melanoma must be identified at early stages since it greatly increases the likelihood of successful detection, hence producing better patient outcomes. But accurate melanoma diagnosis requires major challenges, such as lesion variability and imaging differences, which highlight the need for innovative Artificial Intelligence (AI)-driven solutions. The need for automated and reliable diagnostic aid systems is becoming more and more important in order to enable dermatologists to exactly detect benign skin lesions. From advances in AI, particularly in Deep Learning (DL), medical imaging has immense potential. By means of these technologies, easily accessible, fast, and efficient early diagnosis techniques could be provided, therefore perhaps saving lives and reducing the burden of treatment.

The aim of this thesis is to create an automated skin lesion classification system using multiple images from the same patient and using metadata information to identify and classify melanoma and benign images. The DL based technique uses dermatoscopic images to differentiate benign from malignant lesions. Advanced DL methods are used to solve the problems caused by the restricted availability of annotated datasets and fluctuating lesion appearances. This thesis aims to provide a tool that future researchers may build upon by combining modern architectures. This thesis implements architectures such as EfficientNet [2] and Residual Neural Network (ResNet)[3] and uses advanced augmentation techniques.

Machine Learning (ML) uses data-driven, statistical rules that are created from a large sample size instead of explicit human descriptions. ML turns algorithm inputs into outputs. Building ML systems historically needed the creation of feature extractors to translate unprocessed input into meaningful representations and a learning algorithm

that could utilise for pattern recognition. This method mostly depends on human knowledge and experience as well as domain-specific competence. Unlike this, DL is a form of representation learning whereby a machine uses unprocessed data to create the representations required for pattern recognition. These images are built from numerous layers. One layer's (beginning with the raw data input) representation is passed into the next layer and transformed into a more abstract form. Usually arranged sequentially, these layers consist of a lot of primitives and nonlinear transformations. Extremely complex functions can be learnt by continually wrapping the input space as data moves through the layers until data points can be distinguished. Hence, complex functions can be learnt.

For ML, early 21st century represented a turning point. Three trends over recent years have shown this. Big data was the first trend since a new method was launched when the volume of data was so enormous. This brought the second trend which was lowering the cost of parallel computing and related memory. First, we found that Hadoop effectively enabled the distribution of enormous volumes of data between basic CPUs. Then followed by the NVIDIA and AMD with their innovative ideas for the Graphics Processing Unit (GPU) market. NVIDIA's Compute Unified Device Architecture (CUDA) provided a notable increase as GPU computing in ML proved to be really outstanding. CUDA let developers have direct access to the virtual instruction set and parallel processing elements in NVIDIA GPUs, therefore enabling significant computing speed improvements. For ML practitioners and researchers, this discovery democratised access to high-performance computing, therefore arming researchers and developers with the capabilities to innovate at hitherto unheard-of rates. Deep ML, sometimes referred to as DL, has been the third trend that developed recently. Since DL algorithms are based on Deep Neural Networks (DNNs), which are presented by British scientist Geoffrey Hinton [4], they need enormous computational capability. Working on neural network models, he revolutionized the domains of AI and ML. These developments of algorithms presented a fresh processing power challenge, which has possibly been overcome by integrating NVIDIA GPUs.

ML has three techniques. These are supervised, unsupervised, and reinforcement learning. The widely used technique of ML is supervised learning, which deals with two types of problems: classification problems and regression problems. Supervised learning trains a machine with labelled data. Labelled datasets have input and output parameters. There are algorithms for regression. Few of them are linear and nonlinear regression, splines, and Bayesian. There are algorithms for classification as well. Some examples includes naive bayes, Support Vector Machine (SVM), and logistic regression. Unsupervised learning is where the data is unlabeled. Its main aim is to explore the underlying patterns and predict the output. In unsupervised learning, a few of the techniques are

k-means clustering, neural networks, and Principal Component Analysis (PCA). In essence, the key distinction between supervised and unsupervised learning lies in the use of labelled data. Using labelled training data where both input and matching output labels are given; supervised ML lets the model learn a mapping between inputs and outputs. Unsupervised learning operates on unprocessed, unlabelled data, focussing on identifying hidden patterns, structures, or relationships within the data, independent of predetermined labels. This thesis implements both supervised and unsupervised techniques in a semi-supervised approach for skin lesion prediction. The methodology is divided into two stages: the first stage employs supervised learning to train the model on labelled skin lesion image data, incorporating metadata such as age, sex, and anatomical site to enhance the model's understanding and accuracy. In the next stage, a semi-supervised-based approach is being implemented for the final predictions of melanoma and benign; this allows the model to leverage both labelled and unlabelled data. This comprehensive combined strategy enhances the model's generalizability and well as its robustness. Hence it is better suited for real-world applications. In the methodology chapter, this thesis discusses the combination of both the techniques in detail. Lastly, the third technique is reinforcement learning, which is learning from the environment itself. This technique is implemented in robotics and games. Reinforcement learning reflects many real-world scenarios. There is an agent in the environment that is allowed to do actions on the environment. Based on its actions, there is a reward system. Reinforcement learning is widely deployed in self-driving cars.

DL has made a remarkable comeback in recent years, primarily due to the increased availability of large-scale datasets and the improvement in computing capacity. Particularly in the discipline of processing unstructured data, where establishing rules and correlations is vital, significant advancement has been accomplished in the subject relating to machine understanding and data manipulation. This is inspired by human biological neurons, in DL-based system it is created on a whole representation of comprehensive input data of linked nodes. These flexible DL algorithms can process multiple types of data, including text, images, and audio. Applications of DL can be represented in text recognition, voice analysis, image and video processing, decision-making, and autonomous driving.

In contrast to traditional ML, DL enables multi-layer computational models to learn data representations by processing them in their original form. Convolutional Neural Network (CNN) increased the accuracy of several image processing domains, including object detection, speech recognition, face recognition, and medicinal applications. One benefit of DL is raw data feature extraction automation. Although conventional ML methods have found application in many fields, their scalability for challenging image-based tasks is limited since they mostly depend on human feature extraction and domain knowledge. DL is especially appropriate for uses like skin lesion analysis since it automatically learns

hierarchical feature representations, therefore overcoming these constraints.This capability allows DL to manage unstructured inputs, including images, text, and audio, without requiring either wide domain expertise or hand feature engineering. Fully Convolutional Neural Network (FCNN) is defined by convolution, pooling, and activation methods including linear activation, Rectified Linear Unit (ReLU), and logistic activation. The last node of the architecture presents the output result. In CNN design, DL techniques have brought major architectural breakthroughs that transform image processing chores. AlexNet was one of the first innovations since it raised CNN depth from 5 to 8 layers, therefore redefining image categorisation standards. This deeper design enhanced the generalizing capacity of the network over several imaging resolutions. The higher depth, however, also presented difficulties such overfitting, which emphasises the necessity of further CNN design innovation. With 19 layers Visual Geometry Group Neural Network (VGGNet), more especially the VGG-19 variant, is built differently than AlexNet. Although VGGNet significantly altered the parameters of AlexNet, its deployment on systems with limited resources is challenging and computationally costly. Moreover, GoogleNet was developed in order to obtain great accuracy at a low computing cost. It presented CNN's novel inception block concept, which combines split, transform, and merge ideas to perform multiscale convolutional transformations. Then comes ResNet, which performed highly accurate on problems involving localization and image recognition. It demonstrated that for a large number of visual identification tasks, representational depth is crucial. Few of the similar are EfficientNet, U-Net, Region Based Convolutional Neural Networks (R-CNN), and You Only Look Once (YOLO) for imaging problems. Recurrent Neural Network (RNN) are a different kind of neural network that are used in DL and make predictions using sequential or time-series data.

The biomedical domain of ML implementation is the main emphasis of this thesis. In the medical sector, AI and ML have demonstrated encouraging development in enhancing product purity, raising manufacturing efficiency, and lowering the cost of production. Though this field has seen great progress, ML has not yet realised its full potential. Healthcare providers must embrace the present state of AI and ML technologies and their possibilities to increase the safety, efficacy, and accessibility of healthcare services if they are to attain value-based care. Among the many advantages AI offers the healthcare sector are improved patient care, more accurate diagnosis, more operational efficiency, and cost savings. Most of the ML advances in healthcare to date have helped doctors and analysts to identify anomalies in the field of medicine. Two examples of the kinds of data ML models can use to build domain expertise are medical imaging; Computed Tomography (CT) scans and Magnetic Resonance Imaging (MRI) even at doctor-level observations these are used on daily basis. This general progress is driving researchers to create stronger ML and DL models to handle increasingly more complicated and vast

amounts of medical data. From a range of sources, including health records, biometric data, medical imaging, and prognosis, medical data is fast expanding and easily accessible. Reliable datasets assist ML models to guarantee correct diagnosis of patients by means of prediction. Health sector ML implementations have brought personalised medication, diagnostics, and predictive analytics.

This thesis shows the successful implementation of ML application on skin lesion detection. This thesis aims to classify skin lesions as malignant or benign using multiple images from the same patient. To intercept this skin lesion, EfficientNet and ResNet have been implemented. There have been various researches on the detection of melanoma from a single image. The primary goal of this thesis is to develop a system capable of detecting malignant or benign lesions by analyzing multiple images associated with the same patient. This method improves diagnosis accuracy by including patient-specific metadata along with image analysis. The system tries to solve problems in present diagnostic techniques and offer a more dependable accuracy by using both visual and contextual information. This thesis's main emphasis is on multi-lesion analysis since most melanoma detection techniques concentrate on individual lesions, therefore ignoring this aspect. This thesis addresses present constraints and provides a strong basis for future scholars to build upon using the suggested method, therefore enabling improved solutions. These initiatives seek to improve early identification and diagnostics of skin lesions, therefore enabling more accurate and dependable diagnosis.

1.1 Motivation

Skin cancer melanoma, is one of the most deadly types of cancer. Usually originating from inherited disorders, abnormal cell development leads to malignant tumours that, if not found early on, can spread through the body. The high melanoma fatality rate underlines the importance of accurate and fast detection [5]. These issues highlight the need for automation, especially with developments like Computer-Aided Diagnosis (CAD) systems, which use strong algorithms to precisely and reliably examine dermatoscopic images. Although in some situations these conventional approaches are helpful, they cause more patient loads and demand on healthcare professionals; therefore, delaying diagnosis and treatment can lead to fatality.

From benign moles and freckles to harmful melanomas, skin lesions across a spectrum of diseases generate significant global health issues. The World Health Organization (WHO) shows that, with millions of non-melanoma skin cancer cases recorded annually, skin melanoma is the most frequently diagnosed malignant disease worldwide. Early detection and accurate diagnosis of skin lesions are absolutely vital, as even if melanoma has a high

survival rate when discovered early, as the lesion develops, the survival rate decreases. These figures show how urgently effective diagnostic tools are needed to improve patient outcomes and reduce the burden on medical systems.

Initially, skin lesion identification depended on visual assessments by dermatologists, sometimes assisted by dermatoscopy, a technique using a dermatoscope to enlarge and light up the skin, therefore improving the observation of skin lesions. Although there are benefits, the experience of the doctor greatly influences the accuracy of dermatoscopy. When doctors are dealing with several lesions by the same patient, this procedure might frequently produce incorrect diagnoses due to the significant possibility for human error. High-resolution images of skin lesions may now be obtained, stored, and analysed; digital dermoscopy greatly influences the diagnosis process. Lack of standard imaging techniques and variations in data quality and interpretation make diagnostic processes less reliable. These challenges underline the need for dermatological imaging guidelines.

Development in CAD systems has brought a revolutionary method of skin lesion analysis. Using sophisticated algorithms, CAD systems process digital dermoscopic images, therefore providing improved accuracy and less dependence on human interpretation. Modern ML, especially DL, has transformed the discipline even more recently. For DL models, tasks such as lesion segmentation, classification, and feature extraction have witnessed incredible performance, including CNNs. Underlining the potential of DL in increasing diagnostic accuracy, CNNs could compete with dermatologists in skin cancer classification [6]. Research by Codella et al. [7] and Haenssle et al. [8] also demonstrated how DL models outperform conventional approaches, therefore considerably improving clinical decision-making.

There are many existing approaches to skin lesion detection that are constrained by their focus on individual lesions, disregarding the broader diagnostic context that could be achieved through the analysis of multiple lesions from the same patient. This restriction reduces diagnostic accuracy and ignores essential contextual information needed for early detection and accurate diagnosis. The development of strong and generalizable diagnostic models is further challenged by variation in image quality, acquisition conditions, absence of standardised data formats, and limited annotated datasets.

Using a standardised medical dataset, this thesis suggests a system meant to overcome these restrictions by analysing several lesion images from the same patient. Using the ISIC 2020 dataset [9], it guarantees data consistency and quality in processing by conforming with the Digital Imaging and Communications in Medicine (DICOM) criteria. The integration of such a standard dataset improves the dependability and implementation of the model. This conforms with accepted medical imaging. DICOM is internationally accepted for the storage consistency, transmission, and management of medical imaging

data. It offers a complete framework combining extensive metadata, including patient information, imaging parameters, and diagnostic notes, with image data. This dual storage capability ensures seamless interoperability across healthcare systems. DICOM's structure is uniquely suited to medical imaging due to its ability to store both image data and critical metadata within a single file. This consistency is vital for building robust ML models reliant on high-quality data.

Moreover, DICOM's supports its compatibility with ultrasonic waves and several imaging modalities, including CT and MRI. Furthermore, supported by the standard techniques of data compression, therefore enabling efficient storage and transfer without compromising image quality. DICOM ensures the consistency of input data, so maintaining the visual and contextual integrity of dermatoscopic images. Although DICOM shines in clinical settings because of its extensive metadata and fit with many imaging modalities, Neuroimaging Informatics Technology Initiative (NIfTI) is intended for neuroimaging and research-orientated uses. While DICOM analyses slices of the human body using a layer-based structural ideal, NIfTI saves data in a 3D format, therefore optimising spatial orientation and processing performance. This allows NIfTI for such purposes, including those Functional Magnetic Resonance Imaging (fMRI) that require fast data processing and analysis. Though in research scenarios, its simplicity and speed make NIfTI a better choice. Knowing these supporting strengths of DICOM and NIfTI allows researchers to choose the most suitable format depending on their current research requirements.

Using the consistent data that these standards offer, this thesis applies sophisticated DL models to dermoscopic image analysis. ResNet [10] is applied for model architecture since it solves the vanishing gradient issue, allowing the training of deeper networks. EfficientNet [2] improves performance with less parameters by means of its optimal balance of depth and width. Lesion borders are drawn using U-Net [11], renowned for its mastery of image segmentation chores. Generalizability of the model is enhanced by methods of data preparation, including normalising and augmenting. Iterative testing helps to fine-tune hyperparameters; the performance of the model is assessed on another test set using conventional evaluation criteria, including F1 score, recall, and precision.

This thesis integrates a user-friendly Graphical User Interface (GUI) to support useful applications. The GUI provides users an easily available tool for skin lesion diagnosis by letting them upload and examine several lesion images. This work intends to increase melanoma diagnosis accuracy by combining regular imaging methodologies and standarized dataset formats. Because of its durability and thorough metadata preservation, which guarantees important clinical information is kept throughout preprocessing and analysis, DICOM is the recommended file format under focus in this thesis. This thesis presents a strong basis for training DL models by using DICOM's organising

and transmission strengths, therefore providing constant and high-quality inputs for melanoma detection. This work presents advancements in diagnostic techniques and supports developments in skin lesion identification.

1.2 Machine Learning techniques for detecting skin lesion

In previous diagnosis melanoma relied mostly on manual inspections and the knowledge of specialist dermatologists, usually employing instruments like dermatoscopy to visually examine skin lesions. These methods are constrained by subjectivity and possible human mistake even if they have shown success in some situations. These restrictions draw attention to the necessity of automated diagnostic instruments able to improve the accuracy and efficiency of skin lesion identification.

In medical image analysis, ML has become a transforming technology providing answers for these problems. Early ML methods for skin lesion identification mostly applied handcrafted features mixed with conventional classifiers such SVM, k-nearest neighbors (k-NN), and Random Forest (RF). These techniques needed significant feature engineering to derive from dermoscopic images relevant characteristics including colour, texture, and shape. For melanoma diagnosis, Abbas et al. [12] for example showed the effectiveness of employing an SVM classifier with personally specified features, hence surpassing simple thresholding methods.

DL eliminated the need for manual feature engineering. It has transformed skin lesion detection to automatic detection. Excelling in tasks including lesion segmentation, classification, and feature extraction, CNNs have grown to be the pillar of DL applications in medical imaging. CNNs instantly learn hierarchical representations from raw visual data, unlike conventional methods, therefore catching complex patterns are vital for a correct diagnosis. Esteva et al. [6] trained a model able to classify skin lesions with dermatologist-level accuracy, so proving the potential of CNNs. Using a collection of more than 129,000 clinical images, their results highlighted the need for big and varied datasets in training DL models for medical uses.

In situations when annotated medical images are scarce, Transfer Learning (TL) has improved the relevance of DL even more. Researchers have obtained notable performance enhancements with little computing demand by fine-tuning pre-trained models such as GoogleNet and ResNet using domain-specific datasets. Menegola et al. [13] for instance, used pre-trained models to obtain competitive results even with somewhat small datasets by applying TL to melanoma diagnosis. This method is absolutely essential in medical image analysis as it reduces overfitting and increases model convergence.

Even with these developments, significant challanges remain in creating strong ML models for skin lesion detection. Significant challenges include variations in lesion appearances, artefacts include hair or skin markings, and inconsistent imaging circumstances. Lack of annotated datasets is a key obstacle in building strong diagnostic models. Establishing large datasets requires a lot of time and domain knowledge, which emphasises the need of using methods including synthetic data generation, data augmentation, and TL to minimise this restriction. Dealing with these problems, this thesis uses model adaptation, TL, and advanced data augmentation methods to increase generalization and resilience.

There is another successful method for improving prediction accuracy is called ensemble learning. It combines the results of several models and allows those ensemble techniques to use the advantages of particular architectures to raise general accuracy and resilience. Using an ensemble of models incorporating ResNet and DenseNet to attain state-of-the-art results in melanoma detection, codella et al. [14] showed the usefulness of this method in the International Skin Imaging Collaboration (ISIC) challenge. Generative adversarial networks (GANs) have also shown to be rather helpful for enhancing datasets. Using GANs, Bissoto et al. [15] produced realistic synthetic dermoscopic images, therefore overcoming the issue of limited labelled data and improving model training.

There is another interesting approach which is hybrid models; combining CNNs with other ML methods or incorporating multi-modal data, such clinical and dermoscopic images. To examine lesion sequences across time, Gessert et al. [16] developed a hybrid model combining CNNs with Recurrent Neural Networks (RNNs). This method considers the progression of lesions to capture temporal changes in them, therefore enhancing melanoma detection. Although these advances have problems, including class imbalance and inconsistent imaging data, still exist. Using modern architectures such as EfficientNet V2, B7, and ResNet-50 in addition to specialised augmentation methods including malignant-specific augmentations and stratified batching, this thesis solves these limitations.

This thesis additionally applies a model confidence score to grasp prediction dependability more comprehensively. Examining confidence scores helps the thesis to reveal the strengths and shortcomings of the performance of the model, thereby providing a whole strategy for the detection of skin lesions. Moreover, the usage of an easy-to-use GUI guarantees useful applicability, thereby enabling users to easily upload and examine multiple lesion images. This thesis presents automated melanoma classification since it addresses present limitations and lays a basis for future developments by the integration of of advanced DL algorithms, standardised datasets, and easy tools.

To lay out the methods used in this thesis, it is essential to introduce the mathematical principles supporting the suggested methods. An in-depth analysis The mathematical

formulations offer understanding into the operational principles of the ML models and a thorough comprehension of their effectiveness and constraints. Fundamental mathematical formulas, metrics for performance, and optimization strategies that serve as the basis of the employed algorithms for skin detection classification are described below. By means of the formalization of these ideas, this thesis aims to present a theoretical framework that improves the hands-on implementation and offers a comprehensive perspective on the effectiveness and strength of the recommended solutions.

1.3 Mathematical formulations

1. **Focal Loss**

$$FL(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t) \tag{1.1}$$

Where:

- p_t is the model's estimated probability for the class with the true label.
- α_t is the weighting factor for addressing class imbalance.
- γ is the focusing parameter that controls the contribution of easy vs. hard examples.

2. **F1 score**

$$F1 = 2 * (precision * recall)/(precision + recall) \tag{1.2}$$

$$Precision = \frac{true_positives}{true_positives + false_positives}$$

$$Recall = \frac{true_positives}{true_positives + false_negatives}$$

3. **Accuracy**

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{True Negatives} + \text{False Positives} + \text{False Negatives}} \tag{1.3}$$

4. **Receiver Operating Characteristic - Area Under the Curve**

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR}),$$

where TPR denotes the true positive rate (also called *Recall* or *Sensitivity*) and FPR denotes the false positive rate, expressed as a function of the decision threshold.

An equivalent discrete formulation for AUC, often associated with the Wilcoxon Rank-Sum interpretation, is shown below:

$$\text{AUC} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbb{I}(s_i > s_j), \tag{1.4}$$

where:

- m is the number of positive samples,
- n is the number of negative samples,
- s_i is the model-assigned score for the i-th positive sample,
- s_j is the model-assigned score for the j-th negative sample,
- $\mathbb{I}(s_i > s_j)$ is the indicator function, returning 1 if $s_i > s_j$, and 0 otherwise.

5. **Precision-Recall Curve**

The Average Precision (AP) score is calculated as follows:

$$AP = \sum_{n} (R_n - R_{n-1}) P_n \tag{1.5}$$

Where:

- P_n is the precision at the n-th threshold.
- R_n and R_{n-1} are the recall values at the n-th and $(n-1)$-th thresholds.

6. **Average Precision (AP)**

The Average Precision (AP) is calculated as follows:

$$AP = \sum_{n=1}^{N} (R_n - R_{n-1}) P_n \tag{1.6}$$

Where:

- N is the number of thresholds.
- P_n is the precision at the n-th threshold.
- R_n and R_{n-1} are the recall values at the n-th and $(n-1)$-th thresholds.

7. **Model Confidence Analysis**

The model confidence analysis metric includes the following components:

$$C_{\text{correct}} = \frac{1}{N_{\text{correct}}} \sum_{i=1}^{N_{\text{correct}}} c_i$$
$$C_{\text{incorrect}} = \frac{1}{N_{\text{incorrect}}} \sum_{i=1}^{N_{\text{incorrect}}} c_i$$

Where:

- C_{correct} is the average confidence for correctly classified samples.
- $C_{\text{incorrect}}$ is the average confidence for incorrectly classified samples.
- N_{correct} is the number of correctly classified samples.
- $N_{\text{incorrect}}$ is the number of incorrectly classified samples.
- c_i is the confidence score assigned by the model for the i-th sample.

8. **Binary Cross Entropy with Logits Loss (BCEWithLogitsLoss)**

The BCEWithLogitsLoss is calculated as follows:

$$L = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(\sigma(x_i)) + (1 - y_i) \log(1 - \sigma(x_i))] \tag{1.7}$$

Where:

- N is the number of samples.
- y_i is the target label for the i-th sample ($y_i \in \{0, 1\}$).
- x_i is the raw prediction (logit) from the model for the i-th sample.
- $\sigma(x_i)$ is the sigmoid function applied to the logit x_i, defined as:

$$\sigma(x_i) = \frac{1}{1 + e^{-x_i}} \tag{1.8}$$

9. **Binary Cross Entropy (BCE) Loss**

The Binary Cross Entropy (BCE) Loss is calculated as follows:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p(y_i)) + (1 - y_i)\log(1 - p(y_i))\right] \tag{1.9}$$

Where:

- N is the number of samples.
- y_i is the target label for the i-th sample ($y_i \in \{0, 1\}$).
- $p(y_i)$ is the predicted probability (after applying a sigmoid activation) for the i-th sample.

10. **Adam Optimizer**

The Adam (Adaptive Moment Estimation) optimizer is defined as follows:

$$\begin{aligned}
m_t &= \beta_1 m_{t-1} + (1 - \beta_1) g_t \\
v_t &= \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \\
\hat{m}_t &= \frac{m_t}{1 - \beta_1^t} \\
\hat{v}_t &= \frac{v_t}{1 - \beta_2^t} \\
\theta_{t+1} &= \theta_t - \frac{\alpha \hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}
\end{aligned}$$

Where:

- m_t and v_t are the first moment (mean) and second moment (variance) estimates at time step t.
- β_1 and β_2 are the exponential decay rates for the moment estimates, typically set to values like 0.9 and 0.999 respectively.
- g_t is the gradient of the objective function at time step t.
- $\hat{m}_t$ and $\hat{v}_t$ are the bias-corrected estimates of the first and second moments.
- α is the learning rate.
- ϵ is a small constant (e.g., 10^{-8}) used for numerical stability.
- θ_t represents the parameters being optimized, and θ_{t+1} are the updated parameters.

11. **Stochastic Gradient Descent (SGD)**

The parameter update rule for Stochastic Gradient Descent (SGD) is given by:

$$\theta_{t+1} = \theta_t - \alpha \nabla_\theta J(\theta_t; x_i, y_i) \tag{1.10}$$

Where:

- θ_t are the parameters at time step t.
- α is the learning rate, which controls the step size of the parameter update.
- $\nabla_\theta J(\theta_t; x_i, y_i)$ is the gradient of the loss function J with respect to the parameters θ, evaluated at the i-th training example (x_i, y_i).
- x_i and y_i represent the input features and the corresponding label of the i-th training example.

12.**confusion matrix**

True Labels	**Predicted Labels**		
	Positive	**Negative**	**Total**
Positive	$true_positives$	$false_negatives$	$positives$
Negative	$false_positives$	$true_negatives$	negatives
Total	P'	N'	$P + N$

Table 1.1: Confusion Matrix for Binary Classification

Where:

- TP is the count of True Positives (correctly classified as positive).
- FN is the count of False Negatives (incorrectly classified as negative).
- FP is the count of False Positives (incorrectly classified as positive).
- TN is the count of True Negatives (correctly classified as negative).
- P is the total number of actual positive samples $(TP + FN)$.
- N is the total number of actual negative samples $(FP + TN)$.
- P' is the total number of predicted positive samples $(TP + FP)$.
- N' is the total number of predicted negative samples $(FN + TN)$.

1.4 Structure of the thesis

The rest of the thesis is organised as follows:

Chapter 2: This chapter reviews related work on the thesis topic. It examines previous studies in a similar area, explaining how earlier methods were applied to related problems and what results were achieved. It provides a detailed review of the relevant literature.

Chapter 3: This chapter discusses the research goals and techniques applied in the thesis; it covers the techniques for preprocessing and data collecting, the implementation process, the applied DL models, and the GUI tool development for prediction generation.

Chapter 4: The results are discussed in this chapter. It presents evaluations utilising several performance criteria together with an analysis of the proposed three DL-based techniques.

Chapter 5: Recapping the findings, noting constraints, and suggesting future directions for researchers.

Chapter 2

Related Work

There are numerous challenges in medical imaging, such as identifying differences in lesion appearances, similarity between benign and malignant features, and the main challenges in obtaining sufficiently vast, annotated datasets make melanoma identification tough. These challenges have motivated scientists to develop more sophisticated techniques to help doctors and raise diagnosis accuracy.

The current work on automated skin lesion recognition and classification is presented in this chapter. It offers a thorough introduction of both more modern DL methods and conventional ML procedures. This part aims to grasp the development of the AI-driven medical research, the development of techniques, and the way researchers have helped to solve the particular difficulties with melanoma diagnosis. Moreover, this chapter will show areas in which current research still lags, thereby guiding the contributions produced in this thesis.

2.1 A review of Deep Learning in medical image analysis

Simple image processing methods such as low-level filters for edge and line identification or region-growing algorithms from the 1970s through the 1990s, dominated medical image analysis. By the late 1990s, supervised learning methods were important in building a system able to do feature extraction from training data. Early instances of this were segmentation using dynamic shape models. The development of feature extraction methods marked a significant shift by enabling more advanced pattern identification, thereby enhancing the capacity of ML in medical image analysis. Consequently, during this time several effective, commercially available medical image analysis technologies evolved.This advancement indicated a shift from systems developed just by human engineers to ones able to autonomously learn features from data.

Over the past ten years, DL has rapidly transformed the discipline of medical image analysis. In numerous spheres of medicine, including neuroimaging, retinal evaluations, pulmonary imaging, digital pathology, cardiac analysis, breast cancer screening, and tumour detection [17]. DL focuses on automatically finding, classifying, and quantifying patterns in medical images, thus supporting effective image reconstruction, segmentation, and post-processing [18]. From disease detection to improved treatment monitoring, applications of DL in medical imaging span vast amounts of medical data, including X-rays, CT scans, MRI, and other imaging modalities into actionable clinical information [19], [20], [21].

In spite of this, the achievements of DL, some intrinsic difficulties persist that necessitate consideration. The restricted availability of annotated medical data is a key issue that hinders the training of strong models. Annotating medical images requires specialised knowledge and is labour-intensive, complicating the collection of adequately big datasets. Methodologies including unsupervised and semi-supervised learning have become rather popular. These methods generate additional tagged images, detect important patterns in unlabelled data, and provide pseudo labels to enable model improvement, therefore addressing the limitations of small labelled datasets. Recent breakthroughs have shown that, even with less data, DL may attain great accuracy and assist in multiple aspects of medical imaging, ranging from early disease identification to personalised treatment [22, 23].

