FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Multi-Modal Tasking for Skin Lesion Classification using Deep Neural Networks

Rafaela Garrido Ribeiro de Carvalho

Mestrado Integrado em Bioengenharia

Supervisor: Tudor Nedelcu, PhD, Fraunhofer Portugal

Co-Supervisor: João Pedrosa, PhD, FEUP | DEEC

June 21, 2021

Multi-Modal Tasking for Skin Lesion Classification using Deep Neural Networks

Rafaela Garrido Ribeiro de Carvalho

Mestrado Integrado em Bioengenharia

June 21, 2021

Abstract

Skin lesion describes any abnormal skin tissue and it can be indicative of cancer. Skin tumors are divided in non-melanoma, and melanoma, a very life-threatening condition despite accounting for a minority of cases. Given its tendency to metastasize, early diagnosis is of extreme importance. The process of diagnosis involves visual inspection by dermatologists, however it portrays subjective results. This has motivated the development of automated skin lesion analysis systems.

The encouraging advent of artificial intelligence has allowed for the development of intelligent solutions for skin lesion classification, firstly with machine learning algorithms and, in recent years, with deep learning networks capable of matching human level performance. Nonetheless, there is still room for improvement in order to incorporate this type of computer-aided diagnosis solutions in a clinical setting. Another crucial limitation is the lack of standardized test datasets, run by a third-party organization, for evaluation and comparison of systems.

This dissertation introduces an automated system for skin lesion classification using deep neural networks with innovative aspects. In opposition to the majority of methods described in the literature which focus on detecting melanoma, the work aims to classify several types of skin lesions. To build a robust deep learning model which meets the aforementioned requirements, several techniques were explored: transfer learning, multi-tasking and multimodal learning.

Firstly, the performance of a multi-layer perceptron with hand-crafted features based on the ABCD rule of dermoscopy was compared to a model with deep learning generated features extracted by the EfficientNet-B3 pre-trained on the ImageNet dataset. After inferring the superiority of the latter, a multi-task model with auxiliary related tasks was implemented, providing superior results in all considered metrics. To handle the imbalance of the dataset, oversampling was applied in addition to the introduction of class weights in loss functions; no improvement was verified in the area under the curve values but the models with weighted loss functions originated significantly higher sensitivity for melanoma and seborrheic keratosis. The role of segmentation in classification was also assessed and it was concluded that it was detrimental to performance. Later, manually extracted asymmetry ratio and border gradient were divided in classes and used as auxiliary targets due to their possible correlation to melanoma but no overall increase in performance was observed, only a rise in sensitivity of seborrheic keratosis and specificity of nevus and melanoma. Finally, multimodal learning was studied with the implementation of early fusion technique (combination of the dermoscopic image with its corresponding lesion mask), and the late fusion strategy (concatenation of hand-crafted asymmetry and/or border with deep learned features). The first produced poorer results, specifically for melanoma where the model is just making random presumptions; the second allowed to increase area under the curve of all classes.

This project proved the feasibility of these techniques and their application in skin lesion diagnosis systems, however there is still a clear window of opportunity for further developments.

Keywords: skin lesions, multimodal learning, multi-task learning, deep neural networks.

Resumo

Lesão cutânea descreve qualquer tecido de pele anormal e pode indicar cancro. Estes tumores são divididos em não melanoma e melanoma (condição que põe em risco a vida, apesar de representar a minoria de casos). Dada a tendência para metastizar, diagnóstico precoce é de extrema importância. O processo de diagnóstico envolve inspeção visual por dermatologistas mas é subjetivo. Tal motivou o desenvolvimento de sistemas automatizados de análise de lesões da pele.

O encorajador surgimento da inteligência artificial permitiu o desenvolvimento de soluções inteligentes para a classificação das lesões da pele, primeiramente com algoritmos de aprendizagem automática e, mais tarde, com redes neuronais profundas capazes de igualar peritos humanos. No entanto, há espaço para melhorias a fim de incorporar este tipo de soluções de diagnóstico assistido por computador em contexto clínico. Outra limitação crucial é a falta de bases de dados padronizadas, geridos por entidades externas, para avaliação e comparação de sistemas.

Esta dissertação introduz um sistema automatizado de classificação de lesões da pele, utilizando redes neurais profundas inovadoras. Este trabalho visa classificar vários tipos de lesões, em oposição aos métodos descritos na literatura que detetam melanoma. Para construir um modelo robusto de aprendizagem profunda que satisfaça os requisitos mencionados, foram exploradas várias técnicas: aprendizagem por transferência, aprendizagem multi-tarefa e multimodal.

Primeiramente, comparou-se o desempenho de um *perceptron* multicamadas com características manuais baseadas na regra ABCD de dermatoscopia com um modelo com características extraídas pelo modelo EfficientNet-B3 pré-treinado na base de dados ImageNet. Após inferir a superioridade deste, implementou-se um modelo multi-tarefas, com resultados superiores em todas as métricas consideradas. Para lidar com o desequilíbrio entre classes da base de dados, foram aplicadas sobreamostragem e pesos de classe nas funções de perda; não se verificou qualquer melhoria nos valores da área sob a curva, mas os modelos com funções de perda ponderada originaram sensibilidade mais elevada para melanoma e dermatite seborreica. O papel da segmentação na classificação foi também avaliado, concluindo-se que é prejudicial. Depois, o rácio de assimetria e o gradiente do limite foram divididos em classes e utilizados como alvos auxiliares devido à sua possível correlação com melanoma, mas não se observou qualquer aumento global do desempenho, exceto na sensibilidade de dermatite seborreica e especificidade de nevos e melanoma. Finalmente, estudou-se a aprendizagem multimodal com a implementação da técnica de fusão precoce (combinação da imagem dermatoscópica com a máscara de lesão correspondente), e de fusão tardia (concatenação da assimetria e/ou limite com características de aprendizagem profunda). O primeiro produziu resultados mais pobres, especificamente para o melanoma onde o modelo faz presunções aleatórias; o segundo permitiu aumentar a área sob a curva de todas as classes.