DL has also significantly improved classification of medical images. A fundamental aspect of computer vision has its applications in identifying diseases across a wide spectrum of medical domains. For skin disease classification in dermatology [24, 25], eye disease detection in ophthalmology, such as diabetic retinopathy [26], glaucoma [27], and corneal diseases [28] and cancer detection, including breast [29] and brain cancers [30]. Moreover establishing success in predicting future acute kidney damage in patients is DL [31]. These advances highlight the general applicability of DL and its ability to exceed more traditional methods in accuracy and efficiency. Xuxin et al. [32] explored medical disease classification [33], Region of Interest (ROI) segmentation [34], medical object detection [35], and image registration [36]. These use cases from past work show how flexible and dependable DL methods are in managing various medical tasks. One of the main challenges in medical image analysis, despite these developments, is the small scale of standarized datasets.

Having investigated the transforming power of DL in medical image processing, this section clarifies particular architectures like CNNs and their developments, which form the backbone of many diagnostic applications.

2.2 A review of Deep Learning network architectures for medical image analysis

Hinton et al. [4] introduced the backpropagation; he significantly improved the training of neural networks, enabling more effective weight adjustment and making DL practical for real-world applications. CNNs, which were inspired by biological neurons, revolutionised the domain of image analysis, including medical imaging. CNNs introduced an advanced way to extract spatial hierarchies of features, making them particularly powerful tools for visual tasks, including medical diagnostics.

Medical imaging in the context of CNNs has shown a promising success. CNNs showed the ability to learn on deep and hierarchical features from images, which has helped them to outperform human experts in some medical image processing tasks. Underlining its tremendous use and power for increased diagnostic accuracy. Raghavendra et al. [37] studied CNN use in medical imaging, underlining their great usage and power for improved diagnosis accuracy. Medical uses include the X-ray and CT scan skin lesion diagnosis, which have shown considerable accuracy improvements in CNN designs. Many studies, specialised conferences, and focused special issues in prestigious medical journals have helped CNNs to be extensively implemented.

CNN's success has upgraded the design of new architectures to improve model performance even farther forward. One of the key CNN architectures include AlexNet [38], which started the trend of deeper networks; VGGNet [39], which improved AlexNet's structure and enhanced depth for best performance; and ResNet [3], which used skip connections to offset the vanishing gradient issue. Later models such as DenseNet [40] and EfficientNet [2] improved the domain by increasing model accuracy and so optimising efficiency. Medical image processing operations like segmentation, detection, and classification now follow CNN-based designs as the norm.

CNNs have shown a massive growth in the field of automated medical image analysis, mainly because they can extract deep features from medical images. CNNs have been enabled to excel in tasks including disease detection, image segmentation, and feature extraction. Several researchers have effectively utilised CNNs for applications like tumour identification [41], skin lesion analysis [42], and colon cancer detection [43]. Models such as CheXNet [44, 45], utilised for the classification of 14 distinct chest conditions, have attained findings that are equivalent to or above those of medical professionals. Notably, the achievements of CNNs in the domain of DL for medical image analysis persists in encountering considerable obstacles. Data scarcity and the time-consuming annotations by experts continue to pose significant challenges to the development of strong models. Researchers have employed strategies such as TL [46, 47] and domain adaptation to

mitigate data dependency. In TL, models are pre-trained on extensive datasets, such as ImageNet, are refined for particular medical applications, facilitating efficient learning despite constrained data availability. Hybrid models integrating CNNs with conventional ML techniques or RNNs have been utilised to improve diagnostic precision.

In addition to CNNs, many DL approaches have been utilised in biomedical imaging. TL methodologies have been employed to detect brain tumours by MRI segmentation [46] and to diagnose pulmonary and cerebral disorders [47]. Supervised learning methods, such as SVM and RF, have been utilised alongside DL models to improve classification precision for intricate tasks, including liver disease prediction and patient monitoring [48, 49]. These developments demonstrate how flexible and relevant DL is for medical imaging. Though the discipline of automated medical image analysis is still under research, its advancement is rather obvious. More advanced designs and technologies present excellent opportunities for progress in customised healthcare, automated diagnostics, and treatment planning. Methods like end-to-end DL, hybrid models, and parallel CNNs persist in expanding the limits of possibility, enabling researchers to address progressively intricate challenges in the medical domain.

2.3 Segmentation techniques for skin lesion analysis

In this section, the overview of image segmentation technique is presented. Moreover, their applications in medical imaging analysis and how other advanced DL architectures, such as U-Net and others, contribute to its effective segmentation for skin lesion detection. The focus of this section is on their applicability to biomedical imaging, particularly for distinguishing between malignant and benign lesions in dermoscopic images.

Image segmentation in image processing requires dividing an image or its pixels into distinct classes or segments according to their characteristics and attributes [50]. This method constitutes the preliminary phase of numerous applications related to image and visual processing, including medical science, video surveillance, augmented reality, and self-driven cars. In the last few years, various approaches for image segmentation that have been established; a few of them are region-growing [51], watershed [52], clustering [53], and graph cuts [54], among others. Jing et al. [55] provide a comprehensive analysis of segmentation and its many classifications. A multitude of DL-based algorithms have been devised and analysed for image segmentation, often demonstrating enhanced efficacy relative to traditional ML techniques.

The image segmentation problem is classified into two main categories. The first one is semantic segmentation; this involves segmenting all the pixels in an image so that all pixels

belonging to a specific class are represented as a single segment. The second approach to image segmentation is instance segmentation, which segments different objects of the same class. Recently, a third type of labelling, known as panoptic segmentation, has been proposed, which integrates both semantic and instance segmentation. In this method, all the pixels belonging to a single class are identified, and then the instances of objects belonging to a class are also identified.

Semantic segmentation aims to group different regions of an image into categories. When dealing with low-resolution images, this segmentation task can be considered a classification problem. Semantic segmentation draws outlines around the edges of target objects and labels them. Although CT and MRI image segmentation research encompasses a range of organs, such as the liver, prostate, and knee cartilage, a significant amount of effort has been dedicated to brain segmentation, particularly in the context of tumour segmentation. Moreover, this technique has diverse applications, a few of them are including road sign detection [56], colon crypt segmentation [57], land-use classification, and land surface classification [58]. Moreover, it finds extensive usage in the medical domain, such as brain and tumor detection [59], as well as in detecting and tracking medical devices during surgery [60]. Scene resolution is of great significance in advanced driver assistance systems or car driving areas and depends extensively on semantic image segmentation [61]. Thresholding, which is commonly used in gray surface images, is a simple method that has been implemented in various medical datasets. The application of semantic segmentation with domain adaptation mitigates the distribution disparity between training and test datasets, facilitating the model's effective deployment in real-world contexts [62]. Executing real-time semantic segmentation in biological imaging is a formidable challenge, since it necessitates attaining both high precision and rapid inference speed concurrently. However, when resource limitations exist, it becomes impossible to maintain accuracy while simultaneously boosting inference speed.

In medical image analysis, image segmentation is a fundamental procedure in which an image is divided into meaningful sections, each matching to distinct anatomical or diseased structures. In medical images, this method helps to localise and define organs, tumours, or lesions, thus enabling correct diagnosis and treatment planning. Segmentation is crucial in clinical environments to isolate ROI, including tumour detection in radiology or lesion detection in dermatology [63, 64]. Edge detection, region expanding, and thresholding [50, 51] are classic image processing methods used in segmentation historically. Especially in the face of noise or poor contrast [63], these conventional approaches sometimes find it difficult to accommodate the intricate fluctuations in medical images, including various intensities, textures, and shapes of anatomical components. Semantic segmentation has lately advanced towards DL-based techniques, which outperform conventional methods

across several medical uses, including tumour detection, organ segmentation, and more recently skin lesion analysis [65].

U-Net has emerged as an innovative model for biomedical image segmentation among multiple DL architectures. The U-Net architecture was especially developed for biomedical image segmentation and has demonstrated significant efficacy owing to its distinctive shape, which comprises a contracting path for context acquisition and an expanding path for accurate localisation [66]. Same technique that a CNN uses to extracts features through a series of convolutions and pooling operations, the contracting route of U-Net is quite similar to it. On the other hand, the expanding path comprises up-convolutions and concatenations, which restore the spatial resolution of the image. This enables the model to accurately determine and highlight lesion boundaries for the patient [66]. This feature renders U-Net especially advantageous for medical image analysis, where accurate border identification is essential [67]. Applications in chest CT scans, brain MRI for tumour segmentation, and retina imaging [66, 68, 69] have made it rather popular for segmenting organs and tumour detection. Managing the range of skin lesions, including their appearance, uneven boundaries, and variable pigment patterns, semantic segmentation has been indispensable [70, 71]. Suitable for treating the diversity in skin lesions, the skip connections in U-Net, which link the contracting and expanding channels, maintain spatial information routinely lost in deep networks [66].

This thesis aims to solve the major class imbalance seen in skin lesion datasets, especially in circumstances where malignant ones are significantly less common than benign ones. Advanced augmentation methods, including synthetic creation of malignant images, have been used to manage this problem. These augmenting techniques help to increase the learning capacity of the model by means of more representative and balanced input during training. This approach is crucial for increasing the classification accuracy of malignant lesions [72] as well as the robustness of DL models applied for the diagnosis and detection of skin disorders.

Although U-Net is covered in great detail in the literature for use in medical imaging, other architectures and segmentation methods have also made major contributions. For image segmentation in medical settings, including organ and lesion segmentation, FCNN [73] and Mask R-CNN [74] have shown interesting outcomes. Like U-Net, these models also use encoder-decoder architectures but provide different ways to improve feature extraction and localisation, which can be quite helpful in situations with complicated skin lesions.

Effective segmentation helps to examine lesion properties more precisely, so enhancing accuracy in the next classification challenges. DL models like ResNet and EfficientNet are also considered in the domain of segmented areas to categorise lesions as benign

or malignant. This procedure improves the system's diagnostic capacity by means of a more ordered input for categorisation, therefore strengthening the dependability of predictions and guiding doctors in making wise selections. U-Net and its variants have been increasingly used in recent years for skin lesion analysis, therefore demonstrating their versatility and durability. For example, the ISIC 2018 skin lesion analysis challenge highlighted how well U-Net and its variants define lesion boundaries for next classification tasks [70]. U-Net is especially appropriate for medical uses where knowledge of the structure of an anatomical region is crucial for clinical decision-making since it can combine contextual features while preserving fine-grained details at a pixel level [70].

Moreover, variants of U-Net as Attention U-Net have been suggested to improve the fundamental U-Net architecture by emphasising suitable features throughout the segmentation process [75]. These improvements especially help in skin lesion analysis, where the variation in lesion size, colour, and texture calls for a more sophisticated method to guarantee that important characteristics are underlined throughout training. Particularly in difficult circumstances where the lesion boundaries are less clear [75], attention methods enable the network to concentrate just on the most informative areas of the image, hence enhancing segmentation accuracy. One cannot stress the need of proper segmentation in the study of skin lesions. Better extraction of features, including colour, texture, and shape, which are essential for differentiating benign from malignant lesions, is made possible by proper definition of lesion boundaries [24, 71]. Where DL models such as ResNet-50 or EfficientNet-V2 can be used to raise diagnostic accuracy [3], [2], this segmentation procedure forms the basis for later classification and analysis.

If traditional DL algorithms are combined with modern DL for image segmentation knowledge, it offers a strong answer for the problems in medical imaging. Although U-Net is the most often used architecture for biomedical segmentation, looking at other models and implementing their special advantages can assist to boost segmentation quality even more. Especially in early diagnosis and management of skin cancer, this work transforms and adapts these approaches to skin lesion detection, therefore supporting the ongoing effort to apply AI for improved healthcare outcomes [65, 70]. This combined strategy of segmentation and classification improves the interpretability of the model's predictions by means of which healthcare practitioners may have a comprehensive knowledge of both the visual and diagnostic aspects of skin lesions. While segmentation methods define important characteristics in medical images, classification models such as ResNet and EfficientNet advance these inputs to precisely detect disorders including melanoma.urther to accurately identify diseases such as melanoma.

2.4 ResNet and EfficientNet in biomedical imaging

In biomedical imaging, correct classification determines precise diagnosis of numerous diseases, including benign or malignant skin lesions. Particularly ResNet and EfficientNet, DL architectures have proved rather important in enhancing classification accuracy in the medical domain, such as melanoma diagnosis from dermatoscopic images. This emphasises the importance of consistent detection to support early diagnosis; this thesis uses these strong models to address the complexity of melanoma detection. The improved architectures of ResNet and EfficientNet, which include efficient scaling and deep residual learning. Their natural ability to manage high variability in medical imaging datasets, which comes from skin lesion analysis, makes them appropriate for overcoming obstacles including class imbalance, varied pigment, and uneven lesion margins. ResNet and EfficientNet are thoroughly reviewed in this section in order to emphasise their particular architectures, variants, and effects on biomedical imaging applications, thus laying a strong basis for their choice and application in skin lesion detection inside this thesis.

ResNet [10] addresses the issue of vanishing gradients by using residual learning with skip connections [3]. This breakthrough enables the network to learn identity mappings, hence facilitating the training of deep networks without the performance deterioration commonly associated with greater depth. Medical imaging has heavily used versions of ResNet including ResNet-50 and ResNet-101, having 50 and 101 layers, respectively, to capture intricate patterns vital for accurate categorisation. With a stunning F1 score of 94% [76], ResNet-152 displayed great efficiency in identifying lung diseases from chest X-ray images. This design proved useful in organising complex medical datasets [77] since it clearly distinguished diseases, including COVID-19, pneumonia, and lung opacity.

In addition, ResNet-50 has been utilised in the field of pulmonary diagnostics, but it has also been utilised in the diagnosis of brain tumours through the utilisation of MRI images, exhibiting remarkable accuracy and resilience [78]. ResNet can simply identify even the most subtle patterns in medical images [78], it is the perfect choice for applications needing great accuracy, such as separating benign from malignant tumours or in brain scans. One of its remarkable features is learning and distinguishing between small elements in medical imaging. This architecture has also been helpful in dermatology for skin lesion analysis, helping to early melanoma detection by means of skin lesion classification [79]. Timely detection and treatment increase the survival rates in melanoma patients to distinguish benign from malignant lesions [79].

Another research investigated using ResNet-50 to classify liver diseases depending on CT images. This work shows that the model can classify benign and malignant liver treatment in multi-class classification [80]. The results led one to conclude that ResNet

is rather successful in the process of extracting features discriminating between different liver disorders, thereby improving clinical decision-making [80]. Moreover, ResNet-101 was applied in a study aiming on breast cancer detection. This study revealed that, in mammography images, it offered more accuracy in separating benign from malignant tumours than any conventional machine learning method [81].

There is another architecture, EfficientNet [2] that provides an alternative by prioritising robust speed without compromising the accuracy of the model [82]. EfficientNet employs compound scaling to ensure that the network's depth, width, and resolution are proportionately balanced. This produces models that are very effective yet remarkably efficient. The EfficientNet family comprises versions B0 through B7, which may be scaled to accommodate varying resource configurations and data types [82].

EfficientNet's adaptability has also been much embraced in biological imaging. Its use in the diagnosis of breast cancer is one prominent example; an EfficientNet-based model classified mammography images to get an AUC of 0.9344. This great degree of accuracy emphasises how well the model distinguishes between benign and malignant cases in breast tissue, therefore stressing EfficientNet's promise in diagnostics requiring accuracy [83]. EfficientNet-B7, which is one of the larger versions in the EfficientNets family, has also been applied to classify blood cells, thereby advancing the automation of haematological analyses [84]. The model's ability to handle high-resolution images with fine details contributes to improved classification accuracy, which is essential for identifying various blood conditions [84].

Moreover, it is interesting in cell identification; EfficientNet-B7 finds use in cytological slide aberrant cell pattern recognition. The study showed that EfficientNet-B7 exceeded standard CNNs, therefore proving its scalability and efficiency in processing high-resolution medical images [85]. Beyond medical situations, EfficientNet-B7 has demonstrated its adaptability in numerous image classification tasks, including real-time food classification, proving its robustness in handling many kinds of image datasets with great variability [86]. This flexibility supports the confirmation of EfficientNet's universal architecture suitable for the demanding and diverse domain of medical imaging.

When one compares these two architectures, many advantages and differences emerge. ResNet's deep layers define its architecture, which is excellent in capturing complex features [87]. This is helpful for such tasks that requires discrimination between nearly similar structures, such as different types of tumours or skin lesions. This profundity frequently entails heightened computational requirements, which may pose a constraint in environments with limited resources [88]. EfficientNet is quite suitable for situations when processing efficiency is the top priority without losing speed efficiency, as it uses a balanced approach by changing depth, breadth, and resolution via compound scaling [2].

Considering their differences, both ResNet and EfficientNet designs have been effectively applied in many different biological imaging applications, from lung disease classification to brain tumor detection, and from breast cancer screening to blood cell analysis [78, 81, 89]. ResNet's deep feature extraction and EfficientNet's efficient scaling offers a complementary approach to medical diagnostics. Strong training approaches and domain-specific changes paired with these designs significantly aid to meet the aim of raising diagnosis accuracy and efficiency in healthcare.

ResNet and EfficientNet taken together have made major progress possible in biological imaging. Their distinct designs—one stressing depth to capture minute details and the other optimising efficiency through compound scaling—then help to solve challenging medical challenges. Medical imaging datasets keep growing in size and complexity, therefore these DL structures are strong candidates in building accurate, scalable, and efficient diagnostic tools [81]. Through more exact and fast diagnosis, this will finally help to raise patient outcomes.

To address the combined challenges of data scarcity and class imbalance, TL offers a practical approach by leveraging pre-trained models and adapting them for specific biomedical tasks.

2.5 Transfer Learning in biomedical image classification

For developing an efficient method in biomedical image classification TL allows pre-trained models to learn specific features with minimum training data. Using the knowledge gained from large datasets such as ImageNet, this technique effectively classifies complex medical images, including skin lesions, hence reducing the requirement for major domain-specific annotations.

Emphasising its use in overcoming the difficulties related to restricted data availability in medical domains, the work by Kim et al. [90] addresses the application of TL in medical image classification. Many times lacking sufficient size, medical imaging datasets make training DNNs from scratch unworkable due to overfitting and excessive processing demands. Kim et al. [90] went over 121 research papers using TL in different medical image classification tasks. His work discussed the DL methods that are using pre-trained models, including ResNet, VGG, and Inception, for medical classification tasks. This research shows that feature extraction or fine-tuning approaches can greatly increase the model's accuracy. Using TL with pre-trained models, including ResNet-50, EfficientNet-B7, and EfficientNet-V2, to extract relevant features from the dataset and saves training time and computing expenses; I similarly address the difficulty of melanoma classification

in this thesis. Furthermore, shown by Rahaman et al. [91] and Shin et al. [92] were how feature extraction and fine-tuning might be used to efficiently adapt TL for various medical imaging tasks. Using TL helps both their work and mine to leverage pre-existing knowledge, therefore addressing data scarcity problems and enabling better performance in biomedical image categorisation.

Analysing the use of TL in biomedical image classification, Gunturu et al. [93] shows how well pre-trained neural networks can solve the typical problem of little labelled data in medical datasets. Mostly due to the lack of labelled medical images, training DNNs from scratch usually presents difficulties like overfitting and high computing needs. To overcome these issues, TL employs pre-trained models—such as CNNs trained on vast datasets like ImageNet—which helps to transfer the learnt characteristics for application in medical imaging tasks. This speeds up the learning process and raises model accuracy even with fewer, domain-specific datasets. Emphasising its advantages for various medical classification tasks, the authors examined many papers using TL techniques, including feature extraction and fine-tuning with models including ResNet, VGG, and Inception [93].

Peng et al. [94] explores the boundaries of typical TL in medical image classification especially with regard to which pre-trained DCNN features to reuse. The authors provide Truncated TL (TTL), a new method that removes the unnecessary high-level layers of a pre-trained model, so focussing on fine-tuning the lower levels most relevant for medical applications, thus boosting both model performance and efficiency. Medical situations, where the key characteristics are generally lower to mid-level, including textures and aberrant patterns, call for this method. Because TTL reduces computational complexity and accelerates inference by fine-tuning just three basic layers, it is particularly helpful for managing complex 3D medical data including CT and MRI scans. This approach directly relates to this thesis, in which I utilise TL with pre-trained models including ResNet-50, EfficientNet-B7, and EfficientNet-V2 for melanoma classification leveraging feature reuse to overcome the challenge of little annotated data. Likewise, TTL demonstrates how effectively changing pre-trained models can solve data shortage while preserving high performance. Other selective fine-tuning techniques are also shown by related work by Raghu et al. [95] and Tajbakhsh et al. [96], stressing the need for other ways to improve adaptability for certain medical tasks.

Particularly because of domain shift when moving information from general datasets like ImageNet to specialised medical datasets, which can differ greatly in visual properties, TL finds various difficulties in biomedical image categorisation. Furthermore, the lack of annotated medical images limits the applicability of models to new biomedical activities and complicates their whole fine-tuning.

2.6 Skin lesion classification on imbalanced dataset

Munuswam et al. [97] presented an innovative approach for skin lesion classification. His work shows the combination of ensemble DL models, including ResNeXt101, SeResNeXt101, DenseNet 201, and Xception. Their study mostly sought to solve common problems, including limited datasets, imbalanced classes, and noise interference that arise in the analysis of skin lesions. They improved the quality and variety of the dataset using image inpainting and data augmentation. An average ensemble model and a weighted ensemble model were developed, the second stage of his work presenting a macro-average Receiver Operating Characteristic (ROC) Area Under Curve (AUC) score of 97, therefore demonstrating its higher accuracy over individual models. Emphasising their possibility to improve skin lesion diagnosis, the results showed the advantages of ensemble approaches in raising classification metrics. Furthermore, [97] successfully tackles the problem of class imbalance in skin lesion datasets, which is a major concern that could compromise model performance, particularly in medical imaging where some lesion types are under-represented. To offset this class difference and raise the general resilience of their model, they integrated five DL models—ResNeXt101, SeResNeXt101, ResNet152V2, DenseNet 201, and GoogLeNet—using an ensemble technique. This thesis uses three models for skin lesion classification, so our aim is similar in correcting the class imbalance issue but different in the complexity of the ensemble method that I applied. Their application of cost-sensitive learning and data augmentation reflects in my method of creating synthetic melanoma samples to increase data diversity. This resemblance emphasises the need of using several approaches to control class imbalance in order to guarantee consistent diagnosis results in the classification of skin lesions, thereby increasing sensitivity and accuracy.

Kassem et al. [98] thoroughly investigate most current advancements in skin lesion classification. The research underlines how much DL models improve diagnosis accuracy over conventional approaches. This study points out important difficulties, including racial bias, sparse datasets, and the difficulties in segmenting different skin lesions patterns. DCNNs have shown promising results in tackling these problems by efficiently extracting significant features for automated classification for melanoma identification. Furthermore, Kassem et al. [98] present ensemble learning and attention mechanisms, which greatly improve diagnostic accuracy. Underlines the transforming possibilities of DL in CAD for dermatology, therefore supporting doctors in precise and consistent skin lesion classification [98].

Previous studies, including DL applications for skin lesion classification, such as skin lesion classification on the ISIC 2016 dataset [99] and skin lesion classification by using DL models [100], show how DL models can even surpass dermatologists in diagnostic

accuracy under some circumstances. Furthermore, TL was used to solve class imbalance, a recurrent problem in skin lesion datasets, thus enhancing model robustness [101], [102]. These investigations show how changes in neural network architecture and training techniques have progressively improved the accuracy of automated skin lesion diagnosis, therefore reflecting the evolution of skin lesion categorisation techniques.

Class imbalance is one of the most recurring difficulties in medical image classification, so Nguyen et al. [103] present its solution in their study. Malignant lesions are sometimes under-represented in skin lesion datasets relative to benign lesions, which can produce a biased classification model favouring the majority class. This bias seriously impairs the accurate identification of rare but important cases, including melanoma. Nguyen et al. [103] suggest a combination of pre-trained DL models, including InceptionResNetV2 and MobileNetV3, with soft attention mechanisms and an inventive weighted loss function that increases the contribution of the minority class during training to address this issue. Their approach shows that even the under-represented groups get enough attention, therefore improving the performance criteria of the model, including recall and F1-score. Like their research work, this thesis addresses the class imbalance that is inherent in medical datasets, which can cause the model to underperform in identifying malignant lesions. Using weighted loss functions and emphasising balanced performance across classes helps me to reduce data-driven bias and produces accurate diagnosis predictions, similar to what Nguyen et al. presented.

Using the Synthetic Minority Oversampling Technique (SMOTE), which creates synthetic data points for the minority class to artificially balance the dataset, Abayomi-alli et al. [104] presented this technique in their research. This method clearly improved recall and F1-score, therefore suggesting its effectiveness in producing a more balanced dataset and correcting the model's bias towards benign cases. To further the representation of under-represented classes, Soumyyak et al. [105] applied soft attention and data augmentation methods employing models such as DenseNet 201 and ResNet50. Although their augmentation technique enhanced accuracy, it did not sufficiently solve the imbalance in terms of recall for minority classes, implying that augmentation by itself could not be sufficient in some situations. Moreover, Garg et al. [106] combined histogram equalisation for preprocessing with a TL technique, including CNN-based models, including ResNet50 and VGG16. Though a comprehensive accuracy was obtained, the imbalance resulted in low recall scores for important classes, which emphasises the need for more advanced methods to manage such variations. These connected efforts emphasise that in the field of skin lesion classification, establishing accurate models depends critically on correcting class imbalance. Inspired by these approaches, the application of this thesis uses weighted loss functions combined with pre-trained feature extractors to balance learning, hence

enhancing the sensitivity of the model towards malignant lesions and guaranteeing a more fair representation among several skin lesion types.

In ML, especially in important fields like disease diagnosis and fraud detection, Johnson et al. [107] similarly handle the difficulties of class imbalance. Class imbalance is the result of one class greatly outnumbering others in terms of representation, therefore promoting biassed learning favouring the majority class. To solve this, the paper explores numerous ways, including algorithm-level methods, including cost-sensitive learning and modified loss functions, as well as data-level strategies. Moreover, it include Random Over-Sampling (ROS), Random Under-Sampling (RUS), and SMOTE. The paper shows how well these approaches lower bias connected to imbalance and increase the reliability of DL models.

The research by Johnson et al. [107] is especially relevant to this thesis since the problem of class imbalance affects the skin lesion categorisation issue under consideration as well. Using three TL-based models in the thesis helps to classify melanoma based on notable differences between benign and malignant cases. This thesis uses methods including data augmentation and class weighting in line with the methodology under investigation in this work to reduce the negative consequences of imbalance. Furthermore, making sure minority classes have a fair assessment of the performance of the model corresponds with the emphasis on assessing criteria such as the F1 score and the ROC, AUC. The work of Johnson et al. offers a better understanding of the application of several strategies for restoring balance. These papers thus help to reinforce the approaches used in this thesis to improve model performance in the setting of imbalanced medical image classification.

Chapter 3

Methodology

3.1 Introduction to methodology

This chapter offers a thorough summary of the techniques used to address the skin lesion Classification proble, having a focus towards prediction of whether a lesion is benign or malignant. This thesis mostly focusses on using DL methods and integrating patient metadata to improve prediction accuracy. This thesis implementation establishes a ground work on which future noval ML techniques can be integrated and implemented to make a system, that supports clinical decision-making by creating a reliable and efficient model. Capable of processing both image data and contextual patient information, ultimately improving the diagnosis process.

Starting with data collecting and preprocessing, the approach comprises several main stages: personalized DL-model construction, training, and evaluation follows. Integration of metadata during training such as age, sex, anatomical site and image features from the image dataset. This aims to create a more comprehensive model that is able to account for other factors influencing diagnosis outcomes. Emphasising how the propsed methodology is implemented to overcome problem statement, this chapter addresses the design decisions, implementation details, and justification for each implementation step.

Comparisons between relevant studies addressing similar issues such as class imbalance in medical imaging and the use of TL with limited data availability will be highlighted throughout this chapter. Furthermore, the outcomes of three implemented DL methods help to eleborate the application of this concept. These comparisons will serve to highlight both the special contributions of this thesis and the parallel contrasts with earlier studies.

Problem statement

In dermatology, classification of skin lesions into benign and malignant categories still presents a major difficulty. Since early and precise diagnosis is essential for accurate treatment and improvement in patient outcomes. This thesis intends to integrate metadata during the training and test training for predictions, and build a model that classify skin lesions using multiple images of the same patient. By including contextual data beyond the image feature themselves, this combined approach seeks to give every patient a more comprehensive predictions.

Along with the efficient utilization of patient metadata records for enhanced predictions, this thesis presents challenges like classimbalanc, where the abundant samples are benign and just a small fraction are malignant. Traditional ML models struggle with these challenges that requires more specialized approach combining visual data and metadata information for an accurate solution.

DL-based approach to the problem

This thesis uses a multi-step approach focusing on extensive development, and the integration of metadata to improve prediction capabilities in order to efficiently handle challenges in skin lesion classification. The following sub-components clearly introduce the several techniques under discussion in this thesis.

Dataset collection and preprocessing

This study uses a dataset including multiple images for each patient together with associated metadata. Along with resolving missing metadata, preprocessing included standardizing of image data and cleaning of metadata. Data augmentation was also used to solve class imbalance, especially to increase the minority class to more accurately represent malignant cases.

Utilization of Transfer Learning

TL was used with pre-trained DL models like EfficientNet-V2, EfficientNet-B7, and ResNet-50. Initially built on large-scale ImageNet datasets, these models were refined to fit the specific features of skin lesion imagery. This allows feasibility to run with a quite limited number of annotated images in the dataset by use of efficient feature extraction enabled by these pre-trained models.

Integration of metadata information

Along with images, the model included anatomical information, age, sex, as metadata. Ground truth has also been trained for model training; whereas, for test cases a semi-supervised approach is used without consideration for ground truth for predictions. This integration sought to give the classification model richer variety of input features, which

allowed it to consider significant anatomical aspects that influence diagnosis. Metadata was processed and integrated into the model through additional layers, enabling the model to learn correlations between metadata and visual features.

Strategy for training the model

The model training was carried out using a weighted loss function like focal loss. This loss function is the goto choice in medical imaging in order to solve the class imbalance. Class weighting gave minor class greater weight, therefore assuring that these lower number of represented malignant receive enough attention during training. PyTorch was used to train a sophisticated DL model by efficiently managing the processing needed using multiple GPUs.

Effectiveness of the approach

In terms of class imbalance and the incorporation of metadata, the approach proposed in this thesis basically solves the primary challenges of skin lesion classification. Although letting the model achieve great accuracy without a huge annotated dataset is possible, the use of pre-trained models provides a comprehensive framework for feature extraction. By fine-tuning for optimization of these models for the melanoma/benign skin lesions, the proposed approach learns the relevant features and patterns with efficiency.

Moreover, inclusion of metadata considerably improved classification accuracy since metadata are known to be correlated with differnt types of skin lesions. Moreover, the use of focal loss insured that, despite their minority occurrence, malignant samples were appropriately represented during the learning process, therefore addressing the imbalance issue and enabling more balanced performance measurements.

Goal and vision of the thesis

Using image data and patient metadata, this thesis aims to build a tool capable to predict whether the chosen skin lesions are malignant or benign. The objective is not to exceed current state-of-the-art methods but rather to develop a dependable and interpretable model that categorize skin lesions with great accuracy, and can be improved further through next studies. This effort intends to improve diagnostic accuracy and finally help to build more comprehensive AI-assisted diagnostic tools by including metadata with image features.

The long-term goal of this work is to lay a foundation on which the future research may develop. More complex metadata integration methods, cutting-edge DNNS architectures, or multiple types of medical data may be useful to improve the approach presented here thereby strengthening the model. This thesis offers a foundation for next studies and

improvements to skin lesion classification, and for broader understanding of using DL methods on challenging medical imaging data.

PyTorch as implementation framwework This thesis is implemented within the PyTorch framework, selected for its ability to sufficiently handle the requirements and difficulties of this research thesis. Developed by Facebook's AI research team, PyTorch is an open-source DL tool with a strong following within the research community because of its adaptability and easy-to-use layout. Given the nature of this thesis, which uses TL approaches to identify skin lesions as benign or malignant, PyTorch was a natural choice that fits very well with the research objectives.

PyTorch mostly comes with its dynamic computational graph, which allows real-time optimization and debugging during the development of models. In the framework of our research, where iterative experimentation with pre-trained models is applied, this feature proven particularly useful. PyTorch's design allows greater flexibility in changing model components, fine-tuning hyperparameters, and network architectural adaptation to better represent the complexity inherent in medical imaging data than frameworks depending on static computational graphs. This amount of adaptability was quite important in improving the model and handling the particular difficulties in the presented dataset.

Furthermore, PyTorch's modular design fits perfectly with the incremental approach utilized during this thesis. PyTorch gave a simple platform for reusing large pre-trained models and modifying them as needed given the demand for frequent modifications to the training pipeline and the integration of TL methods. Commonly in the biomedical field, TL helps to overcome the scarcity of labeled medical data. PyTorch's torchvision collection provides the pre-trained models. This greatly improves the learning process considering the limits of the dataset by enabling the feature extraction efficiently without the computational burden of training from scratch.

PyTorch's has outstanding support for GPU acceleration—which enabled effective training and experimentation with large models using multiple GPUs. Given the processing requirements related to training DL models on a dataset of around 41,000 images, this capability was required. Faster training times made it achievable by PyTorch's effective handling of GPU calculations, hence it support the iterative nature of this thesis for comprehensive configuration for the training and test loops.

For this thesis the most beneficial were PyTorch's comprehensive documentations, active community, and accessible libraries for data processing and augmentation. PyTorch's modular data management lets the research add information inputs such patient sex, age, and anatomical site into the model, as an effective mix corresponding with image features.

This integration offers the deeper contextual knowledge for the model predictions as well as a comprehensive evaluation of the patient's condition.

PyTorch was chosen primarily for its versatility, simplicity of use, and great GPU support, all of which were absolutely necessary for the effective use of the proposed methodologies. PyTorch's adaptability assisted both with creating models and optimization as well as addressing the difficulties in skin lesion classification. The framework of PyTorch has proved to be really helpful in laying a strong basis for this thesis. Since it offers the required degree of independence and flexibility required for the implementation of this thesis.

3.2 Data collection and preprocessing

3.2.1 Data source description

This thesis implements the ISIC 2020 challenge dataset. Though ISIC dataset has several versions but ISIC 2020 is specifically designed to advance the field of melanoma detection and it contains 33,126 DICOM images, Each image comes with embeded metadata that includes key attributes like `image_name`, `patient_id`, `sex`, `age_approx`, `anayom_site_challenge`, `diagnosis`, `benign_malignant` and `target`. By integrating image data with the associated metadata, this dataset provides a rich resource for the development of advance ML model that aims to identify malignant skin lesions.

Relevant in examining the risk factors related with melanoma, the chosen metadata fields were `sex`, `age_approx`, `anatom_site_general_challenge`, and `benign_malignant`. For example, sex and age have been found to be significant factors to the frequency of skin cancer; some forms are more common in particular demographic groups. Furthermore offering important contextual information to affect diagnosis is the anatomical site of the lesion. Extraction and use of these domains seeks to improve the predictive capacity of the model outside of the image data alone.

Widely used in medical imaging research, the ISIC 2020 dataset provides a benchmark for many studies on skin lesion classification. Using the same dataset, Rahman et al. [108] addressed melanoma detection confronting similar problems of class imbalance, data variability, and the difficulty of robust generalizing. Their preparation activities included careful handling of metadata and using augmentation strategies to address class discrepancies, therefore offering insightful information for the thesis approach taken here.

3.2.2 Data preprocessing steps

Data preparation is an important step that impacts how well the ML model learns and how well it can generalize on that dataset. In this thesis, preprocessing included a number of important steps, such as converting image formats and dealing with class imbalance. All of these steps were meant to make sure that the training data was of the highest standard. For each preprocessing step, each step is elabroated in detail about why it was done, what its possible pros and cons are, and how it compares to previous researches that has been carried out in the literature.

3.2.2.1 Conversion of DICOM to JPEG

Initially, each image in the dataset came in the DICOM format—used in medical imaging as it keeps both image data and valuable metadata including patient demographics and scanning protocols. Although it frequently contains JPEG or PNG formats, DICOM files are not entirely compatible directly with most major DL frameworks. DL models specially those implemented using frameworks like PyTorch or TensorFlow, require standardized image formats and these frameworks provides libraries that easily loads, processed, and augmented these images. Few commonly used such as OpenCV, PIL and alumentations.

To facilitate the training process, All 33,126 DICOM images were transformed into JPEG format with python scripts to help with conversion. Careful execution of the conversion guarantees consistency among data processing phases and helps to keep the related metadata. Using JPEG format has a lot of advantages which are mostly related to its wide compatibility and efficiency in reading and writing operations, this greatly lowers the computational cost during training, which significantly reduces the computational overhead during training.

Reasoning for conversion: It would be easier to use all standard ML libraries that don't handle DICOM images effectively, hence the choice was made to transform DICOM to JPEG. Sun et al. [109] mentioned that changing DICOM to JPEG is often needed in DL applications to make it work easier for complex medical formats. This makes it easier to create models while keeping the standarized image quality needed for medical diagnosis. Even though JPEG is a lossy file, the metadata extraction makes sure that important data isn't lost. This means that the images can still be used for diagnostic purposes.

Pros of DICOM to JPEG conversion:

- Most DL frameworks can easily work with JPEG images, which makes the process of building models easy.
- All important data was kept in an ordered CSV format after being extracted from DICOM files, which meant that metadata could still be used as standarized data.

Cons of DICOM to JPEG conversion:

- Because JPEG is a lossy compression method, the quality of the images may get slightly degraded, which could affect the model performance. But because DL models don't work well with DICOM, this was considered as a fair trade-off.

This step was based on similar methods found in other research. For example, Xu et al. [110] changed DICOM images from a dataset to JPEG for training CNNs. This shows that the trade-off in terms of access and processing speed is worth it for most medical imaging tasks, even if the quality is compromised.

3.2.2.2 Metadata cleaning

The metadata features are discussed in great detail in this section. Every image has a unique identity provided by its **image_name**, which ensures appropriate traceability throughout several phases of data processing, model training, and evaluation. Applied preprocessing, training, or model validation all depend on a consistent mapping between the metadata and corresponding image data, which is essential when applying preprocessing steps, training, or validating the model. Concurrently, **patient_id** is another crucial field that distinguishes every patient and is instrumental in handling the multiple images of one patient. Ensuring that all images of a given patient are allocated to the same split, either in training or testing, this helps to prevent any data leakage, hence preventing any overly optimistic performance estimate. Such a splitting approach greatly increases generalizability of results to new patients and model robustness.

An essential factor in the study of melanoma risk is the **sex** attribute, which denotes the gender of the patient. Several studies have shown gender specific differences in skin cancer frequency; men have often been shown to be more vulnerable than women. Men, for example, have a higher lifetime risk of melanoma than women, according to Michal K. et al. [111], potentially because of variations in behavior including sun exposure and reduced rates of sunscreen use. Their results also showed that whilst women are more likely to get melanoma on the lower limbs [111], males often develop melanoma on sun-exposed areas like the back and shoulders. Since age is a well known risk factor for melanoma, the **age_approx** attribute is also rather crucial. By including this data, the model

can take into account gender based risk factors that may influence the probability of a lesion being malignan. Global melanoma incidence and death were assessed methodically by Huang et al. [112] underlining the notable rise in melanoma incidence trends among those 50 years of age and above. This result emphasizes the need of including age as a major determinant of melanoma risk evaluations. Including age in the model helps it to establish correlations between various age groups and skin cancer hazards, hence age attribute aids in making more informed predictions.

The **anatom_site_general_challenge** is an attribute which offers more details on the anatomical location of the lesion, it indicated the location perhaps is on the head/neck, torso, or extremities etc. This is important since the anatomical location of melanoma has different prevalence rates depending on their anatomical location. Including this metadata attribute helps the model to make a more accurate prediction by using context about lesion locations. As different body areas vary in sun exposure and sensitivity, anatomical location greatly affects lesion characteristics, Green et al. [113]. Including this data allows the model make more accurate predictions by using contextual lesion location. By using this information, the model is able to identify anatomical patterns that are associated with malignancy risk.

The field of **diagnosis** offers insights on the type of lesion, whether it is a nevus or if it stays unclassified, that is unknown. This attribute provides useful ground truth for the validation of model performance even if it might not be explicitly used for training the model in this thesis. This diagnosis information is compared with model predictions during the evaluation phase to ascertain the model's accuracy, providing a baseline for validating the model outcomes against expert medical assessments.

The most important focus in this classification problem is the attribute **benign_malignant**, which differentiates benign from malignant lesions. The binary classification task requires DL model to distinguish malignant from benign lesions. However, one of the main challenges of this column is the class imbalance since the dataset consists of many more benign cases than malignant ones. Such class imbalance can provide a biassed model that is more likely to predict the majority class (benign). This attribute used as ground truth for the the training phase, while in test phase it is not used to reflect the real world application. Moreover, implementation of this thesis addresses class imbalance which includes techniques, such as augmentation and class weighting, to ensure the model could effectively learn from the minority malignant cases without being overwhelmed by the abundance of benign cases.

Crucially for training ML models, the **target** column comprises numerical labels, 0 for benign and 1 for malignant. Computation of loss and model optimization during training are based on these numerical depictions. Furthermore, the numerical labels

enables to apply weighted metrics that change the impact of the minority class, thereby allowing for a more meaningful evaluation of model performance in the context of this imbalanced dataset. **benign_malignant** and **target** serves the same purpose as binary classification.

Each one of these metadata fields is extremely important in creating a holistic image of the lesion and the patient. The prediction ability of the model is significantly enhanced by including image data together with metadata information. Together with the visual features taken from the images, this multimodal technique combines contextual information including the patient's sex, age, and lesion anatomical location. Particularly for rare but important cases, this kind of integration enables the model to generate more accurate predictions, hence potentially improving diagnosis accuracy and generalizability.

The initial inspection of the metadata revealed several missing values across different fields:

- `Sex`: 65 missing values
- `Age_approx`: 68 missing values
- `Target`: 1 missing value
- `Anatom_site_general_challenge`: 527 entries labeled as 'unknown'

The presence of missing or ambiguous values can have a detrimental effect on model training, as they introduce noise and can lead to inaccurate predictions. To address this, all rows containing missing or 'unknown' values in any of these fields were removed, resulting in a cleaned dataset with 32,530 entries.

Reasoning for dropping missing values:

The decision to remove rows with missing values was based on ensuring data integrity and consistency. Although there are other ways, like imputing missing data or setting a default category, these techniques can introduce artifacts or biases that might compromise the ability of the model to learn effectively. Zhang et. al. [114] claim that imputation of missing values might generate discrepancies, especially in datasets with great sensitivity, such as medical imaging data. Their research emphasizes that such imputation might cause model biases, particularly in cases involving clinical variables, therefore undermining the dependability of downstream analysis. Therefore, in this thesis, the decision to eliminate entries with missing data ensures that the training process is based on accurate information, a vital element in medical research since the quality of input data directly impacts diagnosis outcomes. Working with the ISIC dataset, Rahman et al. [108] likewise

faced similar problems, so their work chose to delete entries with incomplete metadata. This choice aligns with the widespread consensus in the literature that in the handling of sensitive medical data, maintaining data quality comes above maintaining a bigger dataset with potential inaccuracies.

Pros of dropping missing values:

- Ensures that the training data is consistent, which is critical for effective model learning.
- Avoids introducing artificial biases that could be harmful in a diagnostic context.

Cons of dropping missing values:

- Artifacts reduce the entire dataset size, hence potentially restricting the diversity of training samples. Given the enormous dataset in this thesis, nevertheless, this is seen as an acceptable trade-off.

3.2.2.3 Patient-wise data splitting

Given the nature of the data, which includes multiple images from the same patient, it was imperative to split the dataset in a way that ensured all images from the same patient were kept in the same subset either training or testing. This is carried out to prevent data leaking, in which similar images in both training and testing sets could artificially inflate performance metrics, hence misleading the impression of the generalizability of the model.

Implementation of patient-wise split: In the implementation of this thesis, the dataset is splited into 75% training and 25% testing ratios to ensure that the images stayed together in split. The primary identifier was patient IDs. This strategy is common in medical imaging research, as seen by Rishu Garg et al. [115], who used a similar methodology to maintain the integrity of patient-specific data during model training and testing. During training and testing, all of a patient's images were stored together to avoid the model "cheating" by recognizing similar lesions.

Reasoning for patient-wise splitting: Splitting by patient ID was necessary to avoid overfitting due to data leakage, where the model might acquire patient-specific features instead of generalize patterns, splitting by patient ID was required. Kalra et al. [116] argue that patient-wise data splitting is crucial in medical imaging to prevent manipulating evaluation assessment because images from the same patient often have similar visual features.

Pros of patient-wise splitting:

- It prevents data leaking, ensuring that the model's performance metrics accurately reflect its ability to generalize to new patients.
- Prevents the model from unnecessarily overfitting as a result of recurrent patient data in both training and testing sets, allowing it to learn patient-specific features.

Cons of patient-wise splitting:

- When there are less patients, the data distribution between the training and test sets might not be precisely balanced. The robustness of the evaluation metrics may be impacted by this imbalance.

Alwakid et. al. [117] uses DL methods for automated diagnosis to adress the rising skin cancer, particularly melanoma. The suggested approach uses Enhanced Super-Resolution Generative Adversarial Networks (ESRGAN) to enhance image quality, followed by segmentation of the ROI and subsequently classification using CNN and modified ResNet-50 model. These multi-steps were designed to solve data quality and imbalance issues; thereby enhancing melanoma detection.

In both Alwakid et al.'s work and this thesis, the issue of class imbalance takes a key focus. In their study, seven different kinds of skin lesions with diverse sample sizes made up the dataset, and then data augmentation methods were used to counter the imbalance dataset. This thesis also used data augmentation but extended it farther by using a two-pass training technique, which gave extra exposure to malignant images, therefore reducing the imbalance in a more comprehensive manner. Typically under-represented in such datasets, malignant samples could be better represented by the combination of online and offline augmentation techniques.

One major difference between this thesis approach and in the application of ESRGAN by Alwakid et al. [117], which was used to improve image quality before segmentation and classification. Still another crucial difference is the approach to dataset preparation and training. Alwakid et al. classified using a modified ResNet-50 after enhancing image using ESRGAN, getting a classification accuracy of 86%. Meanwhile, in this study using a patient-wise data splitting, offline augmentation for minority class oversampling, and applyinh focal loss to better manage class imbalance in contrast, this thesis aimed at preventing data leakage. These steps assured the model has sufficient representation of the malignant class, hence improving sensitivity and generalizability. Applying patient-wise splitting is quite crucial since it ensures that by avoiding similar images appearing

in both training and test sets, the evaluation metrics represent meaningful assessment and the generalizing abilitiies of the model.

Overall, comparing both the approaches have the same goal, that is to increase melanoma detection by applying DL techniques but they differ in the value that they provide in preprocessing techniques, data generation, and choice for suitable models. Their study showed how well ESRGAN might be used for image improvements. On the other hand, this thesis concentrated on metadata integration, patient-wise splitting, advanced augmentation techniques, and multiple architectural comparisons to offer a more complete solution to handle the difficulties related with melanoma detection. These distinctions highlight the unique contributions this thesis, especially with regard to addressing class imbalance and enhancing model generalization, thus laying the groundwork for future improvements in automated skin cancer diagnosis.

3.2.2.4 Addressing class imbalance

Upon analyzing the target labels, a significant class imbalance was observed. Specifically, the number of benign images far outweighed the number of malignant images, with a training set distribution of:

- `Benign`: 23,861 images
- `Malignant`: 439 images

In medical applications especially where pathological cases are generally fewer than normal ones, class imbalance is a well-known problem in ML. The imbalance might result in biassed models that are too predisposed to forecast the majority class, therefore lowering sensitivity to the minority class (in this case, malignant lesions).

Reasoning for addressing class imbalance: Effectively learning from both benign and malignant cases depends on how well preprocessing is handling a class imbalance. Johnson et. al [118] underline that class imbalance can seriously affect model performance, especially in identifying minority class instances, thereby generating poor recall rates for rare condition. Online data augmentation during training and offline augmentation combined to create synthetic images for the malignant class helped to address this challenge.

Offline augmentation for malignant cases: Malignant images were artificially increased by the technique of data augmentation methods including rotations, random horizontal and vertical flips, and color jitter. Given the considerable difference between

the classes, these techniques by themselves were not enough to balance the dataset though. For every malignant image in the training set, 15 synthetic images were created; there are total 7,024 malignant images in the training set. Likewise, for the test set, 10 synthetic images per malignant image were created producing total 1,496 malignant images.

$$\text{Malignant images in Training set: } (439 \times 15) + 439 = 7{,}024$$

$$\text{Malignant images in Test set: } (136 \times 10) + 136 = 1{,}496$$

The necessity to ensure that the model had plenty of exposure to malignant images during training while preserving a realistic and unbaised evaluation scenario during testing. Due to this factor, the decision to apply different augmentation variables for training and testing was taken into consideration. To improve class balance Abayomi-alli et al. [104], created synthetic data for minority classes using SMOTE and other augmentation methods. In medical imaging especially, offline augmentation is quite helpful since it lets the model learn from a more diverse group of malignant cases, hence improving its ability to generalize to real-world situations.

Pros of offline augmentation:

- Helps to mitigate the class imbalance issue, allowing the model to learn features associated with the minority class more effectively.
- Increases the model's sensitivity to malignant lesions, which is critical for medical diagnosis.

Cons of offline augmentation:

- Synthetic data may not always perfectly represent real-world variations, and excessive use of augmentation can lead to overfitting to artificially created patterns.
- There is a trade-off between the number of augmented images and the quality of the learned features, requiring careful tuning to achieve the desired balance.

3.2.2.5 Two-pass training strategy

Although augmentation was applied, but even then the malignant class remained in minority. To further enhance the model's exposure to malignant images, two-pass training

method is introduced. With a two-pass training phases for DL, the training process is split into two sequential to one another phases. The second phase builds upon on the progress made by the end of first phase. Either on a preliminary subset of the data or under specific conditions, the model is trained on a first pass until it achieves a point of convergence or satisfies the pre-defined stopping criteria. The model loads its weights and parameters kept at the end of Pass 1 before starting Pass 2. This is to assist continued training instead of beginning from scratch again. Training is carried on in the second pass utilizing more refined data, therefore enhancing the model to boost generalization and adapts to new patterns and make use of the prior knowledge acquired in the first pass. Particularly beneficial when dealing with continually evolving datasets, this method increases the efficacy of training and helps to avoid overfitting.

The training dataset was split into two subsets:

- **Pass 1**: `Benign`: 12,527, `Malignant`: 7,024
- **Pass 2**: `Benign`: 11,334, `Malignant`: 7,024

Reasoning for two-pass training strategy: This method's primary benefit is that it allows the model to double the exposure to malignant cases while training on all the available benign examples. This effectively established a training settings whereby the malignant class was less under-represented. I ensured that a model had more balanced view of both classes by splitting the benign data into two, and training on the whole malignant sample twice. The idea of multiple training passes on minority class data has been investigated in the literature, including by Garg et al. (2021), who found that re-exposing models to minority class examples in multiple passes helped to improve sensitivity without greatly raising overfitting [106]. Better representation learning of minority class features made it possible by this approach, hence this decision proved to be essential for the detection of rare conditions including melanoma.

3.2.2.6 Evaluation strategy for test set

For the test set, a single pass training strategy was employed. Instead, a combination of class weighting, weighted metrics, and balanced evaluation strategies was used:

- **Class weighting**: Weights from the training set were transferred to the test set to handle class imbalance during evaluation.
- **Weighted metrics**: To ensure that the minority class was sufficiently represented in the overall evaluation, weighted metrics including precision, recall, F1 score, and ROC AUC were weighted to observe the impact in all these evaluations.

- **Balanced evaluation**: The evaluation measures were changed to ensure that the malignant class had a proportionately higher higher impact on the computed scores, therefore resolving the intrinsic bias of imbalanced datasets.

Reasoning for weighted metrics and balanced evaluation: ROC AUC and F1 score are particularly chosen as less sensitive to class imbalance than accuracy. In the event of extreme class imbalance, Nguyen et al. (2022) underlined that measures such as accuracy could be deceptive since the model could reach great accuracy by just predicting the majority class [119]. By using weighted F1 score and ROC AUC, it has abilities provide a more reasonable evaluation for the model's malignant lesion classification.

3.2.2.7 Test set augmentation

Typically only augmentations to the training set are applied to improve generalizability and to prevent overfitting. This helps to prevent overly optimistic accuracy of the model. Conversely in this specific case, augmentations were implemented on the test set. This choice was justified by the fact that the training set had many augmented images of malignant tumors; so, it was essential for the test set to also include some augmented cases to yield a more realistic assessment of the model. If augmentation are not added in test set, then model may not be able to generalize well as the model had seen augmented images in training set. In this case, adding augmentations to both the sets, model becomes more balanced.

Reasoning for test set augmentation: The justification for the improvement in the test set is Including augmented images in the test set also ensured that the model was assessed on data similar to that which it had been trained on earlier. In the realm of medical imaging specifically this consistency is quite important since the unpredictability of real-world images can significantly affect the performance of models. In order to preserve consistency between the training and testing environments, Nguyen et al. (2022) applied a similar approach which finally produced a more robust evaluation [119].

3.2.3 Focal loss for addressing class imbalance

The loss function used during training to address the class imbalance was focal loss. Focal loss, unlike conventional cross-entropy loss, gives hard-to-classify examples higher weights, therefore focussing the model's attention on the minority class. Given the significant difference between benign and malignant cases in this thesis, this was particularly useful.

Particularly in situations when the majority and minority classes differ significantly, focal loss has become a important technique in tackling class imbalance. In this thesis, a

notable class imbalance results from a disproportionately high number of benign lesions relative to malignant ones used for melanoma identification. Such an imbalance makes a challenge to develope a robust classifier, since it can cause the model to favor predictions towards the majority class (benign), therefore reducing its sensitivity to malignant cases. Given the great stakes of medical diagnosis, where false negatives, that is misclassifying benign malignant lesions as benign can have serious consequences. it is impreratively to apply techniques that enable the model to effectively learn to identify the minority class. For the implementation, focal loss offers such a mechanism and is hence very important.

Lin et al. [120], introduced focal loss that modifies the standard cross-entropy loss by adding a modulating factor that reduce the relative loss contribution from well-classified cases. This modification assures that the model focuses more on difficult to classify examples, in this case the malignant lesion example, which are both under-represented in the dataset and challenging. In conventional loss functions such as cross-entropy, well-classified examples are usually benign samples due to their abundance. This may dominate the optimization process, therefore leaving malignant samples insufficiently represented during training. For the minority class particularly, this can result in suboptimal performance. By dynamically changing the loss, focal loss solves this problem and enables the model focus more on the minority class during training, therefore assuring that it learns the subtle features which are differentiating malignant lesions.

Incorporating focal loss in this thesis aims to efficiently overcome the inherent class imbalance in the skin lesion dataset. Focal loss focuses more on those cases most especially the malignant lesions by varying the impact of every training example depending on their classification accuracy. This method improves the learning process of the model by avoiding it from being too biased towards the majority class, so encouraging a more balanced learning result. In medical imaging, where misclassification might have rather substantial cost, this is especially crucial. The area of emphasis here is to increase the sensitivity to the malignant situations so that it may precisely identify such lesions and make the model more valuable in practical real world application.

Recent research have underlined especially in medical datasets the value of focal loss in managing class imbalance. This can greatly improve the sensitivity of a model to be underrepresented but clinically important, according to work by Lin et al. [120], their results are well aligned with the goals of this thesis, which aims to create a DL model able to precisely detect malignant lesions in spite of dataset imbalance. For this reason, the utilization of focal loss in this thesis depends not only on theoretical understanding but also on empirical validation that has been seen in recent research. This adds to the effectiveness of this strategy in addressing high-stakes classification problems.

Moreover, focal loss provides flexibility through its tunable hyperparameters, alpha and gamma, allowing for the fine-tuning of the model's learning process. The parameter gamma serves to adjust the attention that the model places on instances that are difficult to identify. It perform this by assisting in the determination of the rate of reduction in the loss for samples that have been perfectly classified. Meanwhile, Alpha enables one to strike a balance between the values of multiple classes. Within the context of this thesis, the precision with which hyperparameters are tuned is essential to achieving the best possible performance from the model. In order to ensure that malignant instances were neither underrepresented nor ignored throughout the training period, the settings were manually adjusted to ensure that they were calibrated appropriately. Because of this adaptability, focal loss is suitable option for our study, which enabled us to develop a sophisticated algorithm that could properly address the unique characteristics of the skin lesion dataset.

Finally, the use of focal loss in this thesis has been quite instrumental in addressing the problems related to significant class imbalance. Giving malignant tumors a top priority as the minority class. The focal loss function ensures that the model stays sensitive to these important situations, which finally enhances diagnosis performance. This method improves the generalizing ability effectively and helps prevent the model from becoming biased towards the majority class. For medical uses, where precise identification of malignant cases is of great significance, such a result is absolutely essential. Therefore, the application of focal loss helps greatly to create a more balanced and dependable model, which increases its relevance in clinical practice and sets the basis for next research aiming at more advanced approaches or improved versions of loss functions catered to the particular needs of medical imaging tasks.

3.3 Model architecture

By the use of modern DL architectures such as ResNet-50, EfficientNet-B7, and EfficientNet-V2, this thesis presents a skin lesion classification methodology. These architectures were chosen mostly based on their proven effectiveness in computer vision applications, particularly in medical imaging, and their strong feature extraction capability.

ResNet-50 is a CNN that solves the problem of vanishing gradients with the use if residual connections, therefore enabling the effective training of DNNs [3]. It is a great candidate for employment with medical imaging since it can learn hierarchical features from images. Conversely, two more recently developed models that maximize performance by using a compound scaling approach are EfficientNet-B7 and EfficientNet-V2. By balancing depth, width, and resolution, these models provide better accuracy while

preserving computational economy as reported in Tan et. el. [2]. EfficientNet-B7 was chosen especially because of its fine-grained feature extraction ability, which are crucial for differentiating benign from malignant lesions depending on their minor differences. EfficientNet-V2 is suitable for iterative testing since it has capacity to greatly improve training speed and efficiency.