Este projeto provou a viabilidade destas técnicas e a sua aplicação em sistemas de diagnóstico de lesões cutâneas, no entanto, existe ainda oportunidade clara para futuros desenvolvimentos.

Palavras chave: lesões cutâneas, aprendizagem multimodal, aprendizagem multi-tarefa, redes neuronais profundas.

Acknowledgements

I would like to gracefully acknowledge the individuals who assisted me in the research and development of this dissertation. Firstly, my deepest appreciation to my supervisors Tudor-Ionut Nedelcu and João Pedrosa for their valuable inputs, guidance and availability. You were truly important in the development of this body of work and you have my gratitude. Thank you Fraunhofer AICOS for giving me the opportunity to develop this project and, in particular, the Derm.AI group, for your critics which helped me to refine my work.

I would like to also thank my parents, sister and family for their continued support, love and patience even when I explained way too many deep learning concepts to you.

Nuno for being my biggest supporter, my ride or die, the rock who always made sure I was feeling happy, motivated or needing a break. Thank you for always being by my side.

My 'og *****', we go way back. I appreciate your friendship, sense of humor, saturday night coffee meets, francesinha dinners and stupid jokes.

To the friends I met in university, thank you for all the laughs, more bearable classes (let's also not forget precious lecture notes), memorable nights and unforgettable praxe experiences and, most importantly, for making FEUP feel like a home away from home. I will forever cherish these amazing 5 years with you.

Rafaela Carvalho

"You have brains in your head. You have feet in your shoes. You can steer yourself in any direction you choose. You're on your own. And you know what you know. You are the one who'll decide where to go."

Dr. Seuss

Contents

1 Introduction **1**
1.1 Context and Motivation . . . 1
1.2 Aim of the Work . . . 2
1.3 Document Structure . . . 3

2 The Skin **5**
2.1 Skin Biology . . . 5
2.2 Skin Lesions . . . 7
2.3 Diagnosis . . . 9
2.3.1 Dermoscopy . . . 11
2.3.2 Dermoscopy Algorithms . . . 11
2.4 Teledermatological Screening . . . 16
2.5 Summary . . . 17

3 Automatic Skin Lesion Analysis **19**
3.1 Artificial Intelligence in Healthcare . . . 19
3.2 Evolution of Skin Lesion Classification . . . 20
3.3 Datasets and the Important Role of Challenges . . . 21
3.4 Machine Learning Systems . . . 24
3.4.1 Background on Machine Learning . . . 24
3.4.2 State-of-the-Art Machine Learning Methods . . . 24
3.5 Deep Learning Systems . . . 28
3.5.1 Background on Deep Learning . . . 28
3.5.2 State-of-the-Art Deep Learning Methods . . . 31
3.6 Towards Robust Lesion Classification . . . 33

4 Methodology **35**
4.1 Dataset Selection . . . 35
4.2 ABCD Rule Feature Extraction . . . 37
4.3 ABCD Rule-Inspired Neural Network . . . 39
4.3.1 Model Architecture . . . 39
4.3.2 Training Strategy . . . 40
4.4 CNN for Skin Lesion Classification . . . 40
4.4.1 Model Architecture . . . 41
4.4.2 Training Strategy . . . 41
4.5 Evaluation Metrics . . . 42

5 Experiments **45**
5.1 Hand-Crafted versus Deep Learning Generated Features 45
5.2 Multi-Task Learning 46
5.3 Optimization of Multi-Task Models with Class Balancing 48
5.4 Assessment of Segmentation Impact in Skin Lesion Analysis 49
5.5 Multi-Task and ABCD Rule Criteria Classification 50
5.6 Multimodal Multi-Tasking 51

6 Results and Discussion **53**
6.1 Hand-Crafted versus Deep Learning Generated Features 53
6.2 Multi-Task Learning 55
6.3 Optimization of Multi-Task Models with Class Balancing 56
6.4 Assessment of Segmentation Impact in Skin Lesion Analysis 58
6.5 Multi-Task and ABCD Rule Criteria Classification 60
6.6 Multimodal Multi-Tasking 61
6.7 Comparison with Benchmark Performances 64
6.8 Lessons Learned, Limitations and Future Work 65

7 Conclusions **67**

References **69**

A Supplementary Tables of Chapter 6 - Results **81**

List of Figures

2.1 Skin structure (from [13]). 5
2.2 Epidermis structure (from [16]). 6
2.3 Hierarchical classification of skin lesions (adapted from [22]). 7
2.4 Examples of pigmented skin lesions: a) Melanoma, b) Melanocytic Nevus, c) BCC, d) SCC, e) Dermatofibroma, f) Seborrheic Keratosis (images retrieved from [25]). 8
2.5 Comparison of imaging modalities: basal cell carcinoma (top) and in situ melanoma (bottom), by clinical photography (left) and dermoscopy (right) (images from [25]). 10
2.6 Dermoscopy assessment. 11
2.7 Examples of criteria for Menzies method (Adapted from [41]). 13
2.8 Examples of the ABCD rule (Adapted from [41]). 14
2.9 Examples of the seven-point checklist (adapted from [41]). 15

3.1 Comparison between machine learning and deep learning pipelines for classification of skin lesions. 20
3.2 Framework of a typical machine learning system (images from [109]). 27
3.3 Example of a typical DNN architecture 28
3.4 Framework of a CNN system (dermatological image from [109]). Conv. refers to Convolution, Norm. to Normalization, Pool. to Pooling and F.C. to Fully Connected. 29
3.5 Multi-task learning for a deep CNN. 30
3.6 Schematic of fusion models for multimodal learning: a) Early fusion, b) Late fusion (dermatological image from [25]). 31