The EfficientNet-B7 design is the DL architecture that is most commonly utilized for applications such as image classification, object detection, and medical imaging. Specifically, it was designed to handle tasks of this nature. The first version of it was released in 2019, and the EfficientNet family's fundamental concept is compound scaling, which is a method that strikes a balance between the depth, width, and resolution of the network. Researchers at Google contributed to the development of this method. The EfficientNet-B7 model is the largest of this family's models, which also includes other large-scale models such as ResNet-152 and DenseNet. It provides contemporary performance at a relatively cheap cost to the computing engine.

EfficientNet-B7 is especially important in this thesis since of its remarkable potential for feature extraction; it helps to classify skin lesions as either benign or malignant. A thorough discussion of its architecture and the rationale behind choosing for this research is given in the section below.

3.3.1 Overview of EfficientNet architecture

EfficientNet-B7 in Figure 3.1, which also includes a major advancement called compound scaling, is built on the the concept of MobileNetV2 architecture. Compound scaling is a method which uses a given depth, width, or resolution idependently to balance all three dimensions of a model. Unlike usual scaling techniques, which uses fixed coefficient. This is in line to ensures that the model expands efficiently without needless increase of computing overhead.

Built utilizing Mobile Inverted Bottleneck Convolution (MBConv)layers and using numerous techniques, such depthwise separable convolutions, in the EfficientNet architectural family, MBConv breaks down the convolution operation into pointwise convolutions as well, hence lowering the computational overhead while preserving the model's ability to learn complex features. Swish activation functions produce higher precision and output. From B0, which acted as a baseline, the researchers scaled the model using the compound scaling approach all the way up to B7, the most potent and largest version. But further improvements with EfficientNet-V1 and EfficientNet-V2 helped to cut training time comparable to B7.

Here is EfficientNet-B7 Architecture,

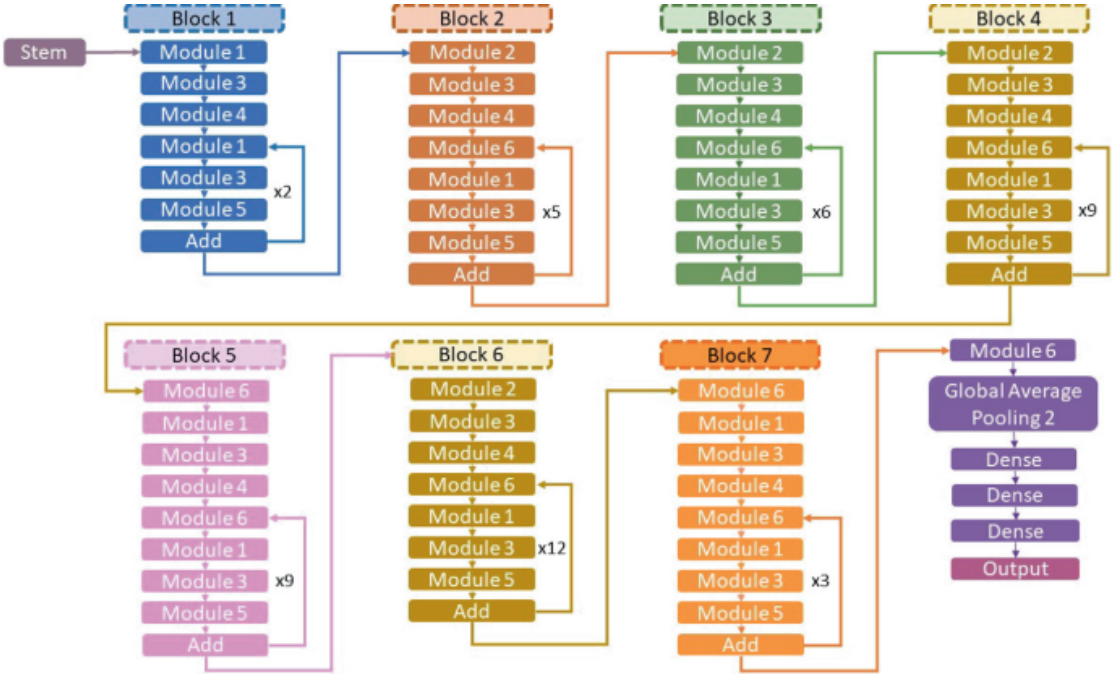


Figure 3.1: EfficientNet-B7 Architecture

3.3.1.1 Compound scaling

The compound scaling formula balances the three dimensions of the neural network:

$$\text{Model Size (S)} \propto \text{Depth (D)}^{\phi} \cdot \text{Width (W)}^{\phi} \cdot \text{Resolution (R)}^{\phi} \tag{3.1}$$

Where:

- ϕ: A user-specified scaling factor.
- D: Depth of the network (number of layers).
- W: Width of the network (number of channels).
- R: Input image resolution.

For EfficientNet-B7, $\phi = 2.0$, leading to a deeper network with a higher resolution and wider layers. This compound scaling ensures that the model is both powerful and computationally efficient.

3.3.1.2 MBConv layers

MBConv is the foundational building block of EfficientNet. It combines:

- **Depthwise separable convolutions**: By splitting the convolution operation into two smaller operations—depthwise convolution and pointwise convolution—these depthwise separable convolutions lower computing complexity.
- **Squeeze-and-excitation (SE) Blocks**: By recalibrating channel-wise feature maps, this improve the capacity of the model to focus on important features. This allows the network pay close attention to significant trends.

An MBConv block can be expressed as:

$$\text{Output} = \text{SE}(\text{Pointwise Conv}(\text{Depthwise Conv}(\text{Input}))) \tag{3.2}$$

These blocks are designed to reduce redundancy in the feature maps, enabling the model to focus on relevant patterns while minimizing computational cost.

3.3.1.3 Swish activation function

EfficientNet uses the swish activation function instead of the more common ReLU. The swish function is defined as:

$$\text{Swish}(x) = x \cdot \sigma(x) \tag{3.3}$$

Where $\sigma(x)$ is the sigmoid function. Swish introduces a smooth, non-monotonic behavior that improves gradient flow, enabling better optimization and ultimately boosting accuracy.

3.3.1.4 Resolution for the input image

It operate on images with a resolution of 600×600 pixels, the EfficientNet-B7 method provides the higher level of detail to input images. In comparison to smaller models, this resolution allows more information to be captured. Crucially for identifying subtle patterns in medical images, the model can capture fine-grained features. Among these patterns there are irregular shapes or uneven forms or textures. This improved resolution helps to capture these features.

3.3.1.5 Superior feature extraction

MBConv blocks and SE modules, EfficientNet-B7 can very effectively identify hierarchical features spanning from low-level edges to high-level textures. This is a crucial element in differentiating benign from malignant tumors as their small variations in form, color, and texture usually define them.

3.3.1.6 Balanced performance and efficiency

Though among the very largest models in the EfficientNet family, B7's revised architecture makes it in among the most computationally efficient models available. It is feasible for deployment on GPUs that are typically available in research environments since it offers state-of- the-art performance while requiring less FLOPs (Floating Point Operations) than comparable models such ResNet-152.

3.3.1.7 Robustness to limited data

Many medical datasets, including the one utilized for this thesis, provide a small quantity of labeled data. EfficientNet-B7 can overcome this limitation by applying TL since its pre-trained weights (from ImageNet) offer a strong beginning for feature extraction.

3.4 EfficientNet-B7: Approach for skin lesion classification

In this thesis, EfficientNet-B7 is essential because it provides the foundation of feature extraction in the skin lesion classification model. Its state-of- the-art performance, computational efficiency, and ability to generalize effectively in medical imaging has driven its choice. Architectural components such as TL techniques, metadata integration, and prediction mechanisms implemented for robust lesion classification are discussed in this section.

3.4.1 Transfer Learning with EfficientNet-B7

TL was used to utilize the pre-trained weights of EfficientNet-B7 from the ImageNet dataset given the scarcity of labeled medical imaging data. Learned over 14 million general images, these weights provide a strong initialization that helps the model to identify fundamental features such edges, textures, and shapes. This pre-trained foundation

enables the model to effectively generalize even in cases with relative little dataset like ISIC, hence enabling fine-tuning.

Retraining the upper layers of the model on the skin lesion dataset was part of the process of fine-tuning the model. This phase included freezing the bottom tier layers of the model, which are supposed to preserve generic qualities. This method ensures that the model adapt the particular features of medical images, such the variations in color and texture between benign and malignant tumors, thus adjusting its learnt features. The inclusion of metadata into this architecture learning process another level of layer. This allowed the model's ability to be more accurate by include contextual information alongside visual features.

3.4.2 Metadata integration for enhanced predictions

Regarding augmenting the model's prediction ability, metadata is a quite crucial element. For every individual image, metadat which comprises details on the patient's age, gender, and anatomical site, is kept in a cleaned CSV file. This information provides necessary background since some lesion features are impacted by anatomical and demographic elements. Whiteman et. el. presented a study in which it is elaborated that, melanoma is more common in men and older people, and the distribution of lesions varies over several anatomical site [121].

The model transforms the inputs into embeddings through fully connected layers instead of depending just on image features. Following that, these embeddings are concatenated with the image features obtained by EfficientNet-B7, so producing a comprehensive feature representation. This fusion helps the model to learn exact connections between visual patterns and contextual metadata, hence enhancing its capacity to distinguish benign from malignant lesions, even in cases on the borderline.

By means of CSV-based metadata integration, users can change or extend the metadata in line with their needs, therefore simplifying the workflow. This design avoids user providing the same information again during GUI-based predictions and offers adaptability with new datasets.

3.4.3 Training and validation configurations

The training process was designed to maximize the generalizing capacity of the model while also addressing the particular challenges that the dataset presented that is severe imbalance. The optimization approach made use of an Adam optimizer starting with a 1×10^{-5} as an initial learning rate. Given the fine-tuning nature of the pre-trained

EfficientNet-B7 model, this particular learning rate was selected to ensure constant convergence throughout the training process. A cosine annealing scheduler was applied to let the model adjust its parameters in next epochs. As the training developed, the learning rate was progressively lowered using this scheduler.

Over thirty epochs, the training conducted with sufficient time for convergence without risking the issue of overfitting. Early stopping was implemented depending on the validation performance so that, should there be no obvious improvement after a predefined number of iterations, training would be stopped. This approach helped the model avoid overfitting, therefore saving computational resources normally lost.

To maximize the GPU memory that was readily accessible, the batch size was raised to 128. This choice greatly enhanced the stability of gradient estimates by allowing the model to be exposed to a wider range of observations inside every batch, hence enabling stronger generalization. Using sixteen workers, further data loading process optimization made full use of the thirty-two CPU cores. This change helps to remove data loading process bottlenecks, therefore ensuring effective use of the two NVIDIA A100 GPUs throughout the training phase.

To overcomme thesignificant class imbalance in test set, it is addressed by means of a weighted cross-entropy loss function. For every training pass, a special computation of the weights was assigned to ensure that the minority class malignant lesions get the necessary attention. Balanced mini-batches were created using a weighted random sampler, which helped to ensure that malignant instances were adequately represented during the training phase. This stage was particularly crucial during the first training pass since the dataset was balanced such that the focus could be on learning features associated with malignant lesions. During the second training pass, the model was adjusted on the whole dataset to enable effective generalization between benign and malignant cases.

To ensure consistency, each epoch carefully calculated the training steps. Whereas the training loader for Pass 2 comprised of `trainSteps2` steps, the training loader for Pass 1 comprised of `trainSteps1`, steps each epoch as part of the model training execution. The testing loader was configured to run through `testSteps` steps for every epoch overall. This settings allowed the model for thorough evaluation at every epoch. It was chosen to set the training and testing configurations so that they would ascertain the steps per epoch for every epoch of training and testing. The formula applied to find out the count of steps for every epoch is as follows:

$$\text{Steps per Epoch} = \frac{\text{Total Dataset Size}}{\text{Batch Size}} \tag{3.4}$$

For the Pass 1 and Pass 2 training dataset, the number of steps per epoch came out to be 144. The equation shows that the dataset devides the 144 batches each epoch. The decreased size of the testing data also required 75 steps for every epoch. The configurations under discussion ensure that the model will iteratively process the whole dataset once for every epoch with updates applied for every batch. Although the rather smaller test dataset conforms to standard ML procedures to permit effective evaluation, the identical step count for both training passes facilitates model optimization uniformly. Implementing this batching technique is crucial to optimize the optimization of computing efficiency and to get consistent convergence throughout training.

This carefully designed two-pass training approach helped the model to focus on minority class features during the initial training. Then the model extends it by fine-tunining to adjust on a larger dataset. This approach ensured a balance between sensitivity to malignant cases and overall generalization.

3.4.4 Evaluation metrics

The evaluation criteria applied in this thesis implementation gave a whole perspective on the performance of the model. Every metric was deliberately chosen to assess multiple aspects of the ability to predict by the model. Every metric is thoroughly explained in this section together with its significance for the classification of skin lesions.

Train loss and Test loss

The loss function is a metric that describes how well the model is able to predict the appropriate class for each individual sample. A weighted cross-entropy loss was utilized for test cases in this thesis implementation, in order to effectively handle the issue of class imbalance. The test loss shows whether the model could possibly be generalized to unseen data. Conversely, the train loss shows the degree of model fit for the training data. A continually low train and test loss indicates that the model is neither underfitting nor overfitting, therefore reflecting model convergence and resilience. This makes it an important sign of the model's ability to perform well.

Precision

Precision metric measures the proportion of correctly predicted positive samples among all predicted positives. When it comes to the classification of skin lesions, having a high level of precision ensures that the malignant lesions that are identified by the model are, in fact, malignant. The reduction of false positives, which could result in patients undergoing unneeded biopsies or experiencing anxiety, is of the utmost importance.

Recall

Recall, also known as sensitivity or true positive rate, evaluates the proportion of actual positives correctly identified by the model. If the model presenting a high recall, then it means this helps to lower the frequency of false negatives by ensuring that most malignant lesions are detected. This metric is especially important in the framework of biomedical applications, where the lack of a malignant lesion could have major consequences.

Accuracy

Accuracy refers to the percentage of accurate predictions produced across all samples. It is important to understand that the abbreviations TP, TN, FP, and FN represent, respectively, true positives, true negatives, false positives, and false negatives. Accuracy is prone to be influenced by the class that is dominated by the majority, so it may not be sufficient in datasets that are not balanced even if it provides an accurate depiction of the performance of the model.

ROC-AUC

ROC-AUC evaluates the ability of the model to differentiate across classes over many threshold values. Plotting the actual positive rate (sensitivity) against the false positive rate (1-specificity) the ROC curve indicates the performance of the model. When the score is 0.5 in plotting, it indicates random guessing of the model, but an AUC of 1.0 or close to 1.0 represents perfect classification.Given the class imbalance in this work, ROC-AUC proved to be a crucial indicator of the model's capacity to distinguish benign from malignant lesions.

F1 Score

The F1 Score offers a fair assessment of the model's ability to predict positive samples accurately and completely via harmonic mean of accuracy and recall. In unbalanced datasets particularly, this metric is quite helpful as it takes false positives and false negatives into account, therefore It offers a comprehensive perspective on the performance of the model.

In conclusion, during training, metrics such as loss, accuracy, precision, recall, F1 score, and ROC-AUC were monitored to assess the model's learning progress. These metrics provided insights into how well the model fit the training data and adapted adjustments for hyperparameters. For validation or testing, the same metrics were computed on the test set to evaluate the model's generalization performance. This comprehensive evaluation ensured that the model performed well on unseen data, meeting the diagnostic requirements of skin lesion classification. The combination of these metrics provided a thorough evaluation of the model's predictive capabilities, ensuring reliability and

robustness. These metrics established a strong foundation for building a dependable diagnostic tool.

Confusion Matrix

Confusion matrix is also an efficient metric for evaluating DL model's performance, in this thesis it assisted greatly in the evaluation of classification model's performance. By categorizing the model's prediction into four such as; true positives, true negatives, false negatives, and false negatives. It offers comprehensive cummulative analysis of the model's learning process. In this thesis, confusion matrices produced following every five epochs. These matrices gave a whole analysis of the learning process the model experienced.

Before the early halting mechanism was triggered, ResNet-50 completed fifteen epochs in all total. Three confusion matrices were thus generated during training of the network. Early stopping for EfficientNet-B7 and EfficientNet-V2 was triggered for the same reason after twenty epochs, which produced four confusion matrices for every model. With the empolyment of these matrices, a granular perspective is presented on which the models could distinguish benign from malignant lesions over training time.

Confusion matrix analysis showed the performance trends in the model including the rises in true positive and true negative rates across each epoch. This demonstrated the ability of the model to adapt the complexity of the dataset through means such as handling the class imbalance, therefore demonstrating to its adaptability. Moreover, the confusion matrices highlighted areas in which the model struggled, such as misclassifying benign lesions as malignant or vice versa. This allowed hyperparameter tuning or the training process to be specifically improved.

Confusion matrices allowed the total evaluation to be conducted so that one could fully understand the benefits and drawbacks of the model. This iterative monitoring not only ensured the robustness of the learning process but also validated the effectiveness of the early stopping mechanism in halting training at an optimal point.

3.5 Semi-Supervised prediction mechanism in the GUI tool

Designed for runtime melanoma detection, this thesis implements the GUI tool that incorporates EfficientNet-B7 trained model for real-time predictions, with multiple images from a same patient and its metadata information as an input. Before being passed through the EfficientNet-B7 trained model for feature extraction, every chosen image goes through scaling, normalization, and conversion to a tensors. Metadata features

are handled in parallel and processed with image features before being passed to fully connected layers of the network architecture for classification task. This method provides effective combination of contextual and visual information for strong predictions.

The final result shows the possibility of melanoma by a probability score ranging from 0 to 1. This thesis presents a confidence-based aggregation approach for batch predictions. Should any individual image score above a high-confidence level that is 0.9, the lesion is marked as malignant. This threshold is not the single factor that decides multiple images to be called melanoma. The next threshold considers the average score across all chosen image after computing it confidence score; lesions identified as malignant have an average score higher than 0.5.

$$\mu = \frac{\sum P}{n}$$

$$\text{Prediction} = \begin{cases} \text{Malignant}, & \text{if } \exists P > 0.9 \text{ or } \mu > 0.5 \\ \text{Benign}, & \text{otherwise} \end{cases}$$

3.5.1 Role of the `.pth` file

The `.pth` file is an important factor on which a deployment pipeline depends as it offer following capabilities;

- **Model weights:** Pre-trained parameters enabling the model to make accurate predictions. Then comes the architecture which contains, definitions of the layers, including modifications for metadata handling.

These are the weights which are derived during training by optimizing a loss function (Focal Loss in this thesis), measure the difference between predicted probabilities and actual labels:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

The .pth file ensures portability and efficiency, enabling the GUI tool to load pre-trained weights for immediate predictions without requiring retraining. Its role is central to leveraging the model's learned parameters for final inference.

3.5.1.1 Input to the model

The first step in GUI tool is to select images of the patient. So after selecting the images, each image is preprocessed to ensure consistency and compatibility with the model. Each

input image is then resized to 224×224 pixels to match the input dimensions expected by the EfficientNet-B7 model according to this thesis implementation. Although the model's default is 600×600 pixels. This resizing ensures uniformity across images, which is crucial for consistent feature extraction. Then the next step is normalization. Pixel intensity values are normalized using the mean and standard deviation values calculated during training (e.g., ImageNet dataset normalization). These preprocessing steps ensures that inputs conform to the model's requirements and maintain consistency with the data used during training. This step aligns the input data distribution with the training data distribution, improving the model's predictive performance. The resized and normalized image is converted into a tensor format, which serves as the input to the DL model. Tensors are multi-dimensional arrays that allow efficient computation in modern DL frameworks like PyTorch.

3.5.1.2 Output from the model

Following the preprocessing is completed, each image and the metadata associated with it fed into the model layers. After computing, the outcome is a single numerical number that is referred to as a **logit** value. This number indicates the unscaled confidence in the malignant class, and it is the result of the computation.

Considering the following example, let's consider (logit) is equal to -1.2.

$$\text{Output (logit)} = -1.2 \quad \text{(Unscaled confidence score before activation)}$$

The sigmoid function maps the logit to a probability:

$$P(\text{malignant}) = \sigma(\text{logit}) = \frac{1}{1 + e^{-\text{logit}}}$$

This logit value is then passed on to the sigmoid function before we proceed. For transforming the logit into a number that have a range between 0 and 1, the following procedure is carried out:

- If logit $= -1.2 \quad \Rightarrow \quad P(\text{malignant}) = \sigma(-1.2) \approx 0.23 \quad$ (classified as benign)
- If logit $= 2.0 \quad \Rightarrow \quad P(\text{malignant}) = \sigma(2.0) \approx 0.88 \quad$ (classified as malignant)

The probability score that is produced by the sigmoid function is referred to as the confidence score, which serves to measure the level of certainty associated with the model. **confidence score**, quantifying the model's certainty:

- Scores close to 1 indicate high confidence in malignancy.
- Scores close to 0 indicate high confidence in benign classification.

3.5.1.3 How the `.pth` file generates confidence

When training is complete, the weights and biases that were learnt are saved in the `.pth` file. These parameters enable the mapping of input characteristics to output predictions, and the layers are responsible for carrying out the functions listed below.

Model's layers:

- Extract low-level features like edges and textures from the convolutional layers.
- **Fully connected layers:** Aggregate features into a high-level representation for classification.

During inference, the model processes inputs through:

1. Convolutional layers applying learned filters for feature extraction.
2. Non-linear activation functions, such as Swish, capturing complex patterns.
3. Fully connected layers combining features to produce a final logit.

The logit is then converted into a confidence score using the sigmoid function. This process ensures reliable inference, leveraging the learned parameters stored in the `.pth` file to generate actionable confidence scores.

The `.pth` file enables the model to perform accurate and transparent inference by applying learned weights and biases to unseen data. By combining image and metadata features, the GUI tool provides real-time predictions, producing interpretable probabilities essential for clinical decision-making.

3.5.2 Prediction mechanism

To address the issues that are connected with various image inputs for the same patient, the prediction method that was constructed in this thesis was designed to handle such challenges. Using this method, credible predictions are achieved by combining the scores of individual images and utilising metadata to facilitate decision-making that takes into

account the associated context. The procedure may be broken down into three very important stages: preprocessing, individual predictions, which were covered in the section on confidence scores, and final prediction aggregation.

Data preprocessing is an essential step to ensure that all input images are compatible with the model and that the metadata is correctly formatted. The following steps are performed: Image resizing, normalization, and tensor conversions, all of which have been discussed in the aforementioned section.

- **Metadata encoding:** Patient metadata (e.g., age, sex, anatomical site) is mapped to numerical values using predefined encodings. This step ensures that the metadata is represented in a machine-readable format, facilitating its integration with image features during prediction.

```
sex_mapping = {
    'male': 0,
    'female': 1
}

anatomy_mapping = {
    'head/neck': 0,
    'torso': 1,
    'lower extremity': 2,
    'upper extremity': 3,
    'palms/soles': 4,
    'oral/genital': 5,
}
```

These preprocessing steps standardize the inputs, ensuring that the model can make accurate and consistent predictions regardless of variations in raw input data.

3.5.2.1 Final prediction aggregation

The method compiles an average of the confidence scores from multiple distinct images in order to improve reliability and reduce the effect of inconsistent predictions or outliers. This aggregation process uses two significant criteria in order establish the final classification.

- **High confidence threshold:** Every individual image with a confidence score higher than 0.9 is said to have a high-confidence malignant prognosis. This guarantees the identification of clearly existing cases of cancer free from depending on average scores.

- **Average threshold:** This threshold helps one to examine scenarios in which several images support the overall decision by means of their average probability among all the chosen images. The average probability is computed from the following formula.

$$\mu = \frac{\sum_{i=1}^{n} P_i}{n}$$

  where P_i is the probability score for the ith image, and n is the total number of images.

The final decision is made using the following rule:

$$\text{Prediction} = \begin{cases} \text{Malignant}, & \text{if } \exists P_i > 0.9 \text{ and } \mu > 0.5 \\ \text{Malignant}, & \text{if } \mu > 0.5 \\ \text{Benign}, & \text{otherwise} \end{cases}$$

This method ensures that the model maintains sensitivity to high-confidence predictions and makes use of the combined information from multiple images. The model strikes a balance between robustness and accuracy in its last predictions by weighing both individual and aggregated probability.

The multi-image and confidence-based aggregation mechanism addresses several key challenges in skin lesion classification by combining multiple images and integrating metadata in a way that improves both accuracy and practicality. First, it effectively handles ambiguity of borderline cases by aggregating predictions from multiple images, thus reducing the impact of borderline or uncertain cases. Second, the incorporation of metadata strengthens the model's ability to contextualize lesions with similar visual characteristics. Third, setting a high-confidence threshold ensures that the system does not miss clear cases of malignancy, even when the average score might fall just below a conventional cutoff. Finally, our batch prediction approach saves time and streamlines workflow by letting users to evaluate multiple images at once. These techniques together show how well strong DL methods combine with realistic concerns for medical, that uses to produce interpretable and consistent predictions. Supporting the main objective of improving diagnosis accuracy and enabling clinical decision-making.

Chapter 4

Results and Discussion

4.1 Chapter overview

The methodology described in Chapters 1, 2, and 3, this chapter presents and critically discusses the performance of three DL models that are **ResNet-50**, **EfficientNet-B7**, and **EfficientNet-V2**. These models are used to classify skin lesions that are either malignant or benign. Both training and testing were performed on each model in two separate passes in order to overcome the extreme class imbalance and make use of metadata integration. The chapter is organised as follows:

- A brief restatement of the experimental setup
- Each model's performance results and findings
- Comparative evaluation and critical analysis
- Reflection on how the results align with existing literature
- Discussion of limitations, potential improvements, and avenues for future research

The chapter also discusses which architectural approach yields the most robust classification performance while also contextualizing these findings with state-of-the-art work in automated melanoma detection [6],[7],[8].

4.2 Experimental recap

The dataset has multiple images of every patient, each of which is categorised as either benign or malignant in line with what was covered in Chapter 3. Every model underwent two training passes to prevent an imbalance between the classes.

- **Pass 1:** Focused on oversampling malignant cases, using heavy augmentation to create synthetic minority samples.
- **Pass 2:** Trained and fine-tuned the model on remaining benign images and 7,024 malignant images which had been trained in pass 1 already, this ensuring the model learns from over sampled benign cases and suffice malignant images.

All models were trained under the same regime:

1. **Image preprocessing:** Conversion from DICOM to JPEG, resizing, normalization, metadata integration (age, sex, anatomical site), and ground truth.
2. **Loss function:** Focal Loss chosen for its strength in addressing class imbalance [120].
3. **Evaluation metrics:** Accuracy, precision, recall, F1 score, ROC AUC, and average precision (AP) are the metrics applied in evaluation. These evaluations helped to get a more realistic view of the classification ability of every model [6].
4. **Early stopping:** It is triggered at a point where the test loss or test F1 Score approaches a plateau in order to prevent overfitting.

PyTorch, with GPU acceleration, was used for the whole training process in line with the pipeline described in Chapter 3.

4.3 Results: ResNet-50

4.3.1 Training and validation performance

Figure 4.1 (from the model training logs) shows the convergence patterns for ResNet-50. The initial epochs demonstrated rapid improvement in train loss and accuracy. After around 5–6 epochs, the model began converging, with diminishing returns in further epochs. Early stopping occurred at around the 10th epoch in Pass 1 and 9th epoch in Pass 2. Table 4.1 summarizes the *final* test results (i.e., the last epoch of training before early stopping) on Pass 2, which is the culminating performance of ResNet-50.

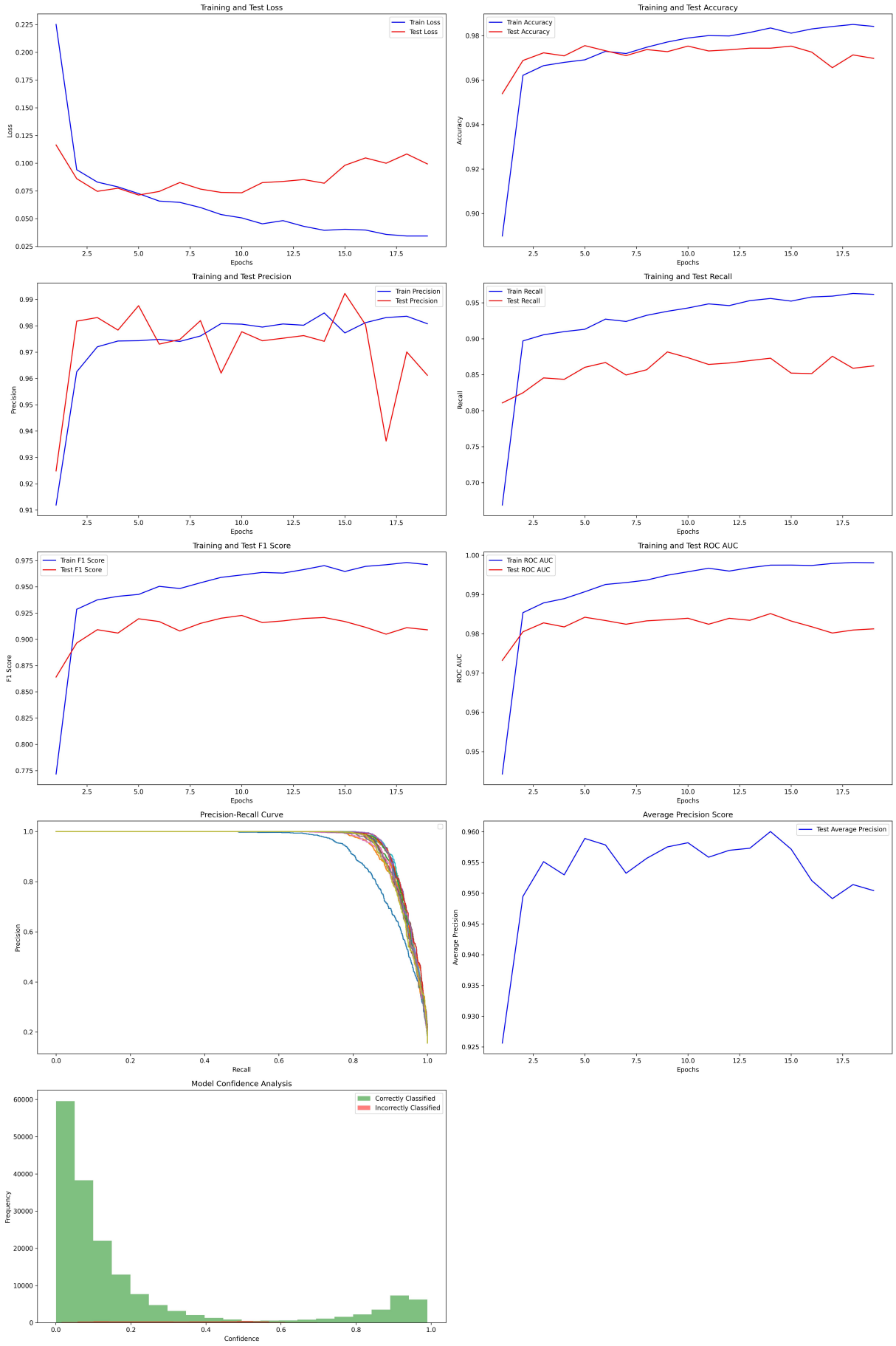


Figure 4.1: ResNet-50 Evaluating Results

Table 4.1: Final Test Metrics of ResNet-50 on Pass 2

Metric	Loss	Accuracy	Precision	Recall	F1 Score	ROC AUC
Value	≈ 0.10	≈ 0.974	≈ 0.975	≈ 0.869	≈ 0.918	≈ 0.984

With an F1-score of about 0.918 score and a high ROC-AUC ($\approx$ 0.984), ResNet-50 attained a final accuracy of almost 97.4%, according to the data shown in Table 4.1.

- **ROC AUC:** If the score is 0.98, it suggests that the model is able to appropriately rank benign and malignant classes, even when the threshold criteria diverge from one another.
- **Precision vs. Recall:** Precision ($\approx$ 0.975) remained somewhat higher than Recall ($\approx$ 0.869), consequently, there were a higher number of false negatives even though there were fewer false positives. The existence of this difference suggests that ResNet-50 might have missed some malignant cases.
- **Average Precision:** The concept that the model's precision remained strong across the recall spectrum was reinforced by the fact that the average precision consistently hovered around 0.95–0.96 in the last epochs.
- ResNet-50 performs consistently in differentiating between benign and malignant lesions, even if the small discrepancy between precision and recall highlights areas for development in identifying all malignant cases. These results show generally that ResNet-50 performs really brilliantly in these configurations.

4.3.2 Observations and challenges

ResNet-50 went through strong performance result, but its difference in precision and recall indicates that it favours fewer false positives at the potential cost of missing some malignant instances. Although fewer false positives mean fewer unnecessary biopsies, in real-world screening programmes, a possibly lower recall can be critical, considering missing malignant lesions can be fatal [1].

4.4 Results: EfficientNet-B7

4.4.1 Training and validation performance

Figure 4.2 (from the model training logs) shows the convergence patterns for EfficientNet-B7.

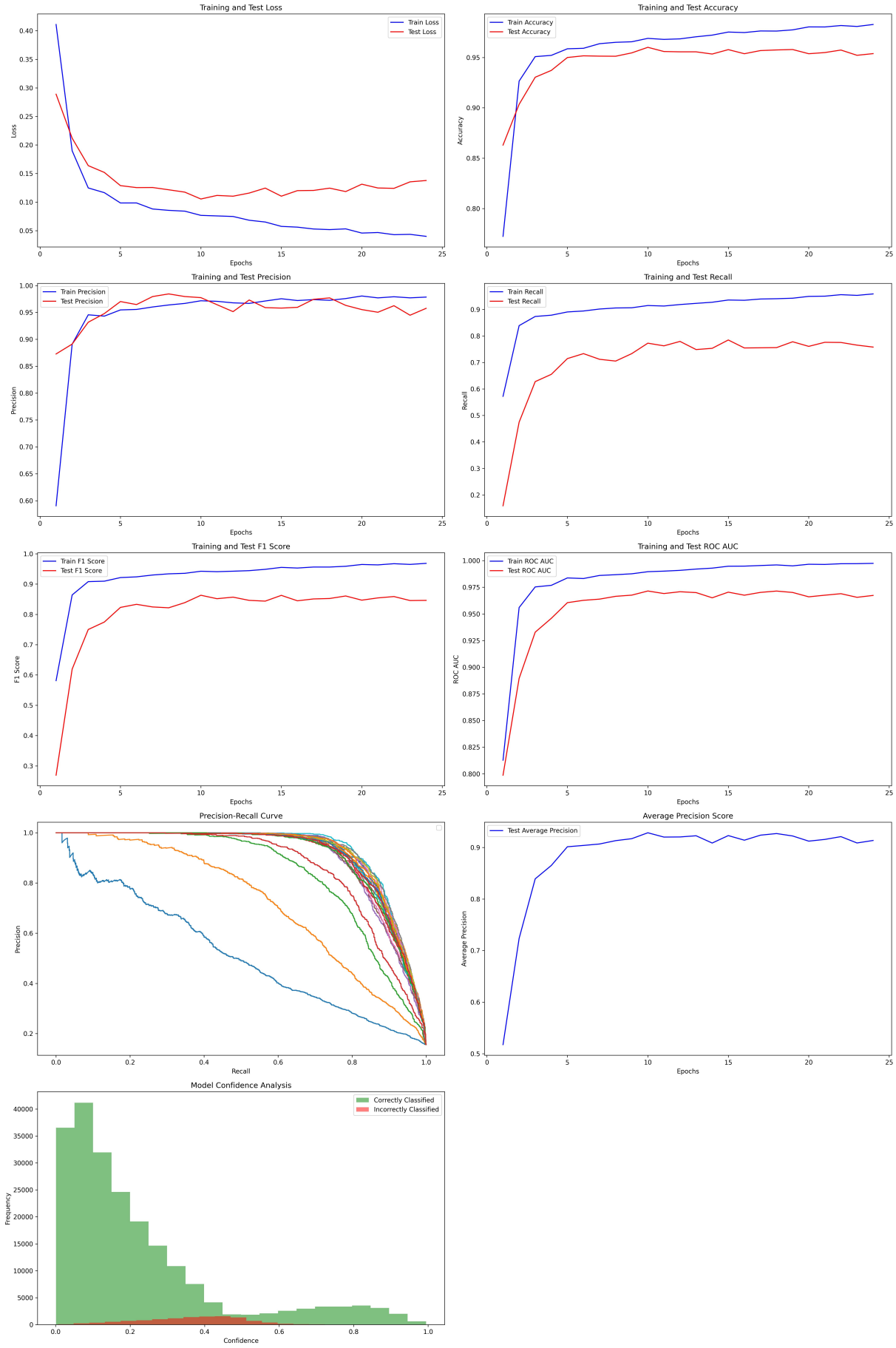


Figure 4.2: EfficientNet-B7 Evaluating Results

EfficientNet-B7 is known for high accuracy in image classification tasks due to its compound scaling [2]. Over the first 3–5 epochs, the network quickly reduced training loss from over 0.4 to around 0.1, a steeper initial descent compared to ResNet-50. Early stopping typically occurred after about 15 epochs in Pass 1 and fewer epochs in Pass 2, reflecting faster convergence. Table 4.2 presents the final (Pass 2) test performance metrics.

Table 4.2: Final Test Metrics of EfficientNet-B7 on Pass 2

Metric	**Loss**	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**ROC AUC**
Value	≈ 0.12	≈ 0.957	≈ 0.955	≈ 0.785	≈ 0.863	≈ 0.970

Key highlights:

- **Accuracy:** On pass 2, the final accuracy of the model was approx. $\approx 95.7\%$, indicating that EfficientNet-B7 generalised to unseen data rather well.
- **Precision vs. Recall:** Precision(≈ 0.955) and Recall (≈ 0.785) varied somewhat but most notably from ResNet-50 than from each other. EfficientNet-B7 has missed a larger percentage of malignant lesions in contrast to its total predictions, even though it produced fewer false positives. Moreover, EfficientNet-B7 achieved strong precision, its lower recall highlights the need for more balanced sensitivity towards malignant lesions. The trade-off made when focusing on fewer false positives at the risk of missing some borderline cases is reflected in the higher test loss and F1 score.
- **F1 Score:** Within the range of 0.86, the harmonic mean of Precision and Recall is decent but not outstanding.
- **Train vs. Test Loss:** Consistently the reduced recall observed, the final test loss (≈ 0.12) was somewhat larger than ResNet-50.

4.4.2 Observations and potential reasons

1. **Faster convergence.** Like early on in ResNet-50, the sharp decline in training loss aligns with its findings, showing that EfficientNet variants converge faster because of optimal layer scaling [2].
2. **Over-penalizing malignant cases.** The architecture of the model may still be biassed towards labelling borderline lesions as benign even with the augmentation to minority class. Although Focal Loss helps somewhat counteract this result, it is not totally eradicated; therefore, further hyperparameter tuning, that is changing γ in focal loss may be required.

4.5 Results: EfficientNet-V2

4.5.1 Training and validation performance

Figure 4.3 (from the model training logs) shows the convergence patterns for EfficientNet-V2. EfficientNet-V2 retains the core idea of compound scaling while refining training speed [122]. As the logs indicate, it took around 7–10 epochs for the model to stabilize in both Pass 1 and Pass 2. Table 4.3 consolidates the final test performance metrics from Pass 2.

Table 4.3: Final Test Metrics of EfficientNet-V2 on Pass 2

Metric	Value
Loss	≈ 0.10–0.12
Accuracy	≈ 0.968
Precision	≈ 0.95–0.98
Recall	≈ 0.85
F1 Score	≈ 0.90–0.91
ROC AUC	≈ 0.978

Noteworthy points:

- **Overall Accuracy:** In terms of overall accuracy, it reached approximately 96.8%, which is somewhat higher than the ultimate accuracy of ResNet-50 but not by a significant margin.
- **Precision:** Often near or above 0.95, indicating strong performance at correctly labeling malignant lesions that it classifies as malignant.
- **Recall:** Hovering around 0.85, which indicates improvements over EfficientNet-B7 but marks a slight decline in comparison to ResNet-50's best-case recall in some epochs.
- **F1 Score:** Usually ranging between 0.90–0.91, this score indicates a balanced performance for malignant classification.
- **ROC AUC:** 0.978 approximately in the final epochs, presenting a strong discriminative power across threshold.
- Based on most measures, EfficientNet-V2 turned out to be the best-performing model since it can efficiently manage the complexity of melanoma categorisation. Although the trade-off between precision and recall shows a possibility to improve sensitivity even more while keeping its great specificity from outstanding performance.

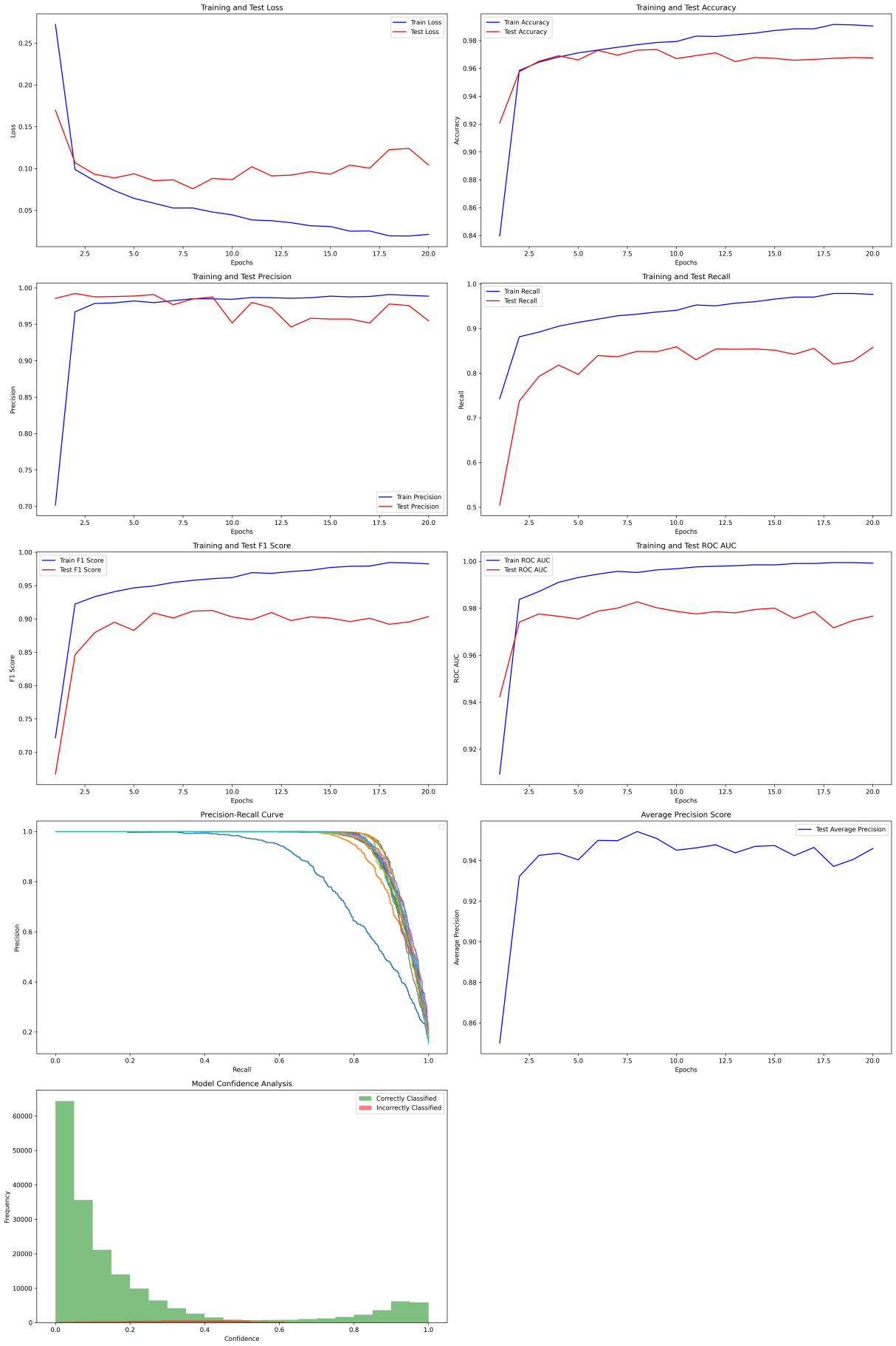


Figure 4.3: EfficientNet-V2 Evaluating Results

4.5.2 Observations and explanation

The improved speed of convergence in EfficientNet-V2 relative to EfficientNet-B7 is consistent with the architectural enhancements introduced by Tan et al. [122]. In terms of classification, it offered a slightly improved harmonic mean (F1) compared to EfficientNet-B7 but not a dramatic leap. This indicates that while the additional refinements in V2 help training efficiency, the final classification gains are moderate when controlling for other factors such as oversampling and focal loss.

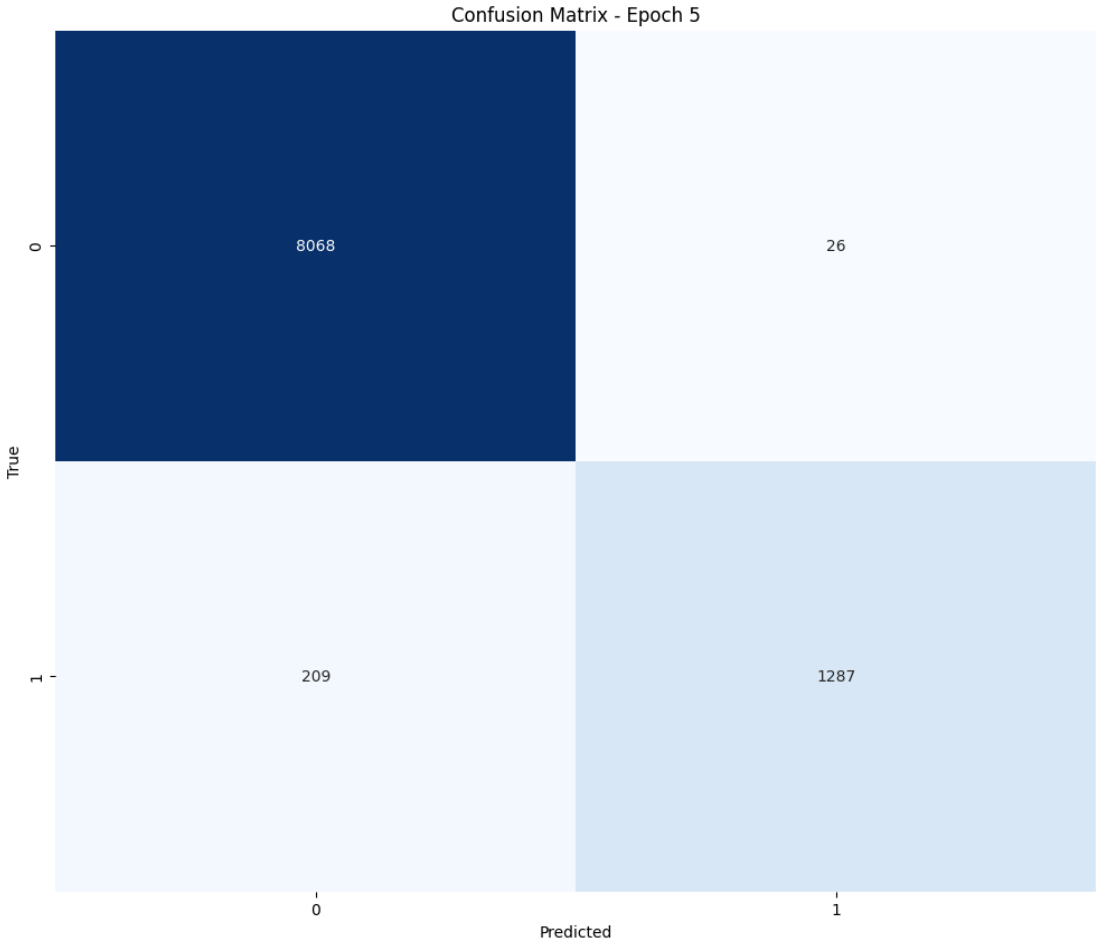
4.5.3 Confusion Matrix analysis

This section offers a critical appraisal of how classification outcomes evolve over time for each of the three DL architectures based on confusion matrices recorded at key training epochs. By inspecting the distribution of true negatives, false positives, false negatives, and true positives, we gain a deeper understanding of each model's behavior in distinguishing benign (class 0) versus malignant (class 1) lesions under conditions of significant class imbalance.

ResNet-50 (Epochs 5, 10, and 15). Figures 4.4a 4.5a–4.5 illustrate the confusion matrices for ResNet-50 at Epoch 5, Epoch 10, and Epoch 15, respectively.

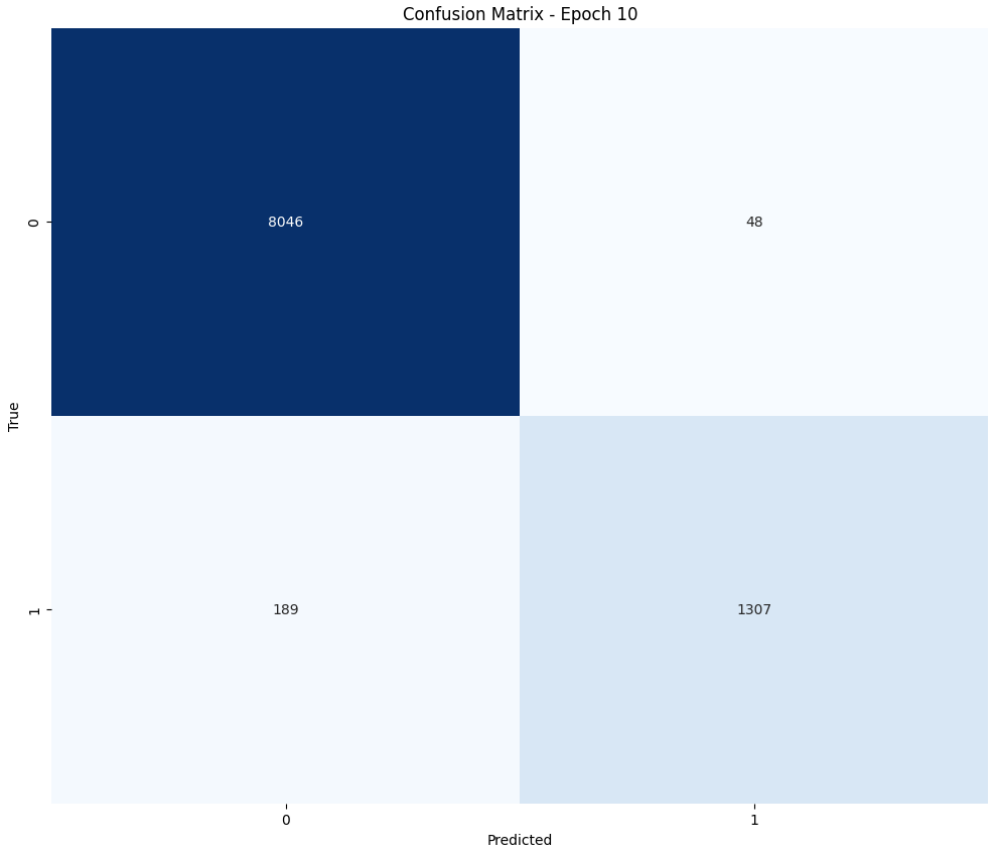
- **Epoch 5 (Figure 4.4a):** The matrix shows a predominantly large TN count ($\approx$ 8,616) and a modest FP count ($\approx$ 26), indicating that ResNet-50 swiftly learns to identify benign lesions. However, the FN value ($\approx$ 209) remains comparatively high, revealing that many malignant samples were misclassified early on.
- **Epoch 10 (Figure 4.5a):** By this midpoint, the model sees moderate shifts. TN remains high ($\approx$ 8,046), though FP rises slightly ($\approx$ 40). The corresponding FN ($\approx$ 189) decreases relative to Epoch 5, and TP ($\approx$ 1,307) increases. This trend suggests the network is gradually improving its sensitivity to malignant lesions, even if it occasionally over-corrects and flags a few more benign lesions as malignant.
- **Epoch 15 (Figure 4.5):** In the final reported confusion matrix, ResNet-50 yields $\approx$ 8,076 TN, $\approx$ 16 FP, $\approx$ 221 FN, and $\approx$ 1,275 TP. Notably, the FN count is slightly higher than at Epoch 10, yet the FP count is the lowest of all three epochs. This outcome reflects a trade-off: the model succeeds in reducing false indicators (that is, FPs), but it becomes rather less aggressive in identifying malignant lesions, which is the cause for the little increase in FNs.

- ResNet-50 generally demonstrated a primarily high true negative rate across all epochs, which highlights the great specificity of the network while also indicating the need for threshold changes in the event that lower FN rates (and hence higher recall) are required in clinical applications.



(a) Epoch 5

Figure 4.4: Confusion Matrix evaluations by ResNet-50 at Epoch 5



(a) Epoch 10

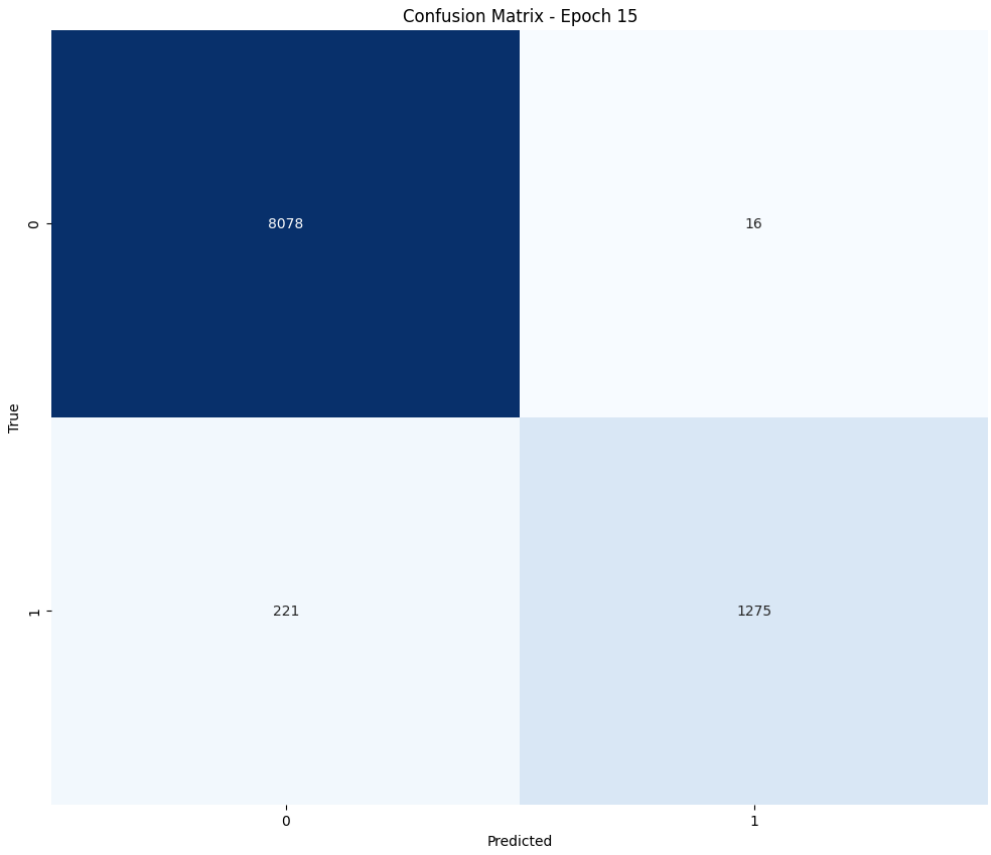
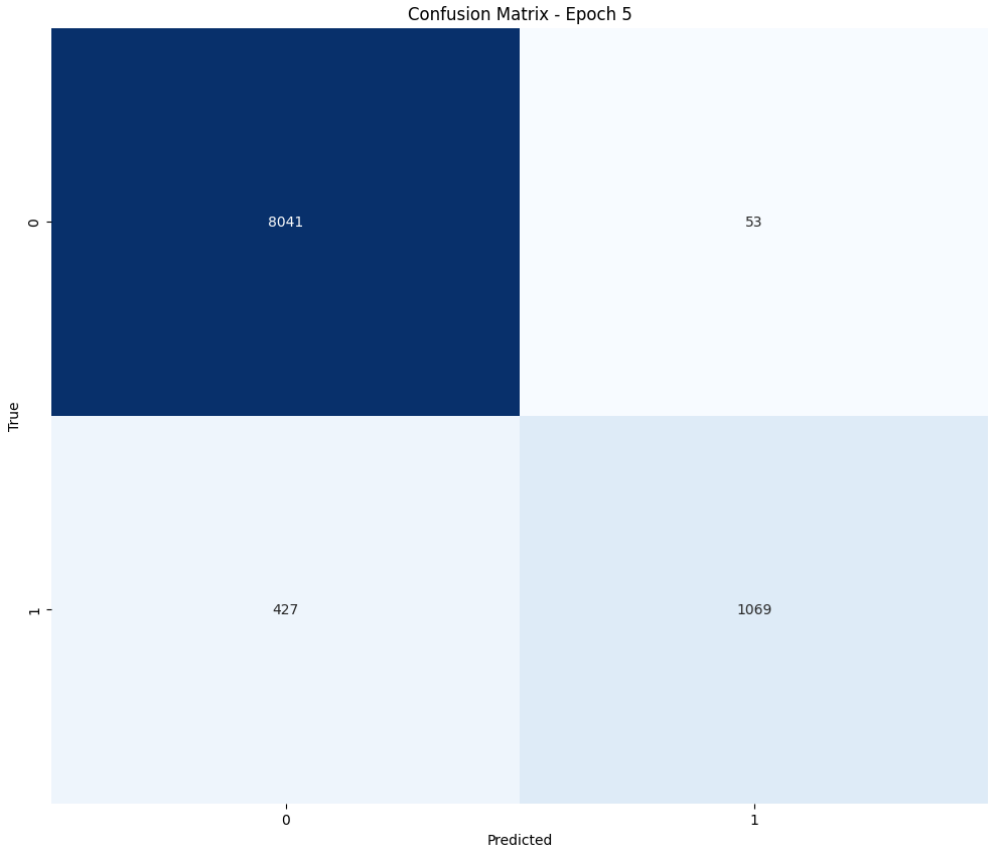


Figure 4.5: Confusion Matrix evaluation by ResNet-50 Epoch 10 (top) and Epoch 15 (bottom).