4.1 Examples of: a) Original dermoscopic image, b) Superpixel-mapped annotations of dermoscopic features (Yellow ■ denotes pigment network, blue-green ■ is negative network, green ■ is milia like cyst and dark blue ■ is streaks), c) Segmentation mask. 36
4.2 ISIC 2017 challenge dataset: samples of nevus in the first row, melanoma in the second and seborrheic keratosis in the third. 36
4.3 Dataset distribution. 37
4.4 Illustration of the asymmetry extraction pipeline. 38
4.5 Illustration of the border extraction pipeline. 38
4.6 Example of image pixels with dermoscopic structures. 39
4.7 Architecture of the ABCD rule inspired neural network for skin lesion classification. 40
4.8 EfficientNet VS other CNNs on the ImageNet dataset (taken from [144]). 41

5.1 Diagram of the neural network model for skin lesion classification. 46

5.2 Diagram of the baseline model for skin lesion classification. 'Avg. Pool' denotes the average pooling layer, 'BN' is Batch Normalization, 'FC' is fully connected layer. 46
5.3 Diagram of the multi-task model. 'Avg. Pool' denotes the average pooling layer, 'BN' is Batch Normalization, 'FC' is fully connected layer. 47
5.4 Examples of the inputs for the multi-task model. 50
5.5 Diagram of the multimodal multi-tasking model with pixel-level image fusion. 'BN' is Batch Normalization, 'FC' is fully connected layer. 51
5.6 Diagram of the multimodal multi-tasking model with late feature fusion. 'BN' is Batch Normalization, 'FC' is fully connected layer. 52

6.1 Comparison between ROC curves of the models with hand-crafted and deep learned features. 54
6.2 Normalized confusion matrices for the models with hand-crafted and deep learned features. 54
6.3 Comparison between ROC curves of multi-task models with dermoscopic features classification as auxiliary tasks. 55
6.4 Normalized confusion matrices for the multi-task models with dermoscopic features classification as auxiliary tasks. 55
6.5 Comparison between ROC curves of the multi-task models with class balancing techniques. 57
6.6 Normalized confusion matrices for the multi-task models with class balancing techniques. 57
6.7 Comparison between ROC curves of multi-task model when segmented images and cropped around the lesion images are used as input. 58
6.8 Normalized confusion matrices for the multi-task model when segmented images and cropped around the lesion images are used as input. 59
6.9 Comparison between ROC curves of the multi-task models with ABCD rule criteria classification tasks. 60
6.10 Normalized confusion matrices for the multi-task models with ABCD rule criteria classification tasks. 60
6.11 Comparison between ROC curves of the multimodal multi-task model with pixel-level image fusion. 62
6.12 Normalized confusion matrix for the multimodal multi-task model with pixel-level image fusion. 62
6.13 Comparison between ROC curves of the multimodal multi-task models with late feature fusion. 63
6.14 Normalized confusion matrix for the multimodal multi-task models with late feature fusion. 63

List of Tables

2.1 Pattern Analysis [37]. 12
2.2 Classification system for Menzies scoring method [40]. 12
2.3 ABCD rule [37]. 13
2.4 Seven-point checklist. 14
2.5 Comparison between three dermoscopy algorithms. 15

3.1 Overview of publicly available datasets 23

4.1 ISIC 2017 challenge dataset description. 35

5.1 Class distribution of dermoscopic structures in the training set. 47

6.1 Results of the models with hand-crafted and deep learned features (bold values highlight the best result for each metric). 54
6.2 Results of the multi-task models, along with the results from the baseline multi-class model (bold values highlight the best result for each metric). 56
6.3 Results of the multi-task models with data balancing techniques (bold values highlight the best result for each metric). 58
6.4 Results of the optimized through data balancing multi-task models with modified images as inputs (bold values highlight the best result for each metric). 59
6.5 Results of the optimized through data balancing multi-task models with ABCD rule related auxiliary tasks (bold values highlight the best result for each metric). 61
6.6 Results of the optimized through data balancing multi-task models with multiple inputs (bold values highlight the best result for each metric). 63
6.7 Comparison between the top performing solution proposed in this work and the best challenge submissions (Avg. denotes Average). 64

A.1 Additional results of the models with hand-crafted and deep learned features (bold values highlight the best result for each metric). 81
A.2 Additional results of the multi-task models, along with the results from the baseline multi-class model (bold values highlight the best result for each metric). 81
A.3 Additional results of the multi-task models with data balancing techniques (bold values highlight the best result for each metric). 81
A.4 Additional results of the optimized through data balancing multi-task models with modified images as input (bold values highlight the best result for each metric). . 82
A.5 Additional results of the optimized through data balancing multi-task models with ABCD rule related auxiliary tasks (bold values highlight the best result for each metric). 82

A.6 Additional results of the optimized through data balancing multi-task models with multiple inputs (bold values highlight the best result for each metric). 82

Abbreviations

ACC	Accuracy
AI	Artificial Intelligence
AUC	Area Under the Curve
ANN	Artificial Neural Network
BN	Batch Normalization
BCC	Basal Cell Carcinoma
CAD	Computer-Aided Diagnosis
CCE	Categorical Cross Entropy
CNN	Convolutional Neural Network
DL	Deep Learning
DT	Decision Trees
DNN	Deep Neural Network
FN	False Negatives
FC	Fully Connected Layer
Fh-AICOS	Fraunhofer Centre for Assistive Information and Communication Solutions
FP	False Positives
FPR	False Positive Rate
GAN	Generative Adversarial Network
GLCM	Gray-Level Co-Occurrence Matrix
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
ISIC	International Skin Imaging Collaboration
KLT	Karhunen-Loéve Transform
KNN	K-Nearest Neighbors
LR	Learning Rate
ML	Machine Learning
MLP	Multi-layer Perceptron
MM	Malignant Melanoma
MTL	Multi-Task Learning
NMSC	Non-Melanoma Skin Cancer
NPV	Negative Predictive Value
NV	Nevus
PCA	Principal Component Analysis
PPV	Positive Predictive Value
ROC	Receiver Operating Characteristic
ROI	Region of Interest
SCC	Squamous Cell Carcinoma
SE	Sensitivity
SF	Store-and-forward

SK	Seborrheic Keratosis
SP	Specificity
SVM	Support Vector Machine
TCGFE	Texture-Colour-Geometry Feature Extraction
TD	Teledermatology
TDS	Total Dermoscopy Score
TN	True Negatives
TNR	True Negative Rate
TP	True Positives
TPR	True Positive Rate

Chapter 1

Introduction

This chapter introduces the general context of the work and the motivation to accomplish the proposed objectives. Innovative aspects of the dissertation's methodologies and the structure adopted are also described.

1.1 Context and Motivation

Cancer is a group of diseases characterized by the uncontrolled growth and spread of abnormal cells. It is, undeniably, a major public health problem worldwide, being the second leading cause of death globally [1]. Skin cancer comprises melanoma (MM) and non-melanoma skin cancer (NMSC), which are ranked the 18th and 5th most common cancers in the world, respectively.

It is widely known that early diagnosis improves prognosis of skin cancer. If diagnosis happens in a localized stage, patients have a 98% 5-year relative survival rate, i.e. 98 out of 100 people are expected to be alive 5 years after the diagnosis, whereas if diagnosed in a distant stage, the survival rate drops to 17% [2]. Therefore, the lives of human beings highly depend on a timely diagnosis.

The difficulty of early clinical diagnosis has led to the development of a non-invasive imaging technique: dermoscopy. This procedure improves the process of diagnosis of skin lesions by dermatologists by revealing dimensions of skin morphologic characteristics imperceptible to the naked eye [3], hence reducing the number of benign lesions unnecessarily biopsied. Algorithms for the identification of dermoscopic criteria which allow to distinguish between melanocytic and non-melanocytic as well as benign or malignant lesions were therefore developed. However, these systems are subjective, with a diagnosis highly dependent on the physician's training, previous experience and interpretation [4]. Moreover, visual differences between benign and malignant skin lesions can be particularly subtle and differentiating between them can be extremely difficult, even for trained professionals. Thus, the success of these methods is limited.

Due to the importance of early diagnosis, the shortage of experts in some regions and the insufficient and subjective nature of diagnosis algorithms, there exists a clear window of opportunity and motivation to develop computer-aided diagnosis (CAD) systems for this problem. Automated

classification systems can help to develop a screening of a large number of patients and reduce waiting times. This can be achieved because by promoting triage of most dangerous cases, it is possible to identify the individuals at higher risk. Ideally, the goal is to detect all skin cancers at an early stage.

Numerous solutions for automatic skin lesion classification have proposed in the literature, among which deep neural networks (DNNs) that have proved to deliver comparable results with medical experts [5]. This has been showcased in multiple skin image analysis challenges, hosted by the International Skin Imaging Collaboration (ISIC), where the top performing algorithms are consistently DNN-based [6, 7]. Recent studies found that the multimodal and multi-tasking training of classifiers are beneficial to performance [8, 9, 10].

Despite the promising results obtained in literature, there still exist limitations in automated systems that must be overcome. This fact stimulates the development of new, faster and more reliable algorithms.

Motivated by these aspects, this work is focused on investigating a multimodal and multi-tasking approach for classification of skin lesions using DNNs.

1.2 Aim of the Work

This dissertation was developed at the Fraunhofer Centre for Assistive Information and Communication Solutions (Fh-AICOS), as part of the Derm.AI: Usage of Artificial Intelligence to Power Teledermatological Screening project[1]. This project aims to improve the teledermatology processes between primary care units and dermatology services in the National Health Service, through a mobile application to acquire macroscopic skin lesion images and the development of AI-powered risk prioritization and decision support platform [11].

To contribute to this project, this work focus on the development of a skin lesion classification system based on a DNN approach, since this type of machine learning (ML) has demonstrated good performance in recent years. Moreover, it is currently the best performing approach, according to results in competitions dedicated to skin lesion analysis.

The main objective of the dissertation is to implement a system for skin lesion classification using DNNs, while making use of the multi-tasking and multimodal methods. Multi-tasking is employed for more efficient training. Furthermore, the fusion of two distinct modalities of data, dermoscopic images and metadata, is investigated to generate a better prediction.

The research in this work aims to contribute with two main innovative aspects.

- *Multi-Class Prediction* - Although the DNNs are delivering satisfactory results for skin lesion classifications, most of the methods described in the literature are focused to detect only one kind of skin lesion (MM detection). In this work, multiple skin lesion types are inspected and discriminated.

[1] http://dermai.projects.fraunhofer.pt/

- *Extraction of Auxiliary Metadata* - The multimodal and multi-tasking methods are limited to the available dataset, which has the skin lesions attributes labelled by medical professionals. In this work, useful features are extracted using computer vision or ML techniques and their impact in the prediction of the classifier is investigated.

1.3 Document Structure

This document is structured as follows:

- *Chapter 2 - The Skin* presents a broad overview of the biological framework of skin physiology and different types of skin lesions, providing the reader with the essential biological insights on the subject. The importance of early diagnosis is reinforced and the most common algorithms for skin lesion diagnosis are explored. A discussion on teledermatology as a helpful tool in triage referrals is also presented.