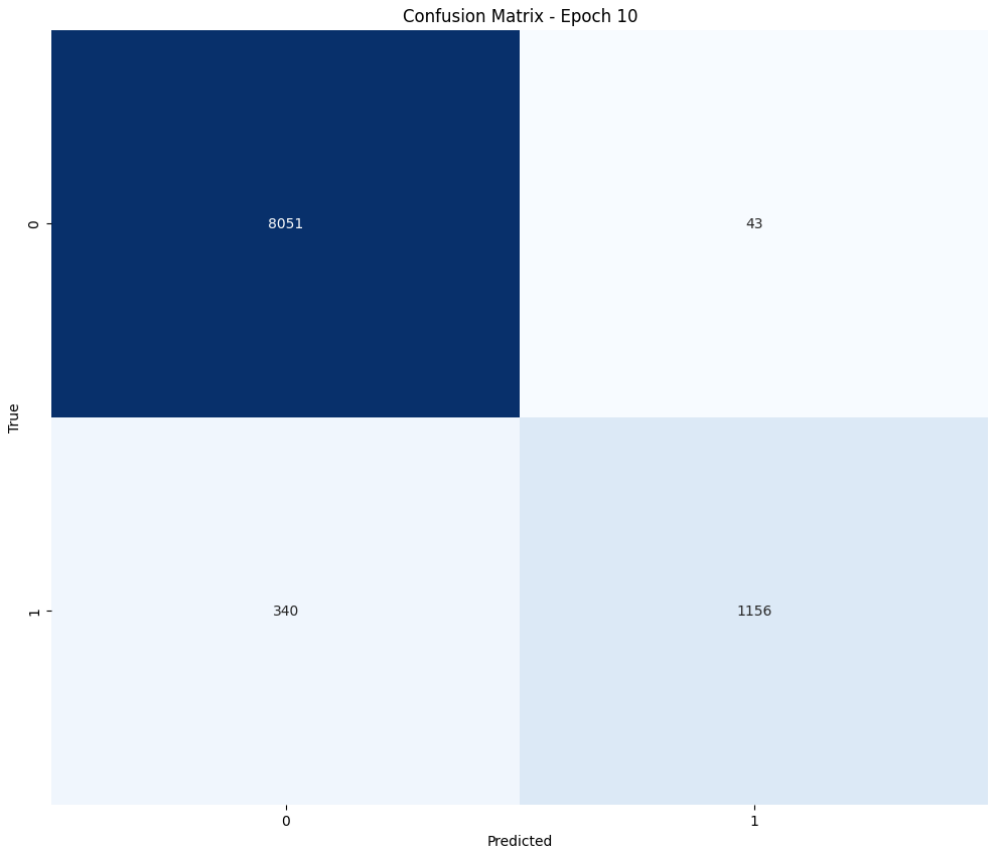
EfficientNet-B7 (Epochs 5, 10, 15, and 20). Figures 4.6a 4.6b–4.7a4.7b show the confusion matrices for EfficientNet-B7 at Epochs 5, 10, 15, and 20.

- **Epoch 5 (Figure 4.6a):** Similar to ResNet-50, B7 rapidly reaches a large TN ($\approx$ 8,041) but exhibits a higher number of FP ($\approx$ 53) compared to ResNet-50's early epochs. Meanwhile, FN ($\approx$ 427) is also substantial, suggesting the model initially struggles to correctly identify many malignant cases.
- **Epoch 10 (Figure 4.6b):** By this stage, TN remains high ($\approx$ 8,051), while FP decreases ($\approx$ 43). Notably, FN dips to $\approx$ 360 and TP rises to over 1,100, indicating an improvement in recognizing malignant lesions.
- **Epoch 15 (Figure 4.7a) and Epoch 20 (Figure 4.7b):** B7's final two matrices show incremental improvements in TP ($\approx$ 1,174 at Epoch 15, $\approx$ 1,138 at Epoch 20), though the difference between these epochs is not uniform. FN remains in the 300+ range. FP similarly hovers around 80–90. These patterns suggest that while B7 converges to a high-precision system (fewer benign misclassifications), it still yields a moderately high FN count. This resonates with the broader observation that EfficientNet-B7, despite strong architectural scaling, can be conservative in flagging malignant lesions, thus maintaining more false negatives.

Hence, EfficientNet-B7 shows fast convergence in early epochs but stabilizes with a precision-oriented bias that may require threshold tuning for applications where recall is paramount.

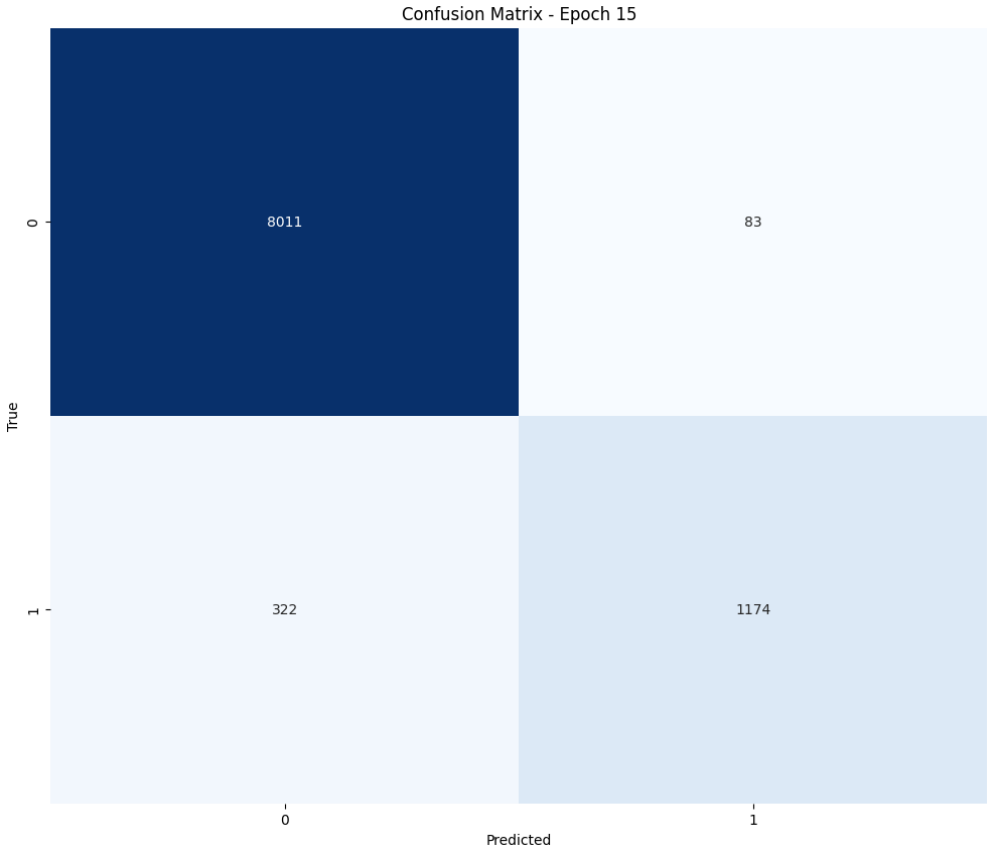


(a) Epoch 5

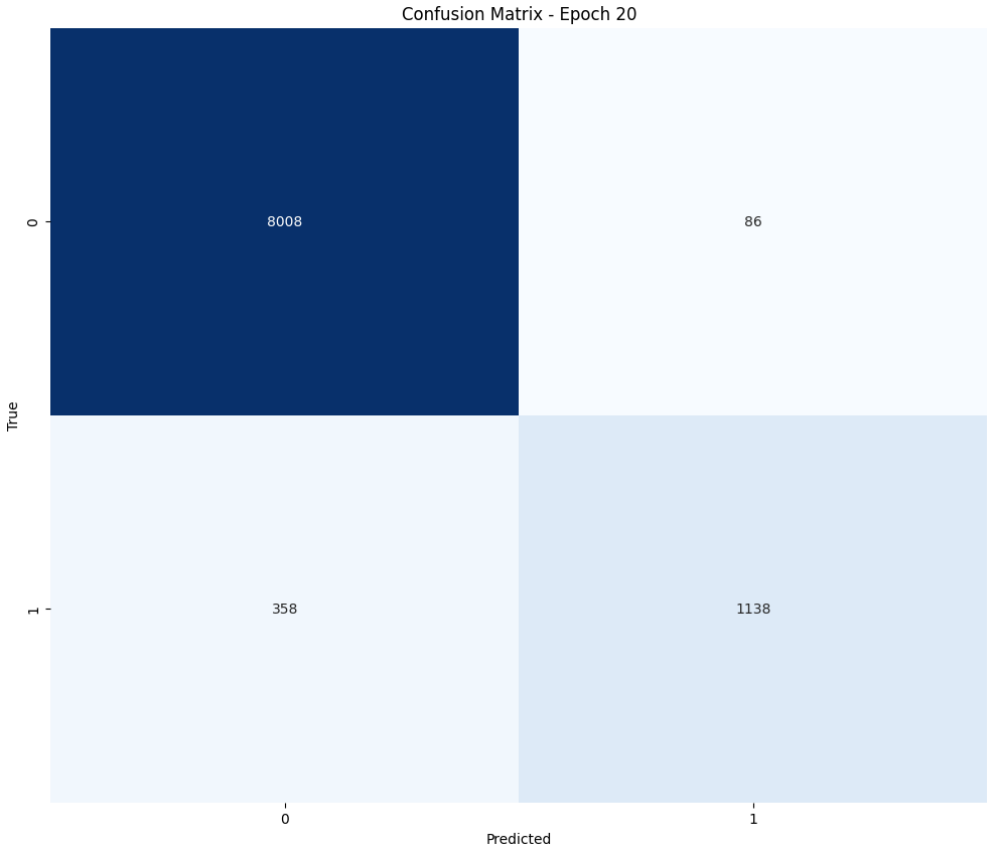


(b) Epoch 10

Figure 4.6: Confusion Matrix evaluations by EfficientNet-B7 at Epoch 5 (top) and Epoch 10 (bottom).



(a) Epoch 15



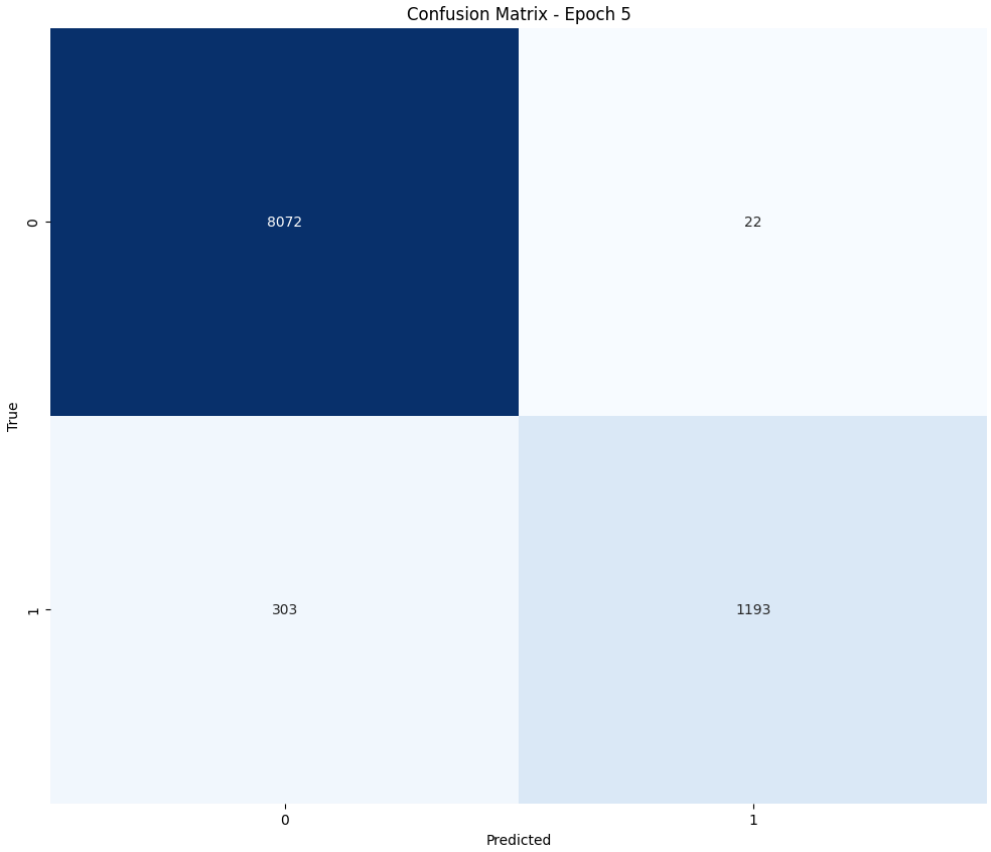
(b) Epoch 20

Figure 4.7: Confusion Matrix evaluations by EfficientNet-B7 at Epoch 15 (top) and Epoch 20 (bottom).

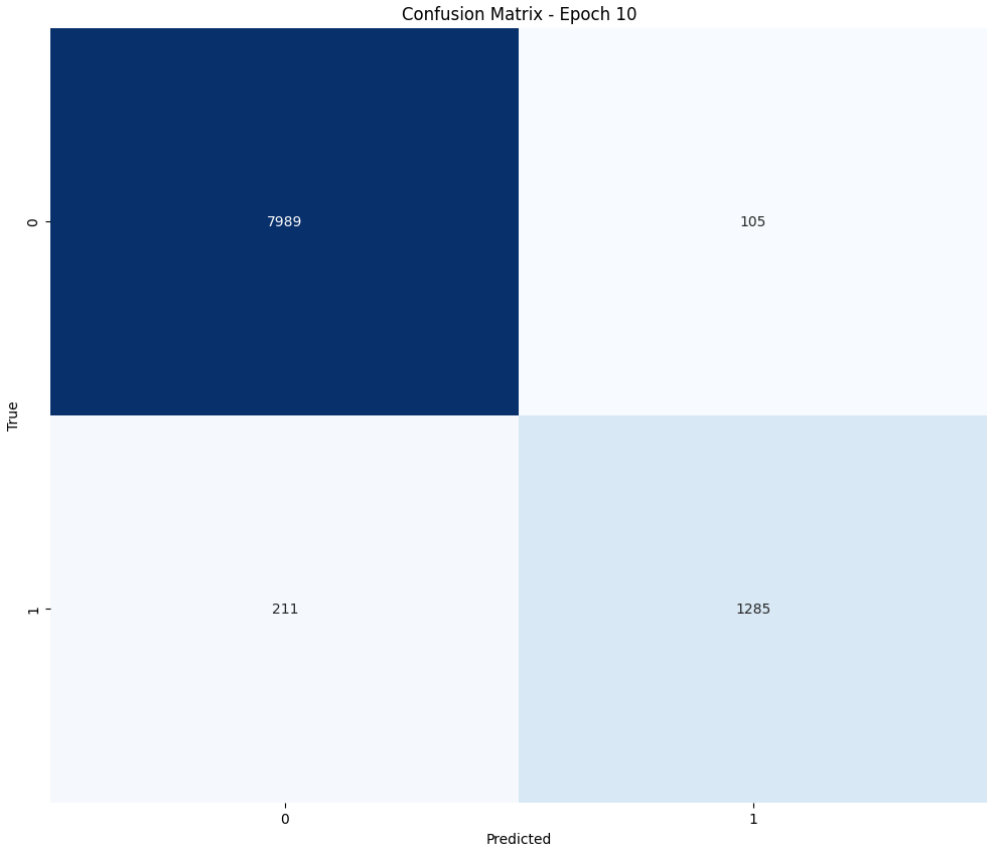
EfficientNet-V2 (Epochs 5, 10, 15, and 20). Figures 4.8a4.8b–4.9a4.9b depict the confusion matrices for EfficientNet-V2 at Epochs 5, 10, 15, and 20.

- **Epoch 5 (Figure 4.8a):** TN is $\approx$ 8,072, FP stands at $\approx$ 22, while FN ($\approx$ 363) is relatively high. Similar to the other models, the early-stage performance indicates a strong initial recognition of benign lesions but substantial misclassifications of malignant cases.
- **Epoch 10 (Figure 4.8b):** TN dips slightly ($\approx$ 7,969), with FP $\approx$ 105. FN reduces to around 211, and TP surpasses 1,200, evidencing a marked increase in malignant detection.
- **Epoch 15 (Figure 4.9a) and Epoch 20 (Figure 4.9b):** By Epoch 15, TN remains large ($\approx$ 8,082), though FP has risen ($\approx$ 92), indicating a shift toward labeling uncertain lesions as malignant. FN ($\approx$ 222) is moderately lower than at Epoch 5, and TP $\approx$ 1,274. At Epoch 20, TN ($\approx$ 7,936) decreases a bit, while FP ($\approx$ 90) stays similar; FN remains around 210–220, and TP hovers in the 1,200+ range. These outcomes reflect a generally balanced strategy where the network retains high specificity but modestly improves recall relative to earlier epochs.

Across these four confusion matrix, EfficientNet-V2 shows steady improvements in TP while maintaining fairly stable FP, aiming for a balanced performance overall.

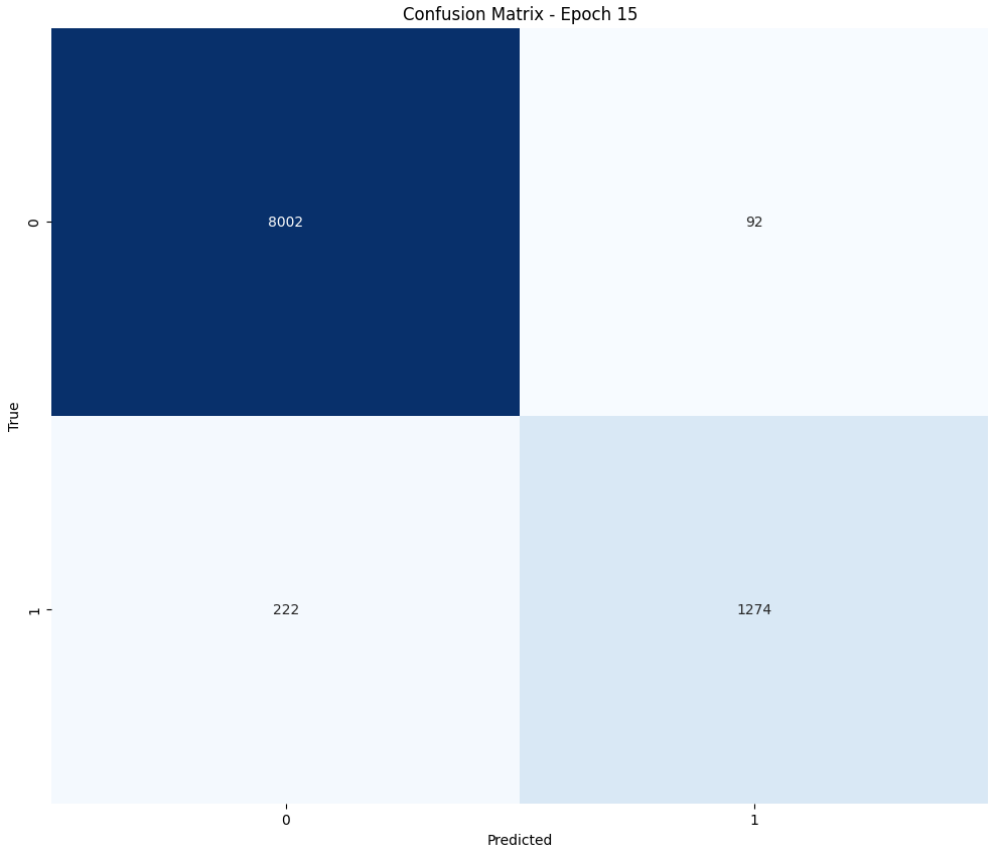


(a) Epoch 5

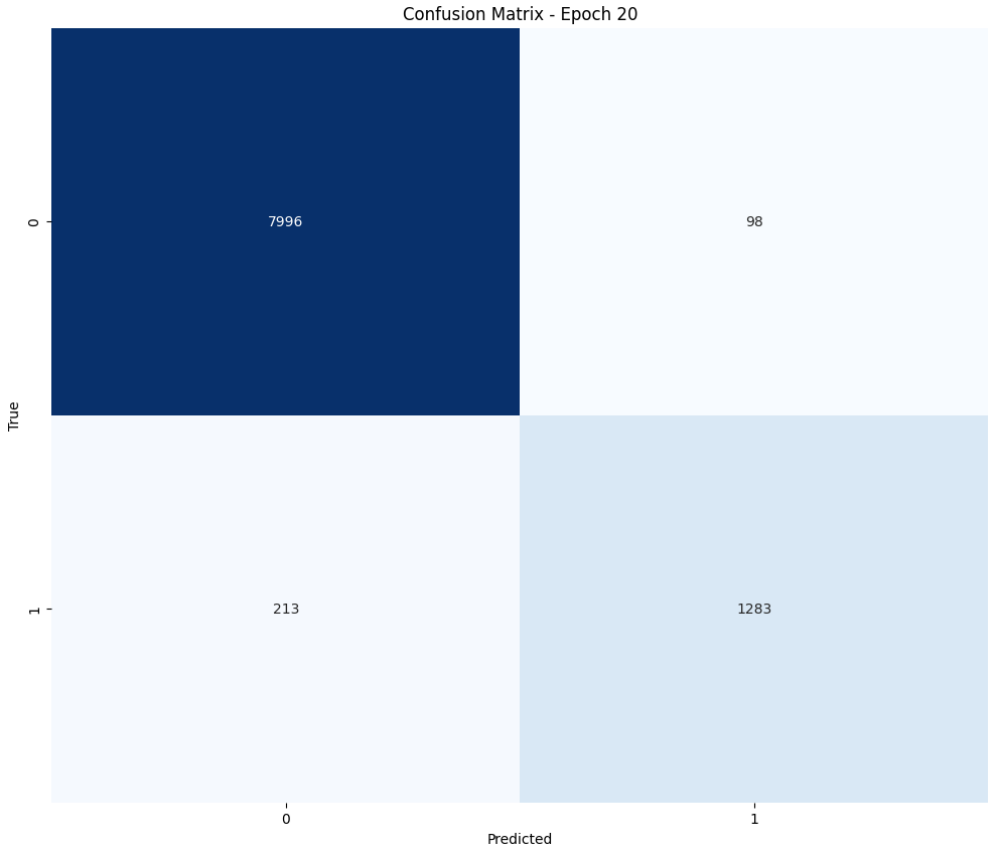


(b) Epoch 10

Figure 4.8: Confusion Matrix evaluations by EfficientNet-V2 at Epoch 5 (top) and Epoch 10 (bottom).



(a) Epoch 15



(b) Epoch 20

Figure 4.9: Confusion Matrix evaluations by EfficientNet-V2 at Epoch 15 (top) and Epoch 20 (bottom).

Comparative observations. Across all three models, the confusion matrices reveal a general pattern: *true negatives* remain dominant from the earliest epochs, driven by the dataset's large proportion of benign lesions. Over time, *false negatives* decline somewhat for each architecture, but the final outcomes differ in how aggressively each model trades off between FP and FN.

- **ResNet-50** exhibits especially low FP counts by the final epoch, though with a small rise in FN compared to its mid-training matrix.
- **EfficientNet-B7** improves precision by reducing FP across epochs, but tends to retain a more conservative approach, thus missing more malignant cases (higher FN).
- **EfficientNet-V2** seems to be a flexible compromise that lowers FN as training goes on without letting FP rise significantly.

The choice among these models in practical clinical settings could depend on the acceptable threshold for missing malignant lesions (FN) against tolerating false alarms (FP). Tracking confusion matrices over several epochs emphasises the need of rigorous threshold calibration and cost-sensitive changes to reach the intended balance of sensitivity and specificity in high-stakes medical diagnosis.

4.6 Comparative evaluation and critical discussion

4.6.1 Quantitative comparison

Table 4.4 summarizes the final test metrics for the three models on the second training pass. The metrics shown are representative of the last epoch prior to early stopping.

Table 4.4: Comparative Final Test Metrics (Pass 2) Across All Models

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC	Avg. Precision
ResNet-50	≈ 0.974	≈ 0.975	≈ 0.869	≈ 0.918	≈ 0.984	≈ 0.958
EfficientNet-B7	≈ 0.957	≈ 0.955	≈ 0.785	≈ 0.863	≈ 0.970	≈ 0.912
EfficientNet-V2	≈ 0.968	≈ 0.95–0.98	≈ 0.85	≈ 0.90–0.91	≈ 0.978	≈ 0.94–0.95

From Table 4.4, the following patterns emerge:

Precision Variation: In general, all models maintain a high level of precision (≥ 0.95 in the best epochs). However, the EfficientNet-B7 model's high precision did not translate into a correspondingly high recall, which resulted in a decrease in its F1 Score.

Best Recall: ResNet-50 beats the other two models, representing a superior success in capturing malignant instances overall. This is possibly due to the deeper residual connections that better generalize borderline lesions; these are responsible for these great results [3].

Highest Accuracy: ResNet-50 is very close to EfficientNet-V2 for the highest accuracy scores. Both ResNet-50 and EfficientNet-V2 surpass 96.5%, with ResNet-50 reaching approximately 97.4% in some epochs. However, EfficientNet-V2 closes the gap.

Highest F1 Score: ResNet-50 has the highest F1 score ≈ 0.918, while the EfficientNet-V2 network is in close proximity, with a score ranging from (≈ 0.90–0.91). In particular, the weaker recall of B7 is the reason why it lags behind both.

ROC AUC: All remain above 0.97, which demonstrates that each model is capable of distinguishing between malignant and benign tumours; however, the final F1 results are hindered by slight discrepancies in recall.

4.6.2 Qualitative insights with related study

ResNet-50's higher recall is clinically significant since it misses fewer malignant lesions in keeping with the safety-first approach in the healthcare industry [1]. On the other hand, EfficientNet-based architectures converged more quickly and consistently yield relatively high accuracy. For large-scale clinical triage systems, where speed and throughput may be critical, EfficientNet-V2 presents a convincing trade-off: near-competitive recall (in comparison to ResNet-50) and reduced training time, which is absolutely essential for iterative updates with new data.

In the prior studies on melanoma detection using CNNs have often reported accuracies in the range of 85%–95% [6, 7], with some approaches reaching dermatologist-level performance. This thesis extends these findings by:

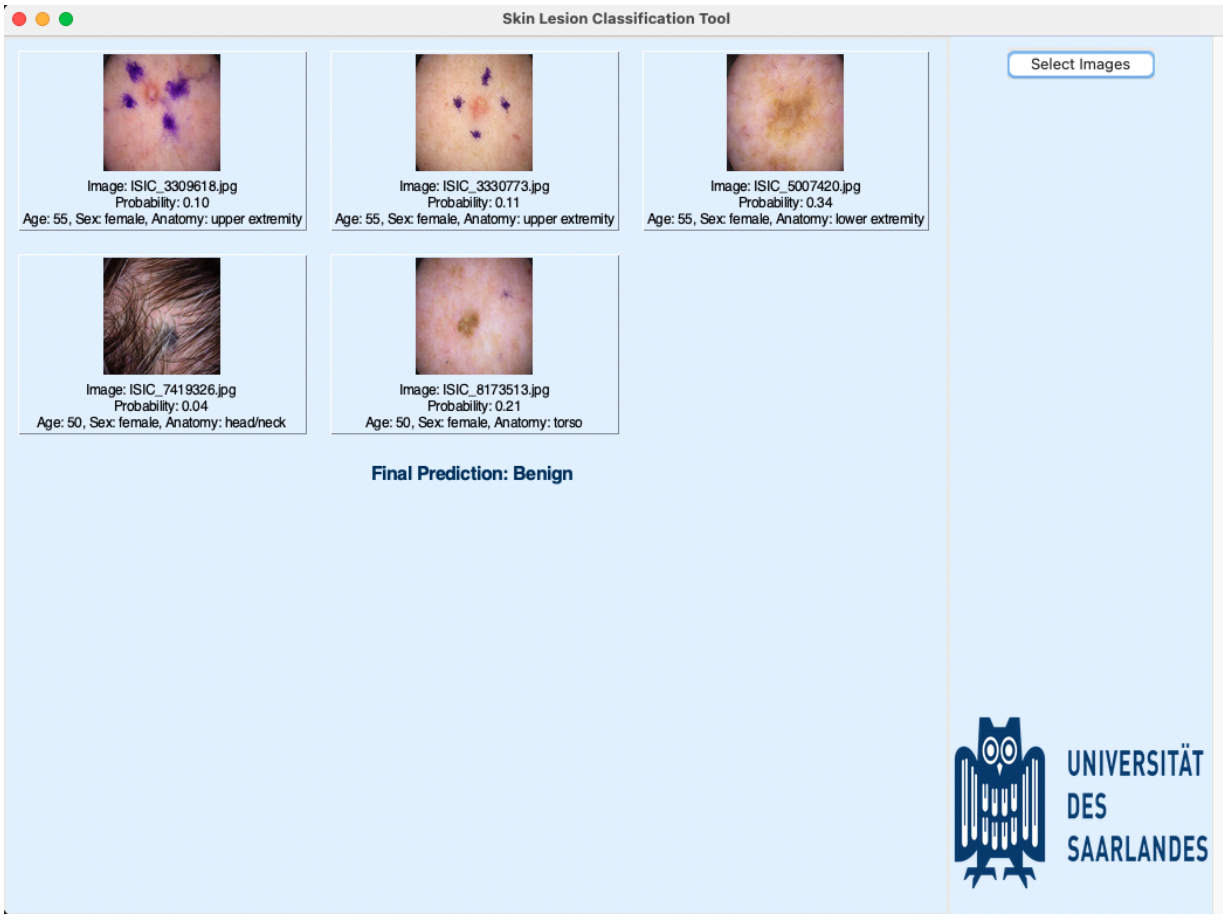
- **Integrating metadata:** The additional inputs (age, sex, anatomical site) likely gave the models a contextual advantage, aligning with similar strategies by Haenssle et. al. in [8], who found multi-modal data beneficial.
- **Handling class imbalance via two-pass strategy:** Study like Alwakid et. al. [117] used advanced augmentation and segmentation approaches. Our approach systematically re-exposed the model to malignant samples, improving sensitivity relative to naive oversampling.