- *Chapter 3 - Automatic Skin Lesion Analysis* introduces a brief reflection on the evolution of automated skin lesion analysis. Furthermore, background information on the typical pipeline of ML systems in the field of skin cancer classification is provided and state-of-the-art DL approaches are summarized. The important role of challenges and publicly available datasets for benchmarking is also discussed.

- *Chapter 4 - Methodology* presents the dataset employed in the work, detailing its composition. Training settings used for the experiments are also described, differentiating between the design of a ML model and a convolutional neural network (CNN) architecture to address the stated problem. Performances measures are introduced as well.

- *Chapter 5 - Experiments* delineates the strategies implemented to build a robust skin lesion classification system, opposing a ML model and a pre-trained CNN, exploring methods to address the class imbalance problem as well as the multi-task and multimodal learning paradigms.

- *Chapter 6 - Results and Discussion* displays the results obtained for all experiments, proceeding to its analysis, discussion and comparison with state of the art methods. The clinical applicability of this study, main limitations and activities to be developed in future research opportunities are also stated.

- *Chapter 7 - Conclusions* concludes the main takeaway of the dissertation.

Chapter 2

The Skin

The biological structure of skin, types of skin lesions, incidence of skin cancer and the importance of its early diagnosis as well as the dermoscopic algorithms are discussed throughout this chapter.

2.1 Skin Biology

Skin is the largest organ in the body and covers its entire external surface. Its structure works as the body's first barrier against pathogens, UV light, chemicals and mechanical injury. Skin also regulates temperature and controls the release of water into the environment [12]. It is composed by three layers: the epidermis, dermis and hypodermis (or subcutis), as observed in Figure 2.1.

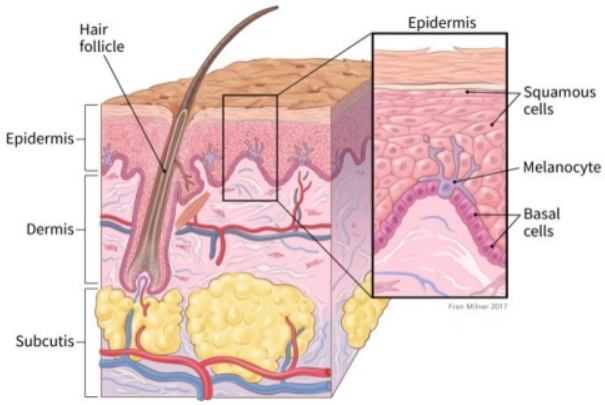


Figure 2.1: Skin structure (from [13]).

The upper skin layer is the epidermis and it can be structurally subdivided, as perceived in Figure 2.2. Its layers include: Stratum Corneum (predominant layer; consists in keratin and horny scales made up of dead keratinocytes. These secrete defensins which are part of the first immune defense of the body), Stratum Lucidum (thin clear layer, present in thicker skin found in the palms and soles), Stratum Granulosum (contains diamond shaped cells with keratohyalin granules and lamellar granules which keep the cells stuck together), Stratum Spinosum (where dendritic cells

can be found) and Stratum Basale (deepest layer; the proliferative capacity of the skin has been observed to be restricted to this layer, which is due to the presence of epidermal stem cells) [12].

The types of cells found in the epidermis are:

- Keratinocytes, the predominant cell type of epidermis. They originate in the basal layer and are responsible for the production of keratin. Upon reaching the outermost skin layer, the keratinocytes have undergone a further maturation process, have lost their nucleus and cytoplasmic organelles and are, from that moment on, referred to as corneocytes [14], responsible for the formation of the epidermal water barrier;
- Langerhans cells, which are the skin's first line defenders, belonging to the skin immune system;
- Merkel cells, oval-shaped cells located in stratum basale which serve a sensory function as mechanoreceptors for light touch, thus being most populous in fingertips [12];
- Melanocytes are neural-crest derived cells and primarily produce melanin, in dedicated organelles known as melanosomes. Melanin is a natural pigment that comes in different forms: brown/black eumelanin (leading type in the skin), red/yellow pheomelanin and brown/black neuromelanin. Differences in skin pigmentation can be attributed to a difference in the amount of melanogenesis and the distribution, size and content of melanosomes [15].

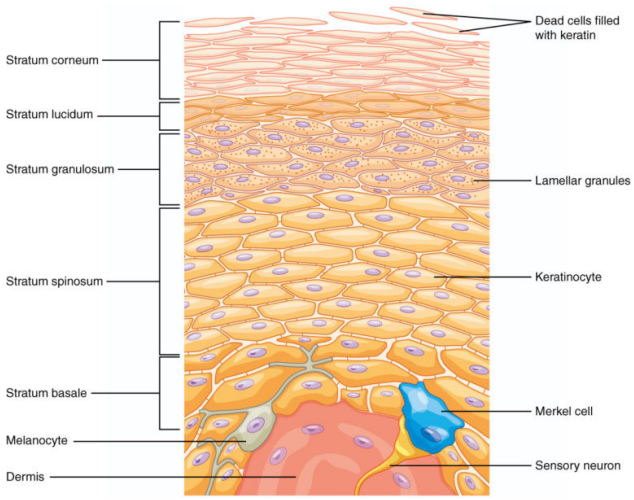


Figure 2.2: Epidermis structure (from [16]).

Dermis is connected to the epidermis at the level of the basement membrane and consists of two layers of connective tissue: papillary (upper and thinner layer, composed of loose connective tissue) and reticular (deeper and thicker, less cellular and with dense connective tissue), which merge together without clear demarcation [12]. The dermis contains the sweat glands, hair follicles, sensory receptors, blood and lymphatic vessels.

The hypodermis, also known as subcutaneous fascia, is the deepest layer of skin, consists of loose connective tissue and contains adipose lobules, thus functioning as an energy reserve.