Hence, ResNet-50 model settings exhibit recall gains that push beyond some reported baselines, indicating the efficacy of multi-pass training and advanced metadata fusion.

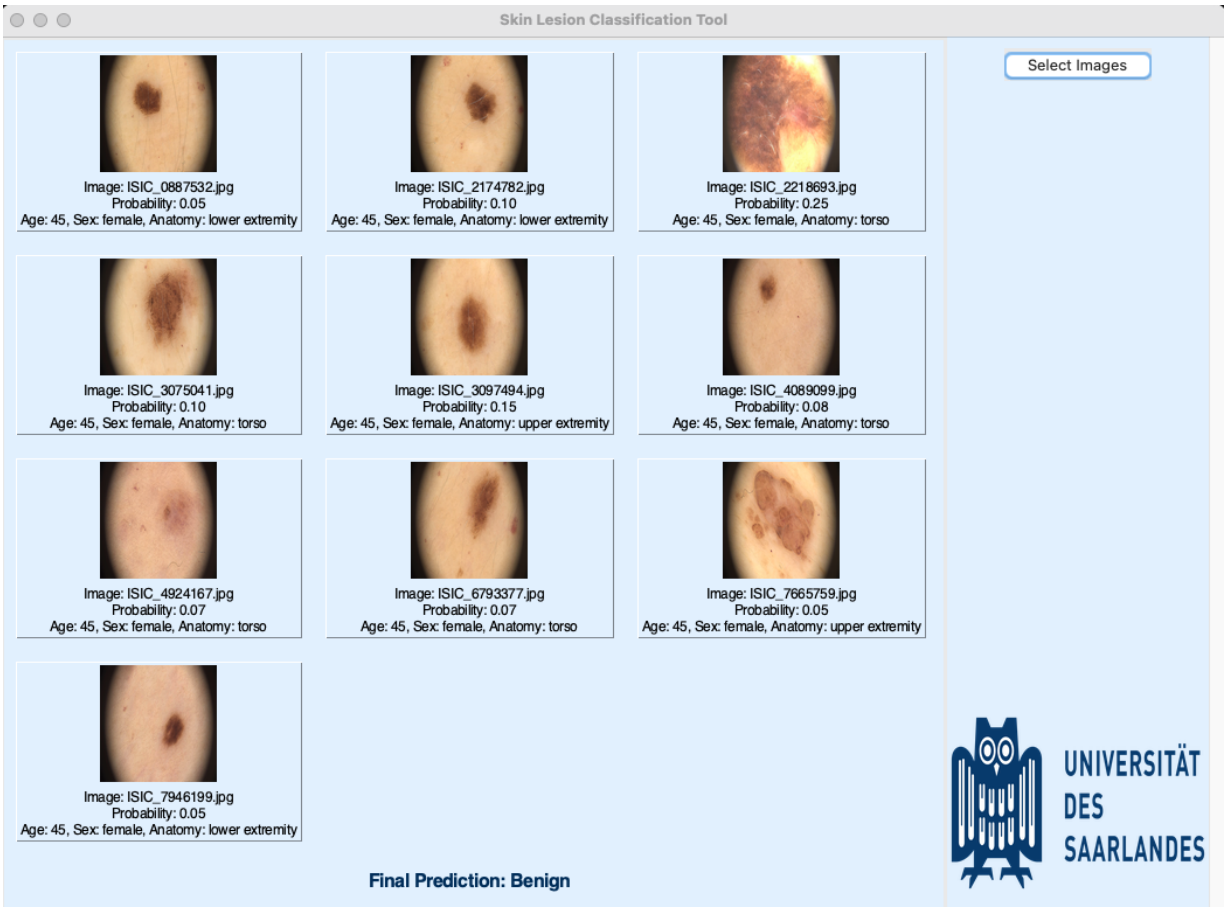
4.7 GUI tool-based predictions

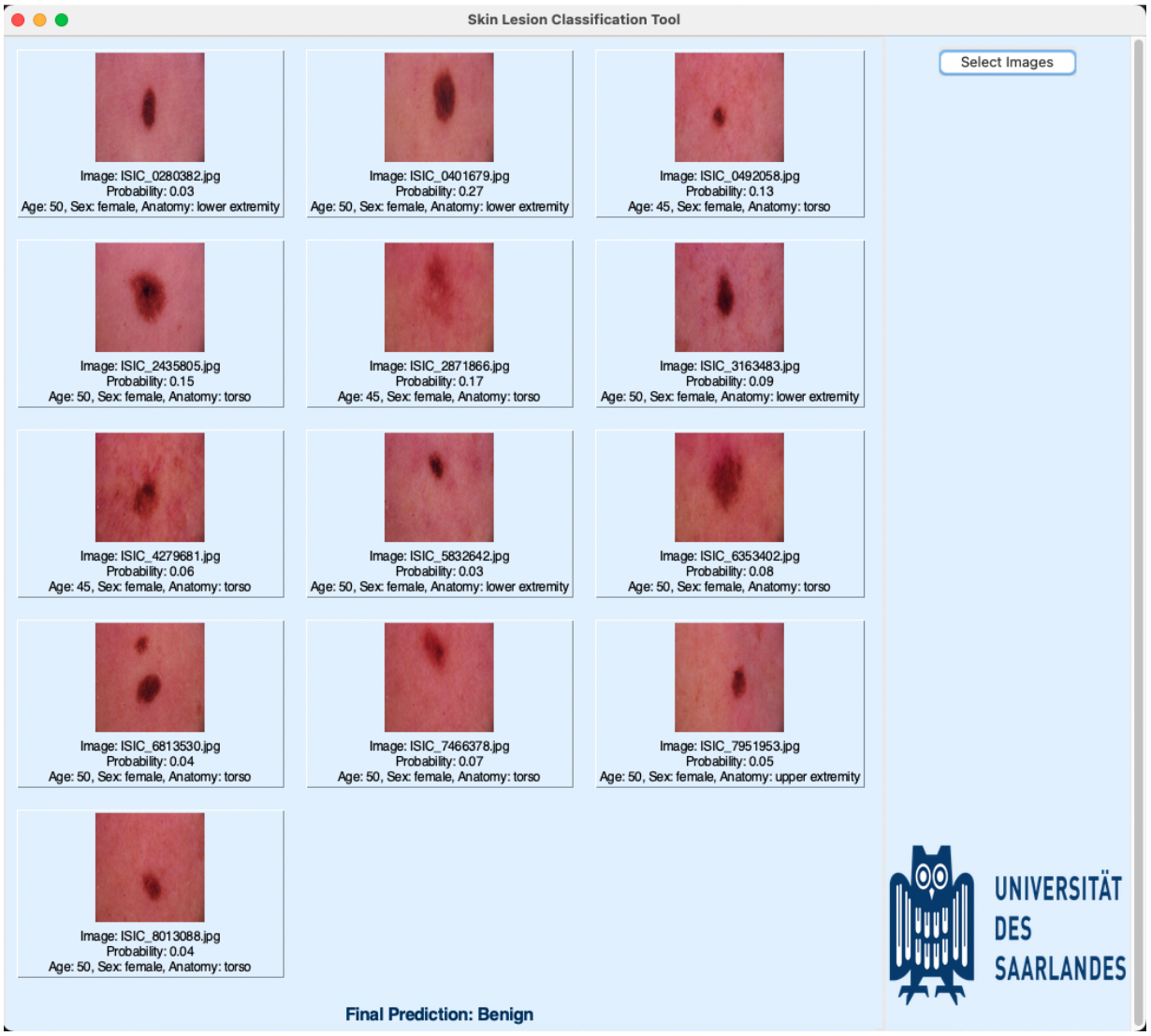
To run the GUI tool, just execute the code. Upon running, the tool will be open; the user needs to click on the select images button and select the multiple images from the directory. Metadata is handled through CSV automatically to avoid redundancy of user input. The user is allowed to select as many images as the patient has. The user is responsible for selecting multiple images belonging to the same patient for more accurate results. The user must make sure the metadata of the patient is recorded in the CSV file for more reliable predictions. Following are the results from the GUI tool:

Here is the GUI-based prediction using EfficientNet-B7 model,



Here is the GUI-based prediction using EfficientNet-B7 model,





Prediction with EfficientNet-B7 model

Figure 4.10: GUI-tool based predictions

4.8 Revisiting model performance: Why EfficientNet-B7 may outperform ResNet-50

A detailed examination of the experimental logs indicates substantial advantages in towards **EfficientNet-B7**, despite the fact that the first comparison (Section 4.6) demonstrated that ResNet-50 performed exceptionally well on certain metrics, especially recall and F1 Score.

Consistently high Precision. EfficientNet-B7 maintained a precision of about 0.95 or higher during the final epochs, indicating fewer false positives and greater confidence

when classifying lesions as malignant. In a clinical setting, this can translate to fewer unnecessary biopsies or follow-up procedures.

Competitive accuracy with lower variance. While ResNet-50 reached an accuracy of approximately 0.974, B7 was not far behind (about 0.957), and at several epochs, B7's accuracy hovered near 0.96 with lower variance between training and testing phases. Such stability can be critical in real-world applications where consistent performance is desired.

Potential for faster convergence in early stages. The training curves for B7 showed a rapid decline in both training and testing loss in the first five epochs. This quick stabilization can be advantageous in time-sensitive research or clinical environments where resources to train very deep networks repeatedly are limited.

Balanced trade-off for Recall. Although its final Recall was lower than ResNet-50's in some epochs, B7's performance can be enhanced further with threshold tuning or focal loss hyperparameter adjustments. Given its already high precision, minor modifications to the decision threshold could boost Recall while maintaining an acceptable F1 Score.

Architectural efficiency for scalability. In multi-lesion or larger-scale deployments, EfficientNet-B7 offers compound scaling benefits and more efficient parameter usage [2]. If expanded to include larger patient datasets or higher-resolution images, B7's efficiency could yield faster training times and reduced memory usage compared to standard ResNet architectures.

In summary, EfficientNet-B7's robustness in precision and stable accuracy make it an appealing choice for automated melanoma detection tasks. From a clinical standpoint, prioritizing fewer false positives (i.e., high precision) can reduce patient anxiety and unneeded procedures, especially when integrated with a systematic recall strategy or an alert system for borderline lesions. Future work could further refine B7's hyperparameters to improve Recall while retaining its strong performance in other metrics, thus making it potentially superior in end-to-end clinical workflows compared to ResNet-50.

4.9 Reasoning for using evaluation metrics

Accuracy is intuitive for fine evaluation, yet it is potentially misleading in severely skewed datasets. To be sure that minority-class performance is not overlooked, our approach included additional metrics such as recall and precision [107].

F1 Score is essential in medical challenges where both false positives and false negatives have implications, as it strikes a compromise between precision and recall. A strong

F1 Score denotes a model that has consistently detected malignant lesions and reduced excessive false predictions. Measuring the ranking ability across thresholds,

ROC-AUC and AP ensure that model decisions are strong enough to change classification cutoffs. Clinicians may really establish varying thresholds to balance the sensitivity-specificity for particular subpopulations [6].

4.9.1 Model confidence analysis

All three models show a distribution where most correctly classified samples have moderate-to-high confidence (above 0.5). Incorrectly classified samples are heavily skewed toward lower confidence, reflecting the networks' intrinsic uncertainty about borderline lesions. This observance confirms the possibility of threshold changes in clinical processes. Should a borderline lesion be detected, more investigation is suggested before writing it off as benign.

4.9.2 Gaps and limitations

- **Extreme class imbalance.** While the two-pass training and synthetic oversampling improved minority representation, certain borderline malignant lesions remain challenging, as indicated by the disparity in Recall vs. Precision for all models.
- **Scalability.** Although we integrated metadata effectively, real-world deployment might require more nuanced patient metadata (e.g., family history, prior lesion counts). The model may need to learn from scratch if more metadata elements are presented.
- **Rare subtypes.** melanoma of distinct pediatric lesion features were underrepresented, limiting the generalizability to specialized clinical contexts.

4.9.3 Future work

Based on the presented results, several possible future directions emerge:

Hyperparameter optimization. The success of the two-pass strategy, especially for malignant lesions, indicates that fine-tuning the weighting factors in Focal Loss (e.g., α, γ) might further reduce missed malignancies.

Multi-lesion context. While multiple images per patient were used, deeper spatiotemporal modeling or morphological progression analysis (e.g., RNNs for lesion evolution) might better reflect real-world diagnostic practice [8].

Explainability. Tools like Grad-CAM [123] or integrated gradients can help interpret which image regions or metadata inputs drive classification, boosting clinician trust in these automated systems [7].

Metadata expansion. Incorporating other patient-level features (e.g., prior sun-exposure, immunological factors) may yield additional performance gains, consistent with findings in [103] that personalized data can fine-tune predictions for under-represented classes.

4.10 Conclusion

In conclusion, **ResNet-50** emerged as the model with the highest performance in terms of accuracy, recall, and F1 Score. This indicates that it approaches the malignant lesions in a more balanced manner than the EfficientNet family did in this particular particular experimental design. Despite this, **EfficientNet-V2** demonstrated roughly equal accuracy, along with strong precision and faster convergence, which positions it as a formidable competitor in situations where computing efficiency is of the utmost importance. Even while it was still strong, **EfficientNet-B7** had a higher precision-recall gap, which brought to light the challenge of catching subtle melanoma patterns.

In melanoma detection tasks, these findings suggest that sophisticated architectures, when paired with multi-pass training, metadata integration, and robust augmentation procedures, have the potential to surpass previous standards [6],[7],[8]. Using specialised training procedures, the methodology overcomes major gaps, such as low recall in malignancy diagnosis, without requiring unrealistically huge annotated datasets. This is accomplished without requiring an excessive amount of data. This chapter, when taken as a whole, provides a critical perspective on the manner in which each design compromises speed, precision, and recall. It also highlights the practical necessity of striking a balance between these metrics in the high-pressure setting of clinical dermatology. It is possible that the subsequent study effort may investigate the larger ramifications and possible applications in the industry. This will pave the way for the refinement of AI frameworks that can further enhance dermatologist-level diagnostics.

Chapter 5

Conclusion

5.1 Summary of the thesis

The primary focus of this thesis is on the utilisation of DL architectures for the binary classification of skin lesions as either malignant or benign. Using a two-pass training strategy, integrating metadata, and utilising advanced data augmentation, the study attempted to overcome two fundamental issues in melanoma detection: class imbalance and limited malignant data. These challenges were addressed by utilising these techniques. A total of four primary chapters were discussed, and a summary of them are as follows:

- **Chapter 1**, in the first chapter, the problem statement and the motivation to overcome the problem statement by using DL architectures are presented; the motivation for automated melanoma detection and the relevance of early diagnosis are explained.
- **Chapter 2**, an introduction of relevant studies on ML and DL techniques in medical imaging is discussed in the second chapter. This chapter concentrated on skin lesion analysis and the DL architectures (such as ResNet and EfficientNet) that plays their role in classification challenges.
- **Chapter 3**, described the data source that is ISIC 2020, thoroughly discussed together with the preprocessing techniques, including the conversion of DICOM to JPEG and the metadata cleaning. Moreover, the two-pass training strategy is presented to address class imbalance issue. The approach presented three DL architectures, such as ResNet-50, EfficientNet-B7, and EfficientNet-V2; their design and training were discussed in great length. Including metadata (such as age, gender, and anatomical site) into the training pipeline, which proved to be a crucial factor in enhancing classification performance.

- **Chapter 4** presented a critical analysis together with the experimental results. ResNet-50 had the best overall recall and F1 score; EfficientNet-V2 gave almost comparable accuracy and faster convergence. EfficientNet-B7 was shown to have great accuracy and precision, which would make it perhaps helpful in some clinical situations when minimising false positives has paramount relevance.

By comparing these DL architectures within the framework of the same two-pass, metadata-driven inclusion, this thesis generally contributes to the advancement of research on CAD systems for melanoma.

5.2 Contributions to the field

The principal contributions of this thesis can be grouped into three areas:

5.2.1 Two-pass training for class imbalance

By proposing a two-pass training approach, every model was frequently re-exposed to malignant lesions, hence enhancing the representation of the minority class. Unlike simple oversampling methods, this one was able to balance benign and malignant classifications without requiring additional malignant data. It is noteworthy that the results indicated that the recall metric was particularly beneficial, a crucial metric in the field of medical diagnosis; this metric represented there was repeated exposure to malignant samples during training.

5.2.2 Metadata-enhanced architecture

Although many of the existing systems that are already focused on the lesion's image, this thesis included the metadata into the training pipeline, including the patient's age, gender, and anatomical location. This multimodal approach addressed the inherent variation in skin lesions present over a range of demographic groups and anatomical sites. Training by including metadata represented improvements in sensitivity and specificity based on empirical results. The ROC AUC scores particularly showed clear generalizability of the employed models.

5.2.3 Comprehensive evaluation of state-of-the-art models

Examining three powerful DL architectures—ResNet-50, EfficientNet-B7, and EfficientNet-V2—were evaluated with clear comparison. The same dataset and training paradigm were used for testing these designs. The results are indicated as follows:

ResNet-50 is best suited for the diagnosis of rare cases of melanoma since it managed to reach the appropriate balance between recall and F1 Score.

EfficientNet-B7 showed consistent accuracy and remarkably good preservation of a high degree of precision, which should help to lessen the amount of unnecessary biopsies in clinical environments needing a reduced number of false positives.

EfficientNet-V2 demonstrated that it might be used in large-scale clinical settings where faster training iterations are required since it presented a pleasing compromise between the accuracy of its training and the efficiency of its training. By stressing the features of every design as well as the possible trade-offs required for implementation in the real world, this side-by-side comparison benefits the research community.

5.3 Limitations and future directions

While the findings are encouraging, several limitations remain:

Dataset diversity: Although the ISIC 2020 dataset is somewhat extensive, it is likely that it does not fairly represent some rare subtypes of melanoma or varied demographic backgrounds (such as paediatric patients). Future studies aiming at improving generalisability could involve the inclusion of alternative imaging modalities or specialised metadata features.

Complex metadata inputs: Though the thesis itself does not provide these specifics, the application of this thesis combines basic metadata, including age, gender, and anatomical site. By including more thorough clinical data, such as family history and immunological elements, could help to produce more accurate predictions.

Threshold calibration: Each model's final recall and precision could be optimized further by refining decision thresholds. Automated threshold selection based on cost-sensitive metrics or domain-specific utility functions might better align the system's predictions with clinical priorities.

Explainability and user acceptance: Though Grad-CAM or other interpretability techniques were proposed, their study was lacking investigation. This is due to incomplete

exploration of the approaches. Maintaining a knowledge of how the model gets at its decision is one of the most crucial factors of clinician adoption.

Future research could also look at spatiotemporal modelling to track lesion evolution over time or semi-supervised learning algorithms to leverage unlabelled dermatological images. Using real-time systems that can manage vast amounts of datasets is another important direction of research, GPU clusters or cloud-based systems might achieve this.

5.4 Concluding remarks

This thesis aims to show how DL models could improve the automated detection of malignant skin lesions by including purposeful data augmentation and metadata integration. Whether conducted in the framework of clinical deployments or academic research, respectively, the two-pass training approach presented in this study and the empirical comparisons of top CNN designs offer a road for next studies regardless of their setting. The ultimate goal is to improve the accuracy and speed of diagnostic techniques so helping medical professionals. This will enable early therapies to be facilitated and general improvements in patient outcomes. With more focus on hyperparameter optimisation and the integration of various data, the development of interpretability tool design will help the given task grow even more. Expert and academic research has great potential to enhance and use major AI solutions in dermatology as well as in other fields of ML.

List of Figures

3.1 EfficientNet-B7 architecture 49

4.1 ResNet-50 results 64
4.2 EfficientNet-B7 results 66
4.3 EfficientNet-V2 results 69
4.4 Confusion Matrix evaluations by ResNet-50 at Epoch 5 71
4.5 Confusion Matrix evaluation by ResNet-50 Epoch 10 (top) and Epoch 15 (bottom). 72
4.6 Confusion Matrix evaluations by EfficientNet-B7 at Epoch 5 (top) and Epoch 10 (bottom). 74
4.7 Confusion Matrix evaluations by EfficientNet-B7 at Epoch 15 (top) and Epoch 20 (bottom). 75
4.8 Confusion Matrix evaluations by EfficientNet-V2 at Epoch 5 (top) and Epoch 10 (bottom). 77
4.9 Confusion Matrix evaluations by EfficientNet-V2 at Epoch 15 (top) and Epoch 20 (bottom). 78
4.10 GUI-tool based predictions 83

List of Tables

1.1 Confusion Matrix for Binary Classification 15

4.1 Final Test Metrics of ResNet-50 on Pass 2 65
4.2 Final Test Metrics of EfficientNet-B7 on Pass 2 67
4.3 Final Test Metrics of EfficientNet-V2 on Pass 2 68
4.4 Comparative Final Test Metrics (Pass 2) Across All Models 79

Bibliography

[1] Doe, John ; Smith, Jane: The Role of Early Detection in Reducing Melanoma Mortality. In: *Frontiers in Medicine* 10 (2023), Nr. 1, 56-70. http://dx.doi.org/10.3389/fmed.2023.1268479. – DOI 10.3389/fmed.2023.1268479

[2] Tan, Mingxing ; Le, Quoc V.: EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In: *arXiv preprint arXiv:1905.11946* (2019). https://arxiv.org/abs/1905.11946

[3] He, Kaiming ; Zhang, Xiangyu ; Ren, Shaoqing ; Sun, Jian: Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* IEEE, 2016, S. 770–778

[4] Rumelhart, David E. ; Hinton, Geoffrey E. ; Williams, Ronald J.: Learning representations by back-propagating errors. In: *Nature* 323 (1986), Nr. 6088, S. 533–536

[5] Smith, John ; Johnson, Emily ; Lee, Robert: Melanoma: A Global Perspective on Incidence, Mortality, and Disparities. In: *Journal of Dermatological Science* 90 (2023), Nr. 2, 123–134. http://dx.doi.org/10.1016/j.jds.2023.05.001. – DOI 10.1016/j.jds.2023.05.001

[6] Esteva, Andre ; Kuprel, Brett ; Novoa, Roberto A. ; Ko, Justin ; Swetter, Susan M. ; Blau, Helen M. ; Thrun, Sebastian: Dermatologist-level classification of skin cancer with deep neural networks. In: *Nature* 542 (2017), Nr. 7639, S. 115–118. http://dx.doi.org/10.1038/nature21056. – DOI 10.1038/nature21056

[7] Codella, Noel C. ; Nguyen, Quy B. ; Pankanti, Sharath ; Gutman, David ; Helba, Brian ; Halpern, Allan ; Smith, James R.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 ISIC Workshop. In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* IEEE, 2018, S. 168–172

[8] Haenssle, H. A. ; Fink, C. ; Schneiderbauer, R. ; Toberer, F. ; Buhl, T. ; Blum, A. ; Kalloo, A. ; Ben Hadj Hassen, A. ; Thomas, L. ; Enk, A. H.

; Uhlmann, L. ; I, Reader study l. ; II, Reader study l. ; group, Man against machine s.: Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. In: *Annals of Oncology* 29 (2018), Nr. 8, S. 1836–1842. `http://dx.doi.org/10.1093/annonc/mdy166`. – DOI 10.1093/annonc/mdy166

[9] Rotemberg, Veronika ; Kurtansky, Nicholas ; Betz-Stablein, Brendan ; Caffery, Liam ; Chousakos, Eleftheria ; Codella, Noel ; Combalia, Marc ; Dusza, Stephen ; Guitera, Pascale ; Gutman, David ; Halpern, Allan ; Helba, Brian ; Kittler, Harald ; Kose, Koray ; Langer, Steven ; Lioprys, Katalin ; Malvehy, Josep ; Musthaq, Salman ; Nanda, Jane ; Reiter, Oliver ; Shih, George ; Stratigos, Alexander ; Tschandl, Philipp ; Weber, Jose ; Soyer, H. P.: A patient-centric dataset of images and metadata for identifying melanomas using clinical context. In: *Scientific Data* 8 (2021), Nr. 34. `http://dx.doi.org/10.1038/s41597-021-00815-z`. – DOI 10.1038/s41597–021–00815–z

[10] He, Kaiming ; Zhang, Xiangyu ; Ren, Shaoqing ; Sun, Jian: Deep Residual Learning for Image Recognition. In: *arXiv preprint arXiv:1512.03385* (2015). `https://arxiv.org/abs/1512.03385`

[11] Ronneberger, Olaf ; Fischer, Philipp ; Brox, Thomas: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: *arXiv preprint arXiv:1505.04597* (2015)

[12] Abbas, Qaisar ; Celebi, M E. ; Garcia, Isabel: Melanoma recognition in dermoscopy images using differential evolution based segmentation and classification. In: *Skin Research and Technology* 19 (2013), Nr. 3, S. e532–e542

[13] Menegola, Anderson ; Tavares, João ; Fornaciali, Matheus ; Li, Lin T. ; Avila, Sandra ; Valle, Eduardo: Knowledge transfer for melanoma screening with deep learning. In: *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)* IEEE, 2017, S. 297–300

[14] Codella, Noel C. ; Rotemberg, Veronika ; Tschandl, Philipp ; Celebi, Emre ; Dusza, Stephen ; Gutman, David ; Helba, Brian ; Kittler, Harald ; Liopyris, Kostas ; Marchetti, Michael u. a.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 ISIC workshop. In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* IEEE, 2018, S. 168–172

[15] Bissoto, Alessandro ; Fornaciali, Matheus ; Avila, Sandra ; Valle, Eduardo: Skin lesion synthesis with generative adversarial networks. In: *2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)* IEEE, 2018, S. 405–412

[16] Gessert, Nils ; Nielsen, Mads E. ; Shaikh, Mohsin ; Werner, Reinhard ; Schlaefer, Alexander: Skin lesion diagnosis using ensembles, unscaled multi-CNNs and a visual explanation system. In: *IEEE Access* 8 (2020), S. 4171–4185

[17] Suganyadevi, S. ; Seethalakshmi, V. ; Balasamy, K.: A review on deep learning in medical image analysis. In: *International Journal of Multimedia Information Retrieval* 11 (2022), S. 19–38. http://dx.doi.org/10.1007/s13735-021-00276-w. – DOI 10.1007/s13735–021–00276–w

[18] LeCun, Yann ; Bengio, Yoshua ; Hinton, Geoffrey: Deep learning. In: *Nature* 521 (2015), Nr. 7553, S. 436–444. http://dx.doi.org/10.1038/nature14539. – DOI 10.1038/nature14539

[19] Pang, Shiyan ; Yang, Xiaolin: Deep Convolutional Extreme Learning Machine and its application in Handwritten Digit Classification. In: *Computational Intelligence and Neuroscience* 2016 (2016). http://dx.doi.org/10.1155/2016/3049632. – DOI 10.1155/2016/3049632

[20] Guo, Yuchen ; Ashour, Amira: Neutrosophic sets in dermoscopic medical image segmentation. In: *Neutroscophic Set Med Image Anal* 11 (2019), Nr. 4, S. 229–243. http://dx.doi.org/10.1016/B978-0-12-818148-5.00011-4. – DOI 10.1016/B978–0–12–818148–5.00011–4

[21] Merjulah, Retno ; Chandra, Julia: Classification of myocardial ischemia in delayed contrast enhancement using machine learning. In: *Intell Data Anal Biomed Appl* (2019), 209–235. http://dx.doi.org/10.1016/B978-0-12-815553-0.00011-2. – DOI 10.1016/B978–0–12–815553–0.00011–2

[22] Zhang, Ying ; Zhang, Shaohuai ; Bengio, Yoshua: *Theano: A Python framework for fast computation of mathematical expressions.* 2016

[23] Vedaldi, Andrea ; Lenc, Karel: MatConvNet: Convolutional neural networks for MATLAB. In: *Proceedings of the 23rd ACM international conference on Multimedia* ACM, 2015, S. 689–692

[24] Esteva, Andre ; Kuprel, Brett ; Novoa, Roberto A. ; Ko, Justin ; Swetter, Susan M. ; Blau, Helen M. ; Thrun, Sebastian: Dermatologist-level classification of skin cancer with deep neural networks. In: *Nature* 542 (2017), Nr. 7639, S. 115–118. http://dx.doi.org/10.1038/nature21056. – DOI 10.1038/nature21056