2.2 Skin Lesions

A skin lesion is an atypical change in the normal appearance of skin tissue. The lesion is normally classified as benign or malignant, according to the non-cancerous, pre-cancerous or cancerous nature of the cell. Skin cancer refers to the abnormal growth of aberrant skin cells.

Risk factors associated with skin cancer are lighter skin, past sunburns, personal or family history of skin cancer. However, exposure to ultraviolet radiation (UVR) is the main cause. The sun's UVR can damage the deoxyribonucleic acid (DNA) in skin cell progressively, resulting in the growth of cancerous cells [17].

Skin cancers can be divided into two main types:

1. **Non-Melanoma Skin Cancer** is the most frequently diagnosed type of skin cancer in Caucasian population and is defined as a malignant neoplasm formed from keratinocytes, subdivided in basal cell carcinoma (BCC) and squamous cell carcinoma (SCC) [18]. The incidence of NMSC is 18-20 times higher than MM [19] and increases with age. About 80% of NMSC are BCC (it is also considered the most common form of skin cancer), whereas SCC represents 19%. The former grows mostly on sun-exposed areas, however they do it slowly and are unlikely to spread to other body parts; the latter also affects sun-exposed areas as well as damaged skin [18].

2. **Melanoma Skin Cancer** is a malignant tumor that arises from uncontrolled proliferation of melanocytes. Cutaneous MM is the most dangerous form of skin cancer [20]. MM used to be a rare cancer, but in the last 50 years its incidence is spreading faster than other cancers. Although it accounts for less than 5% of all cutaneous malignancies, MM is the most lethal, making up the largest portion of skin cancer deaths [20, 21].

Classification of Pigmented Skin Lesions

Skin lesions can be organized in a hierarchical way, as shown in Figure 2.3.

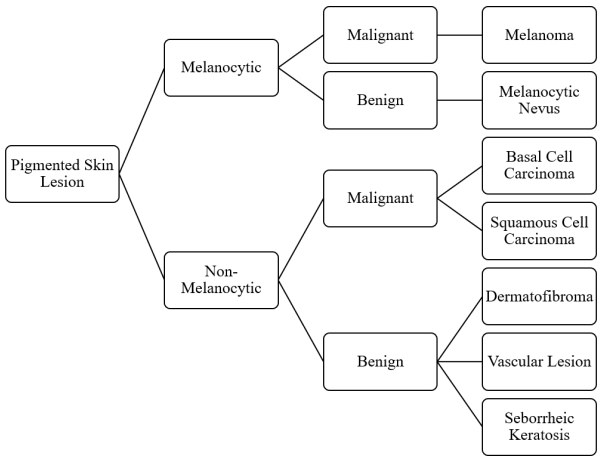


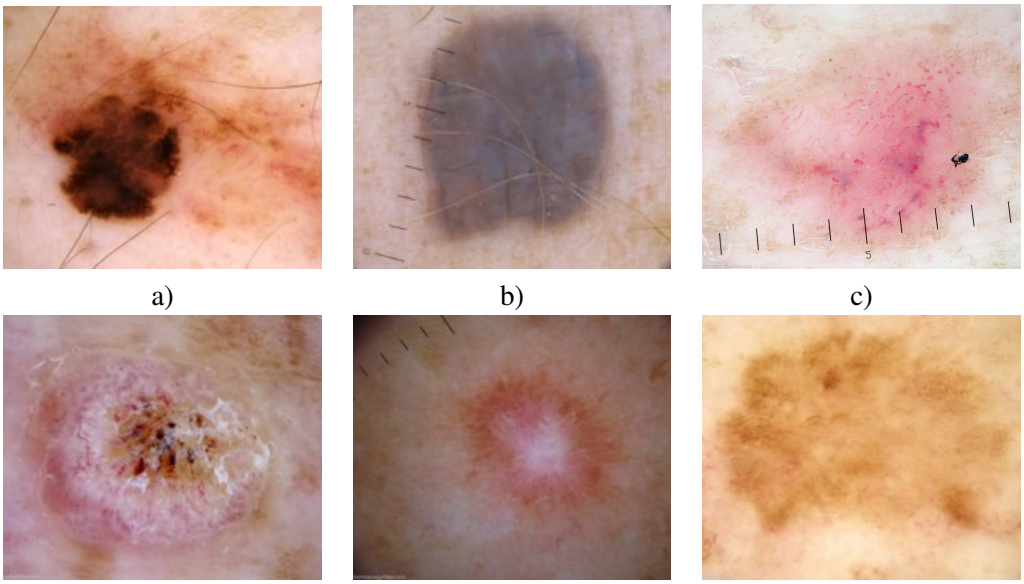
Figure 2.3: Hierarchical classification of skin lesions (adapted from [22]).

Firstly, it is important to identify the type of cells which are the source of the lesion: melanocytic lesions develop from melanocytes (skin cells responsible for the production of melanin); non-melanocytic lesions come from keratinocytes and can be subdivided in different classes according to their location on the epidermis (at the basal or squamous cell layers) [23]. The distinction between the two can be done based on the presence or absence of a set of pre-defined structures. Then, differentiation of lesion type is possible: malignant neoplasm or a benign lesion.

The malignant non-melanocytic neoplasms - NMSC - are described above. In opposition, benign non-melanocytic lesions include dermatofibromas, which are firm nodules whose surface is smooth and are mostly located on the lower extremities; vascular lesions, described as anomalies derived from capillaries, veins, lymphatic vessels and arteries [24]; and seborrheic keratosis (SK), which appear in older people and on any body area.

MM is the malignant form of melanocytic lesions. Their growth occurs at a much faster pace than BCC and they exhibit a remarkable capability to invade tissues and metastasize to other organs. It is of utmost importance to detect MM at an early stage, i.e., when it is located in the epidermis: MM in situ. When in a localized stage, the malignant cells are contained within the epidermis and have no contact with deeper skin layers and the blood stream. Thus, the cancer has not yet metastasized and it can be removed by an excision. The shape of early stage MMs is normally irregular and they exhibit a variety of colors. Invasive MMs may be papular or nodular, ulcerated and present a brown/black coloration with regions of red, white or blue. A melanocytic NV, or mole, is a common benign skin lesion. These lesions may be acquired or emerge at birth and may appear in any layer of the skin. However, it is indispensable to pay attention to this type of lesions, as they can be precursors to cutaneous MM.

Some examples of the referred lesions are represented in Figure 2.4.



a) b) c)

d) e) f)

Figure 2.4: Examples of pigmented skin lesions: a) Melanoma, b) Melanocytic Nevus, c) BCC, d) SCC, e) Dermatofibroma, f) Seborrheic Keratosis (images retrieved from [25]).

2.3 Diagnosis

As described in Section 2.2, there are two main types of skin cancer: NMSC and MM. In 2020, almost 325k new cases of MM were detected worldwide, ranking it as the 18th most common cancer, and over 57k deaths were registered. NMSC is the 5th most commonly occurring cancer globally, after breast, lung and bronchus, prostate and colorectal [26]. This type of tumor accounted for 1.2M new cases and almost 64k deaths, although the number of cases is likely to be much higher since NMSC is often not tracked by cancer registries (Source: Globocan 2020[1]).

Incidence rates of MM skin cancer rose by 44% between 2008 and 2018 with deaths increasing by 32%[2]. Globally, one person in every 26 522 develops MM skin cancer; Australia holds the 1st position, with 1 case per 1 746 individuals, followed by Northern European countries. As stated by European Cancer Information System (ECIS), MM incidence rates across Europe vary greatly, with highest estimated rates in Nordic countries, namely Denmark, the Netherlands and Sweden (with 1 per 2011, 2079, 2398 individuals, respectively); lowest incidence in Bulgaria, Romania, and Cyprus (1 per 11 164, 12 547, 13 988 individuals, respectively) (data from 2020) [27].

Duarte et al. (2018) [28] assessed the clinical and economic burden of MM and NMSC at public hospitals in mainland Portugal and found that, between 2011 and 2015, 6567 and 45 309 patients with MM and NMSC, respectively, were evaluated. Associação Portuguesa de Cancro Cutâneo predicted that 13k new cases of skin cancer would appear in 2020, with over 1000 of these being MMs, which present a mortality rate of 15%.

In addition to the considerable problem in public health, the economic burden of treatment is substantial. From 2002-2006 to 2007-2011, the average annual total cost for skin cancer in the US increased from $3.6B to $8.1B, representing a growth of 126.2%, while the average annual total cost for all other cancers only incremented by 25.1%. Average annual total treatment costs during 2007–2011 were $4.8B for NMSC and $3.3B for MM [29]. The increase in treatment costs results from the number of people treated for skin cancers but also from an increase of cost per capita. The average cost of treatment per patient increased from $1000 in 2006 to $1600 in 2011 [30]. The expenses depend of two factors: location (office treatment is more cost effective than that rendered in a hospital); type of treatment (destruction is the least expensive but with the lowest cure rate, followed by excision, Mohs surgery, superficial radiation treatment, ASC surgical excision, and, above all, treatment in the hospital outpatient department) [31].

These substantial expenses can be notably reduced by means of efficient prevention strategies. Moreover, considering that MM tends to metastasize beyond its primary site, by implementing these strategies for early diagnosis, it would be possible to reduce skin cancer incidence and mortality as well as treatment costs.

Once MM is advanced, surgery is no longer sufficient and it becomes more difficult to treat the disease [32]. Thus, an accurate classification of the type of skin lesion is required when choosing the treatment, as different types require distinct handling.

[1] https://gco.iarc.fr/today/fact-sheets-cancers

[2] https://melanomapatients.org.au/wp-content/uploads/2020/04/2020-campaign-report-GC-version-MPA_1.pdf

The process of diagnosing skin cancer usually begins with analysis of the anamnesis, i.e. medical history, and visual inspection of a suspicious lesion from a clinical expert. In cases where it is difficult to distinguish between a non-cancerous skin spot and skin cancer, the doctor may need to take a tissue sample, a biopsy, and perform histopathological examination under a microscope to confirm the diagnosis.

There are two distinct ways physicians look at a pigmented skin lesion: through a macroscopic (clinical) or microscopic (dermoscopic) view (cf. Figure 2.5).