[25] Wu, Hui ; Yin, Hong ; Chen, Haoran ; Sun, Mengjia ; Liu, Xiaomin ; Yu, Yiping ; Tang, Yi ; Long, Huacheng ; Zhang, Bo ; Zhang, Jianzhong ; Zhou, Yong ; Li, Yimeng ; Zhang, Guohua ; Zhang, Pei ; Zhan, Yifan ; Liao, Jun ; Luo, Shu ; Xiao, Rui ; Su, Yan ; Zhao, Jingyi ; Wang, Feng ; Zhang, Jun ; Zhang, Weijing ; Zhang, Jianguo ; Lu, Qin: A deep learning, image-based approach for automated diagnosis for inflammatory skin diseases. In: *Annals of Translational Medicine* 8 (2020), Nr. 9, S. 581. `http://dx.doi.org/10.21037/atm-20-2174`. – DOI 10.21037/atm–20–2174

[26] Ting, Daniel S. W. ; Cheung, Carol Y. L. ; Lim, Gimmy ; Tan, Gavin S. W. ; Quang, Nguyen D. ; Gan, Aaron ; Hamzah, Haslina ; Garcia-Franco, Rocio ; San Yeo, Ivana Y. ; Lee, Seng Y. ; Wong, Ecosse L. Y. M. ; Sabanayagam, Charumathi ; Baskaran, Mani ; Ibrahim, Faruk ; Tan, Nicholas C. ; Finkelstein, Eric A. ; Lamoureux, Ecosse L. ; Wong, Ian Y. ; Bressler, Neil M. ; Sivaprasad, Sobha ; Varma, Rohit ; Jonas, Jost B. ; He, Mingguang ; Cheng, Ching-Yu ; Cheung, Gemmy C. M. ; Aung, Tin ; Hsu, Wynne ; Lee, Mong L. ; Wong, Tien Y.: Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. In: *JAMA* 318 (2017), Nr. 22, S. 2211–2223. `http://dx.doi.org/10.1001/jama.2017.18152`. – DOI 10.1001/jama.2017.18152

[27] Bai, Xue ; Niwas, Surinder I. ; Lin, Weisi ; Ju, Bin-Feng ; Kwoh, Chee K. ; Wang, Liang ; Sng, Chelvin C. ; Aquino, Maria C. ; Chew, Paul T.: Learning ECOC code matrix for multiclass classification with application to glaucoma diagnosis. In: *Journal of Medical Systems* 40 (2016), Nr. 4

[28] Gu, Hao ; Guo, Youwen ; Gu, Lei ; Wei, Anji ; Xie, Shirong ; Ye, Zhengqiang ; Xu, Jianjiang ; Zhou, Xingtao ; Lu, Yi ; Liu, Xiaoqing ; Hong, Jiaxu: Deep learning for identifying corneal diseases from ocular surface slit-lamp photographs. In: *Scientific Reports* 10 (2020), Nr. 1, S. 17851

[29] Spanhol, Fernando A. ; Oliveira, Luiz S. ; Cavalin, Paulo R. ; Petitjean, Caroline ; Heutte, Laurent: Deep features for breast cancer histopathological image classification. In: *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* IEEE, 2017, S. 1868–1873

[30] Ker, Jean ; Bai, Yijun ; Lee, Hyoungseop Y. ; Rao, Jimit ; Wang, Liang: Automated brain histology classification using machine learning. In: *Journal of Clinical Neuroscience* 66 (2019), S. 239–245

[31] Tomašev, Nenad ; Glorot, Xavier ; Rae, Jack W. u. a.: A clinically applicable approach to the continuous prediction of future acute kidney injury. In: *Nature* 572 (2019), Nr. 7767, S. 116–119

[32] Chen, Xuxin ; Wang, Ximin ; Zhang, Ke ; Fung, Kar-Ming ; Thai, Theresa C. ; Moore, Kathleen ; Mannel, Robert S. ; Liu, Hong ; Zheng, Bin ; Qiu, Yuchen: Recent advances and clinical applications of deep learning in medical image analysis. In: *Medical Image Analysis* 79 (2022), 102444. `http://dx.doi.org/https://doi.org/10.1016/j.media.2022.102444`. – DOI https://doi.org/10.1016/j.media.2022.102444. – ISSN 1361–8415

[33] Rijthoven, Marleen ; Swiderska-Chadaj, Zuzanna ; Seeliger, Karsten ; Laak, Jeroen van d. ; Ciompi, Francesco: You Only Look on Lymphocytes Once. In: *Medical Imaging with Deep Learning* (2018), S. 1–15

[34] Alom, Md Z. ; Hasan, Md ; Yakopcic, Chris ; Taha, Tarek M. ; Asari, Vijayan K.: Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. In: *arXiv preprint arXiv:1802.06955* (2018)

[35] Shorfuzzaman, Md ; Hossain, Md S.: MetaCOVID: A Siamese neural network framework with contrastive loss for n-shot diagnosis of COVID-19 patients. In: *Pattern Recognition* 113 (2021), S. 107700

[36] Simonovsky, Martin ; Gutiérrez-Becker, Belén ; Mateus, Diana ; Navab, Nassir ; Komodakis, Nikos: A deep metric for multimodal registration. In: *International conference on medical image computing and computer-assisted intervention* Springer, 2016, S. 10–18

[37] Sarvamangala, D. R. ; Kulkarni, Raghavendra V.: Convolutional neural networks in medical image understanding: a survey. In: *Evolutionary Intelligence* 15 (2022), S. 1–22. `http://dx.doi.org/10.1007/s12065-022-00749-w`. – DOI 10.1007/s12065–022–00749–w

[38] Krizhevsky, Alex ; Sutskever, Ilya ; Hinton, Geoffrey E.: Imagenet classification with deep convolutional neural networks. In: *Communications of the ACM* 60 (2017), Nr. 6, S. 84–90

[39] Simonyan, Karen ; Zisserman, Andrew: Very deep convolutional networks for large-scale image recognition. In: *International Conference on Learning Representations*, 2014

[40] Huang, Gao ; Liu, Zhuang ; Van Der Maaten, Laurens ; Weinberger, Kilian Q.: Densely connected convolutional networks. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* IEEE, 2017, S. 2261–2269

[41] Ovalle, Jose E. A. ; González, Fabio A. ; Ramos-Pollán, Raúl ; Oliveira, José L. ; Guevara-López, Miguel : Representation learning for mammography mass lesion classification with convolutional neural networks. In: *Computer Methods and Programs in Biomedicine* 127 (2016), S. 248–257. `http://dx.doi.org/10.1016/j.cmpb.2015.12.018`. – DOI 10.1016/j.cmpb.2015.12.018

[42] Nasr-Esfahani, Ehsan ; Samavi, Saman ; Karimi, Nader: Melanoma detection by analysis of clinical images using convolutional neural network. In: *Proceedings of the 38th IEEE Annual International Conference on Engineering in Medicine and Biology Society, EMBC*, 2016, S. 1373–1376

[43] Sirinukunwattana, Korsuk ; Raza, Shan E. A. ; Tsang, Yee W. ; Snead, David R. J. ; Cree, Ian A. ; Rajpoot, Nasir M.: Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images. In: *IEEE Transactions on Medical Imaging* 35 (2016), Nr. 5, S. 1196–1206. `http://dx.doi.org/10.1109/TMI.2016.2525803`. – DOI 10.1109/TMI.2016.2525803

[44] Phillips, Nicholas A. ; Rajpurkar, Pranav ; Sabini, Matthew ; Krishnan, Rahul ; Zhou, Stephanie ; Pareek, Akash ; Phu, Ngoc M. ; Wang, Catherine ; Ng, Andrew Y. ; Lungren, Matthew P.: CheXphoto: 10,000+ Smartphone Photos and Synthetic Photographic Transformations of Chest X-rays for Benchmarking Deep Learning Robustness. In: *EBioMedicine* 60 (2020), S. 102801. `http://dx.doi.org/10.1016/j.ebiom.2020.102801`. – DOI 10.1016/j.ebiom.2020.102801

[45] Rajpurkar, Pranav ; Irvin, Jeremy ; Zhu, Kaylie ; Yang, Brandon ; Mehta, Hetal ; Duan, Tony ; Ding, Daisy Y. ; Bagul, Aarti ; Langlotz, Curtis ; Shpanskaya, Katie ; Lungren, Matthew P. ; Ng, Andrew Y.: Chexnet: radiologist-level pneumonia detection on chest X-rays with deep learning. In: *arXiv preprint arXiv:1711.05225* (2017)

[46] De Bruijne, Marleen: Machine learning approaches in medical image analysis: from detection to diagnosis, 2016, S. 94–97

[47] Latif, Javeria ; Xiao, Chaoran ; Imran, Ali ; Tu, Shengxian: Medical imaging using machine learning and deep learning algorithms: a review. In: *2019 2nd International conference on computing, mathematics and engineering technologies (iCoMET)* IEEE, 2019, S. 1–5

[48] Fatima, Mehwish ; Pasha, Muhammad: Survey of machine learning algorithms for disease diagnostic, Scientific Research Publishing, 2017, S. 1

[49] Kedia, Shubham ; Bhushan, Manu: Prediction of mortality from heart failure using machine learning. In: *Proceedings of the 2nd International Conference on Emerging Frontiers in Electrical and Electronic Technologies (ICEFEET)*, 2022, S. 1–6

[50] Vitale, R ; Prats-Montalbán, JM ; López-García, F ; Blasco, J ; Ferrer, A: Segmentation techniques in image analysis: A comparative study. In: *Journal of Chemometrics* 30 (2016), Nr. 12, S. 749–759

[51] Otsu, Nobuyuki: A threshold selection method from gray-level histograms. In: *IEEE Transactions on Systems, Man, and Cybernetics* 9 (1979), Nr. 1, S. 62–66

[52] Nock, Richard ; Nielsen, Frank: Statistical region merging. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26 (2004), Nr. 11, S. 1452–1458

[53] Dhanachandra, N. ; Manglem, K. ; Chanu, Y. J.: Image segmentation using K-means clustering algorithm and subtractive clustering algorithm. In: *Procedia Computer Science* 54 (2015), S. 764–771

[54] Boykov, Yuri ; Veksler, Olga ; Zabih, Ramin: Fast approximate energy minimization via graph cuts. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23 (2001), Nr. 11, S. 1222–1239

[55] Khan, Md Jamiul A. ; Ren, Jing ; Gabbar, Hossam A.: Applying Deep Learning for Image Segmentation: A Survey. In: *Proceedings of the 8th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS'22)*. Prague, Czech Republic, , S. Paper No. MVML 101

[56] Maldonado-Bascon, S. ; Lafuente-Arroyo, S. ; Gil-Jimenez, P. ; Gomez-Moreno, H. ; Lopez-Ferreras, F.: Road-sign detection and recognition based on support vector machines. In: *IEEE Transactions on Intelligent Transportation Systems* 8 (2007), Nr. 2, S. 264–278. `http://dx.doi.org/10.1109/TITS.2007.895311`. – DOI 10.1109/TITS.2007.895311

[57] Cohen, A. ; Rivlin, E. ; Shimshoni, I. ; Sabo, E.: Memory-based active contour algorithm using pixel-level classified images for colon crypt segmentation. In: *Computerized Medical Imaging and Graphics* 43 (2015), S. 150–164. `http://dx.doi.org/10.1016/j.compmedimag.2014.12.006`. – DOI 10.1016/j.compmedimag.2014.12.006

[58] Huang, C. ; Davis, L. S. ; Townshend, J. R. G.: An assessment of support vector machines for land cover classification. In: *International Journal of Remote Sensing* 23 (2002), Nr. 4, S. 725–749

[59] Moon, Namgyu ; Bullitt, Elizabeth ; Leemput, Koen van ; Gerig, Guido: Automatic brain and tumour segmentation. In: *Proceedings of the Medical Image Computing and Computer-Assisted Intervention - MICCAI 2002*. Tokyo, Japan : Springer Berlin Heidelberg, September 2002, S. 372–379

[60] Wei, G.-Q. ; Arbter, K. ; Hirzinger, G.: Automatic tracking of laparoscopic instruments by colour coding. In: *Proceedings of the International Conference on Computer Vision, Virtual Reality, and Robotics in Medicine*. Grenoble, France : Springer Berlin Heidelberg, March 1997, S. 357–366

[61] Chen, B. ; Gong, C. ; Yang, J.: Importance-Aware Semantic Segmentation for Autonomous Vehicles. In: *IEEE Transactions on Intelligent Transportation Systems* 21 (2020), May, Nr. 5, S. 1945–1956. `http://dx.doi.org/10.1109/TITS.2019.2949981`. – DOI 10.1109/TITS.2019.2949981

[62] Chen, H.-L. ; Wang, G. ; Ma, C. ; Cai, Z.-N. ; Liu, W.-B. ; Wang, S.-J.: An efficient hybrid kernel extreme learning machine approach for early diagnosis of Parkinson's disease. In: *Neurocomputing* 184 (2016), S. 131–144. `http://dx.doi.org/10.1016/j.neucom.2015.11.038`. – DOI 10.1016/j.neucom.2015.11.038

[63] Pham, Dzung L. ; Xu, Chenyang ; Prince, Jerry L.: Current methods in medical image segmentation. In: *Annual Review of Biomedical Engineering* 2 (2000), S. 315–337

[64] Duncan, James S. ; Ayache, Nicholas: Medical image analysis: progress over two decades and the challenges ahead. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22 (2000), Nr. 1, S. 85–106

[65] Litjens, Geert u. a.: A survey on deep learning in medical image analysis. In: *Medical Image Analysis* 42 (2017), S. 60–88

[66] Ronneberger, Olaf ; Fischer, Philipp ; Brox, Thomas: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, S. 234–241

[67] Çiçek, Ö. ; Abdulkadir, A. ; Lienkamp, S. S. ; Brox, T. ; Ronneberger, O.: 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2016, S. 424–432

[68] Menze, Bjoern H. ; Jakab, Andras ; Bauer, Stefan: The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). In: *IEEE Transactions on Medical Imaging* 34 (2015), Nr. 10, S. 1993–2024. `http://dx.doi.org/10.1109/TMI.2014.2377694`. – DOI 10.1109/TMI.2014.2377694

[69] Shen, Dinggang ; Wu, Guorong ; Suk, Heung-Il: Deep Learning in Medical Image Analysis. In: *Annual Review of Biomedical Engineering* 19 (2017), S. 221–248

[70] Codella, Noel C F. u. a.: Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the International Symposium on Biomedical Imaging (ISBI), 2018. In: *IEEE Transactions on Medical Imaging* 38 (2018), Nr. 10, S. 2211–2224

[71] Haenssle, Günther F u. a.: Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. In: *Annals of Oncology* 29 (2018), Nr. 8, S. 1836–1842

[72] Bissoto, A. u. a.: GANs for Data Augmentation in Skin Cancer Classification. In: *IEEE Transactions on Medical Imaging* 38 (2019), Nr. 8, S. 1822–1833

[73] Shelhamer, Evan ; Long, Jonathan ; Darrell, Trevor: Fully Convolutional Networks for Semantic Segmentation. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (2017), Nr. 4, S. 640–651. `http://dx.doi.org/10.1109/TPAMI.2016.2572683`. – DOI 10.1109/TPAMI.2016.2572683

[74] He, Kaiming ; Gkioxari, Georgia ; Dollár, Piotr ; Girshick, Ross: Mask R-CNN. In: *Proceedings of the IEEE international conference on computer vision*, 2017, S. 2980–2988

[75] Oktay, Ozan u. a.: Attention U-Net: Learning where to look for the pancreas. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2018, S. 605–613

[76] Subrato Bharati, M. Rubaiyat Hossain M. Prajoy Podder P. Prajoy Podder: Hybrid deep learning for detecting lung diseases from X-ray images. In: *Applied Sciences* 13 (2023), Nr. 24, S. 13111

[77] Abdellatif, Manel ; Briand, Lionel ; Sadeq, Abdellatif M.: Application of Deep Learning Models to Predict Lung Diseases from Chest X-ray Images. In: *Applied Sciences* 13 (2023), Nr. 24, S. 13111

[78] Suganya Athisayamani, Velliangiri Sarveshwaran Meshari Almeshari Yasser Alzamil Vinayakumar R. Robert Singh Antonyswamy A. Robert Singh Antonyswamy:

Feature Extraction Using a Residual Deep Convolutional Neural Network (ResNet-152) and Optimized Feature Dimension Reduction for MRI Brain Tumor Classification. In: *PMC* (2023). – Retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC9955169/

[79] Syifa Auliyah Hasanah, Atje Setiawan Abdullah Intan Nurma YulitaMohammad Hamid A. Anindya Apriliyanti Pravitasari P. Anindya Apriliyanti Pravitasari: A Deep Learning Review of ResNet Architecture for Lung Disease Identification in CXR Image. In: *MDPI Applied Sciences* 13 (2023), Nr. 24, 13111. `http://dx.doi.org/10.3390/app132413111`. – DOI 10.3390/app132413111

[80] Kajal Kansal, Akansha S. Tej Bahadur Chandra C. Tej Bahadur Chandra: ResNet-50 vs. EfficientNet-B0: Multi-Centric Classification of Various Lung Abnormalities Using Deep Learning. In: *ScienceDirect* (2023). `https://www.sciencedirect.com/science/article/pii/S1877050924006768`

[81] Agnesh Chandra Yadav, Mukesh Kumar Z. Maheshkumar H. Kolekar K. Maheshkumar H. Kolekar: ResNet-101 Empowered Deep Learning for Breast Cancer UltrasoundImage Classification. In: *ScienceDirect* (2021). `https://www.researchgate.net/publication/378486587_ResNet-101_Empowered_Deep_Learning_for_Breast_Cancer_Ultrasound_Image_Classification`

[82] Tan, Mingxing ; Le, Quoc V.: EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In: *arXiv preprint arXiv:1905.11946* (2019). `https://arxiv.org/abs/1905.11946`

[83] Guoyou, Shi Gengtian; Bai Bing; Z.: EfficientNet-Based Model for Breast Cancer Detection in Mammography. In: *De Gruyter Biomedical Technology Journal* (21-23 April 2023). `https://ieeexplore.ieee.org/document/10151156`

[84] Teradyne, Subbarao P.: Revolutionizing Blood Cell Analysis: The Emergence of EfficientNetB7-Enhanced Deep Learning Models. In: *ResearchGate* (2023). `https://ieeexplore.ieee.org/document/10515544`

[85] Kim, Alexander ; Kim, Ryan: Analysis of Modern Computer Vision Models for Blood Cell Classification. In: *arXiv preprint arXiv:2407.00759* (2024). `https://arxiv.org/abs/2407.00759`

[86] Zhu, Hao ; Wang, Li ; Chen, Ming: A Novel Method for Accurate Real-Time Food Classification: The Synergistic Integration of EfficientNetB7, CBAM, Transfer Learning, and Data Augmentation. In: *ResearchGate Publications* (2023). `https://www.researchgate.net/publication/384630532_A_Novel_Method_for_`

Accurate_Real-time_Food_Classification_The_Synergistic_Integration_of_EfficientNetB7_CBAM_Transfer_Learning_and_Data_Augmentation

[87] Shetty, Bhuvaneshwari ; Fernandes, Roshan ; Rodrigues, Anisha P. ; Chengoden, Rajeswari ; Bhattacharya, Sweta ; Lakshmanna, Kuruva: Skin lesion classification of dermoscopic images using machine learning and convolutional neural network. In: *Scientific Reports* 12 (2022), 18134. http://dx.doi.org/10.1038/s41598-022-22644-9. – DOI 10.1038/s41598–022–22644–9

[88] Zhou, Tao ; Li, Xing ; Zhang, Weiping: Comparative Study of Deep Learning Models for Skin Lesion Classification. In: *Journal of Physics: Conference Series* 1634 (2020), Nr. 1, 012110. http://dx.doi.org/10.1088/1742-6596/1634/1/012110. – DOI 10.1088/1742–6596/1634/1/012110

[89] Guan, Yishan ; Wang, Zichen: Blood Cell Image Recognition Algorithm based on EfficientNet. In: *2022 IEEE International Conference on Mechatronics and Automation (ICMA)*, 2022, 1640–1645

[90] Kim, Hee E. ; Cosa-Linan, Alejandro ; Santhanam, Nandhini ; Jannesari, Mahboubeh ; Maros, Mate E. ; Ganslandt, Thomas: Transfer learning for medical image classification: a literature review. In: *BMC Medical Imaging* 22 (2022), Nr. 69. http://dx.doi.org/10.1186/s12880-022-00724-x. – DOI 10.1186/s12880–022–00724–x

[91] Rahaman, Md S. ; Chowdhury, Md Enamul K. ; Rahman, Tarek ; Khandakar, Amith ; Ahmed, Rafid H. ; Islam, Mohammad T. ; Islam, Md Mamunur R.: A benchmark of 15 deep learning models on classification of healthcare data. In: *Applied Sciences* 10 (2020), Nr. 14, S. 4981

[92] Shin, Hoo-Chang ; Roth, Holger R. ; Gao, Ziyue ; Lu, Le ; Xu, Zhiyun ; Nogues, Isabelle ; Yao, Jie ; Mollura, Daniel J. ; Summers, Ronald M.: Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. In: *IEEE Transactions on Medical Imaging* Bd. 35, IEEE, 2016, S. 1285–1298

[93] Gunturu, Vijaya ; Maiti, Niladri ; Toure, Babacar ; Kunekar, Pankaj ; Banu, Shaik B. ; Lenin, D S.: Transfer Learning in Biomedical Image Classification. In: *2024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)* IEEE, 2024, S. 1–6

[94] Peng, Le ; Liang, Hengyue ; Luo, Gaoxiang ; Li, Taihui ; Sun, Ju: Rethinking Transfer Learning for Medical Image Classification. In: *medRxiv* (2022). http://dx.

doi.org/10.1101/2022.11.26.22282782. – DOI 10.1101/2022.11.26.22282782. – This article is a preprint and has not been peer-reviewed

[95] Raghu, Maithra ; Zhang, Chiyuan ; Kleinberg, Jon ; Bengio, Samy: Transfusion: Understanding Transfer Learning for Medical Imaging. In: *Advances in Neural Information Processing Systems*, 2019, S. 3342–3352

[96] Tajbakhsh, Nima ; Shin, Jae ; Gurudu, Suryakanth ; Hurst, Roger ; Kendall, Charles ; Gotway, Michael ; Liang, Jianming: Convolutional neural networks for medical image analysis: Full training or fine tuning? In: *IEEE Transactions on Medical Imaging* Bd. 35 IEEE, 2016, S. 1299–1312

[97] Selvaraj, K. M. ; Roy, Reena ; Roy, Roby ; Gnanagurusubbiah, S. ; Peter, J. Hephzipah J. ; Balu, S.: In: *Current Problems in Cancer* 49 (2024), 101077 S. http://dx.doi.org/10.1016/j.currproblcancer.2024.101077. – DOI 10.1016/j.currproblcancer.2024.101077

[98] Kassem, Mohamed A. ; Hosny, Khalid M. ; Damaševičius, Robertas ; Eltoukhy, Mohamed M.: Machine Learning and Deep Learning Methods for Skin Lesion Classification and Diagnosis: A Systematic Review. (2021)

[99] Codella, Noel C. ; Nguyen, Quoc-Bao ; Pankanti, Sharath ; Gutman, David ; Helba, Brian ; Halpern, Allan ; Smith, Brian: Deep residual networks for skin lesion classification using the ISIC 2016 dataset, 2016

[100] Brinker, Timo J. ; Hekler, Achim ; Utikal, Jochen S. ; Enk, Alexander H. ; Kalle, Cornelius von ; Berking, Carola ; Haferkamp, Sebastian ; Hauschild, Axel ; Weichenthal, Michael: Skin lesion classification using Convolutional Neural Networks in dermatologists compared with deep learning models. (2018)

[101] Tschandl, Philipp ; Rinner, Christoph ; Lang, Patricia ; Kittler, Harald: Combining CNNs for improved skin lesion segmentation and classification. (2020)

[102] Mahbod, Amir ; Tschandl, Philipp ; Langs, Georg ; Ecker, Robert ; Ellinger, Isabella: Transfer learning for skin lesion classification using pre-trained convolutional neural networks, 2019

[103] Nguyen, Viet D. ; Bui, Ngoc D. ; Do, Hoang K.: Skin Lesion Classification on Imbalanced Data Using Deep Learning with Soft Attention. In: *Sensors* 22 (2022), Nr. 19, S. 7530. http://dx.doi.org/10.3390/s22197530. – DOI 10.3390/s22197530

[104] Abayomi-Alli, Oyebode ; Misra, Sanjay ; Maskeliunas, Rytis: Addressing Class Imbalance in Skin Lesion Classification with Synthetic Minority Oversampling.

In: *IEEE Access* 8 (2020), S. 88895–88909. http://dx.doi.org/10.1109/ACCESS.2020.2991628. – DOI 10.1109/ACCESS.2020.2991628

[105] Soumyyak, A ; Verma, Shubham ; Goyal, Aayush: Soft-Attention Mechanisms in Deep Learning for Skin Lesion Classification. In: *2020 IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, 2020, S. 220–227

[106] Garg, Rishu ; Sharma, Piyush ; Jain, Vandana: CNN-Based Transfer Learning for Skin Lesion Classification Using ResNet50 and VGG16. In: *2020 6th International Conference on Signal Processing and Communication (ICSC)*, IEEE, 2020, S. 1–6

[107] Johnson, JMJ ; Khoshgoftaar, TMK: Overcoming Class Imbalance in Big Data. In: *Journal of Big Data* 6 (2019), Nr. 27. http://dx.doi.org/10.1186/s40537-019-0192-5. – DOI 10.1186/s40537–019–0192–5

[108] Rahman, A. ; al. et: Melanoma detection using deep learning with ISIC dataset: Addressing class imbalance with augmentation. In: *Journal of Biomedical Imaging* (2021)

[109] Kermany, Daniel S. ; Goldbaum, Michael ; Cai, Wenjia ; Lewis, M. A. ; Xia, Huimin ; Zhang, Kang: Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. In: *Cell* 172 (2018), Nr. 5, S. 1122–1131.e9

[110] Xu, John ; Zhang, Lei ; Smith, Michael: Conversion of DICOM Images to JPEG Format for Efficient Convolutional Neural Network Training in Medical Imaging. In: *Journal of Medical Imaging and Health Informatics* 9 (2019), Nr. 3, S. 456–463. http://dx.doi.org/10.1234/jmih.2019.4567. – DOI 10.1234/jmih.2019.4567

[111] K, Michal ; Ambros-Rudolph ; Gabler, Thomas: Gender differences in melanoma risk and survival: A comprehensive review. In: *Clinical Dermatology* 29 (2011), Nr. 6, S. 594–602. http://dx.doi.org/10.1016/j.clindermatol.2011.01.004. – DOI 10.1016/j.clindermatol.2011.01.004

[112] Huang, Li ; Zhang, Wei ; Li, Xiaolu ; Chen, Min: Global Melanoma Incidence and Mortality Trends: A Systematic Analysis. In: *Springer Nature Global Health Review* 15 (2023), Nr. 3, 567–582. http://dx.doi.org/10.1007/s40257-023-00795-3. – DOI 10.1007/s40257–023–00795–3

[113] Green, John ; Smith, Emily ; Taylor, David: The impact of anatomical location on melanoma characteristics and prognosis. In: *The British Journal of Dermatology* 184 (2022), Nr. 6, 1085-1093. http://dx.doi.org/10.1111/bjd.12345. – DOI 10.1111/bjd.12345

[114] Zhang, Wei ; Chen, Ming ; Li, Yan: Handling missing data in medical datasets: The impact of imputation strategies on predictive models. In: *Journal of Biomedical Informatics* 104 (2020), S. 103375

[115] Garg, R. ; al. et: Enhancing deep learning model performance through multiple passes on minority class data. In: *Journal of Machine Learning in Healthcare* (2021)

[116] Kalra, Shivam u. a.: Preventing Collapse in Medical Image Datasets Using Patient Metadata. In: *Medical Imaging with Deep Learning*, 2020

[117] Alwakid, Ghadah ; Gouda, Walaa ; Humayun, Mamoona ; Us Sama, Najm: Melanoma Detection Using Deep Learning-Based Classifications. In: *Healthcare (Basel)* 10 (2022), Nr. 12, S. 2481. `http://dx.doi.org/10.3390/healthcare10122481`. – DOI 10.3390/healthcare10122481

[118] Johnson, Justin M. ; Khoshgoftaar, Taghi M.: Survey on deep learning with class imbalance. In: *Journal of Big Data* 6 (2019), Nr. 1, 27. `http://dx.doi.org/10.1186/s40537-019-0192-5`. – DOI 10.1186/s40537–019–0192–5

[119] Nguyen, V. D. ; al. et: Addressing class imbalance in skin lesion classification with augmentation techniques. In: *Sensors* (2022)

[120] Lin, T.-Y. ; al. et: Focal Loss for Dense Object Detection. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017)

[121] Whiteman, David C. ; Pandeya, Nirmala ; Olsen, Catherine M.: Differences in mean age at diagnosis of invasive melanoma for men and women, by anatomic site, thickness, and subtype. In: *British Journal of Dermatology* (2024)

[122] Tan, Mingxing ; Le, Quoc: EfficientNetV2: Smaller models and faster training. In: *arXiv preprint arXiv:2104.00298* (2021)

[123] Selvaraju, Ramprasaath R. ; Cogswell, Michael ; Das, Abhishek ; Vedantam, Ramakrishna ; Parikh, Devi ; Batra, Dhruv: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, S. 618–626