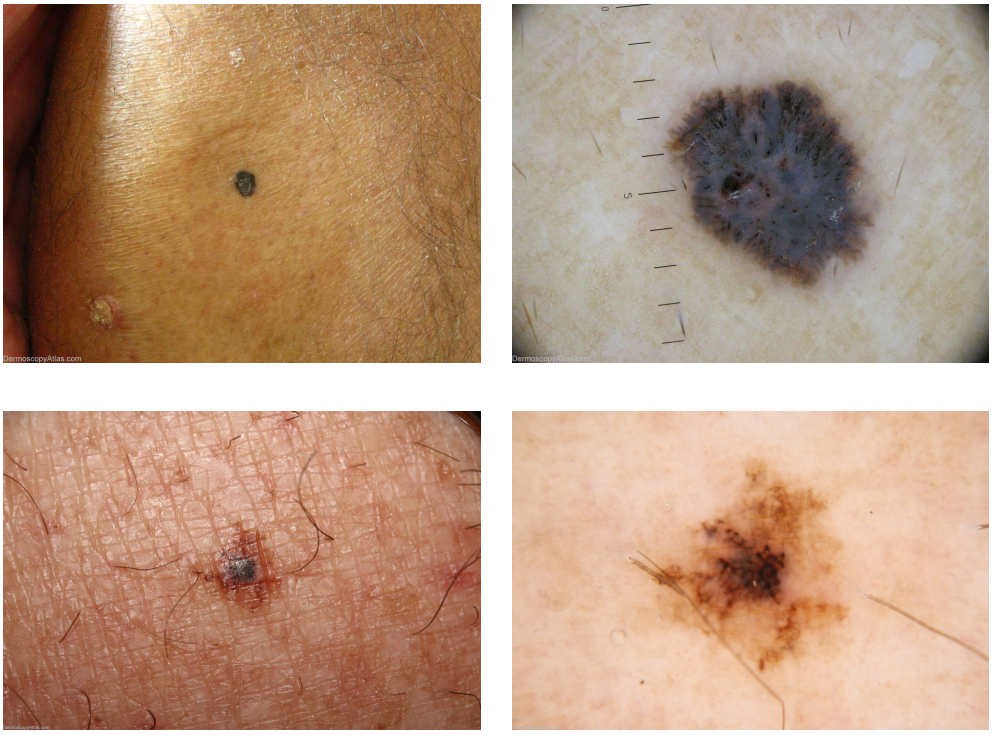


Figure 2.5: Comparison of imaging modalities: basal cell carcinoma (top) and in situ melanoma (bottom), by clinical photography (left) and dermoscopy (right) (images from [25]).

Clinical images are a representation of what the physician observes with the naked eye. On the other hand, dermoscopic images are magnified representations acquired through dermoscopy, a technique which follows the clinical screening and increases the sensitivity for skin cancer detection. Since it is the most commonly used and provides advantages such as reduction of the number of unnecessary biopsies (benign lesions biopsied), and diagnosis of thinner MMs compared to naked eye examination [33], this technique will be further explored in the following subsection.

A major obstacle to a successful diagnosis is the presence of artifacts. Namely, hairs, reflections, shadows, ruler marking, skin lines and air bubbles can confuse diagnosis and hinder achievement of better accuracy in the diagnosis process. Different devices and illumination conditions can lead to misdiagnosis, as well.

2.3.1 Dermoscopy

Dermoscopy is a non-invasive medical technique for in vivo observation of pigmented skin lesions (Figure 2.6), that uses light and magnification for a better evaluation of colors and microstructures of the epidermis, the dermoepidermal junction, and the papillary dermis not visible in plain sight. The identification of specific patterns related to the color distribution and dermoscopic structures can greatly help in the examination of the skin lesion [34]. This technique provides a valuable support in diagnosing skin lesions.

Figure 2.6: Dermoscopy assessment.

During a dermoscopy assessment, the pigmented skin lesion is typically covered with a liquid (oil or alcohol). The application of such fluids is necessary for the reduction of the reflectivity of the skin and enhancement of the transparency of the stratum corneum. This allows visualization of the aforementioned structures and it also suggests the location and distribution of melanin [34]. Afterwards, the lesion is investigated under a specific optical system (dermatoscope, stereomicroscope, videodermatoscope or digital imaging system).

2.3.2 Dermoscopy Algorithms

The major problem of visual assessment of skin lesions is its subjective nature. To address this, several algorithms for classification and diagnosis using dermoscopy have been developed.

In the world of dermatology, there are criteria to distinguish between melanocytic and non-melanocytic lesions and to perform the final diganosis: benign or malignant. These methods are based on the observation of numerous parameters related with dermoscopic structures and colors.

Several different methods of classification have been proposed in the literature but the most used procedures are pattern analysis, ABCD rule, Menzies method and seven-point checklist [35].

Pattern Analysis

Proposed by Pehamberger et al. [36], pattern analysis is the classic dermoscopic method for diagnosing skin lesions.

This procedure progresses in two steps. The first is to classify the lesion as melanocytic or non-melanocytic. This classification is performed based on global patterns (Table 2.1). A reticular

Table 2.1: Pattern Analysis [37].

Global Pattern	Local Features
Reticular Pattern	Pigmented Network
Cobblestone Pattern	Dots / Globules
Starburst Pattern	Streaks
Homogeneous Pattern	Blue-Whitish Veil
Parallel Pattern	Hypopigmentation
Multicomponent Pattern	Blotches
	Vascular Structures

pattern (pigment network that covers most parts of the lesion), a globular or cobblestone patterns (closely aggregated globules) and a starburst pattern (pigmented streaks in a radial arrangement localized at the periphery of the lesion) are usually identifiers of melanocytic lesions. The second step of the algorithm is to distinguish between benign melanocytic lesions and MMs. For this, an analysis of the local features is required (Table 2.1). If a lesion presents atypical features it is considered as malignant; typical structures are connected with benign lesions.

Pattern analysis increases the rate of correct decisions made by dermatologists. Nevertheless, the assessment is still subjective and lacks reproducibility, since its efficiency is correlated to the previous experience of the physician [34].

Menzies Scoring Method

The Menzies method is a simple dermoscopy method for diagnosing MMs [38]. It consists of 11 features, 2 negative and 9 positive, as specified in Table 2.2, which must be scored as present or absent by the observer. When none of the "negative features" and at least 1 of the 9 "positive features" are present, the lesion is classified as MM [39]. Examples of criteria are illustrated in Figure 2.7.

Table 2.2: Classification system for Menzies scoring method [40].

Diagnostic Criteria	
HIGHLY SUGGESTIVE OF MELANOMA	
Absence of both:	**Presence of at least one of the following:**
Pattern symmetry	Blue-white veil
Color uniformity	Multiple brown dots
	Pseudopods
	Radial streaming
	Scarlike depigmentation
	Peripheral black dots/globules
	5-6 colors
	Multiple blue/gray dots
	Broadened network

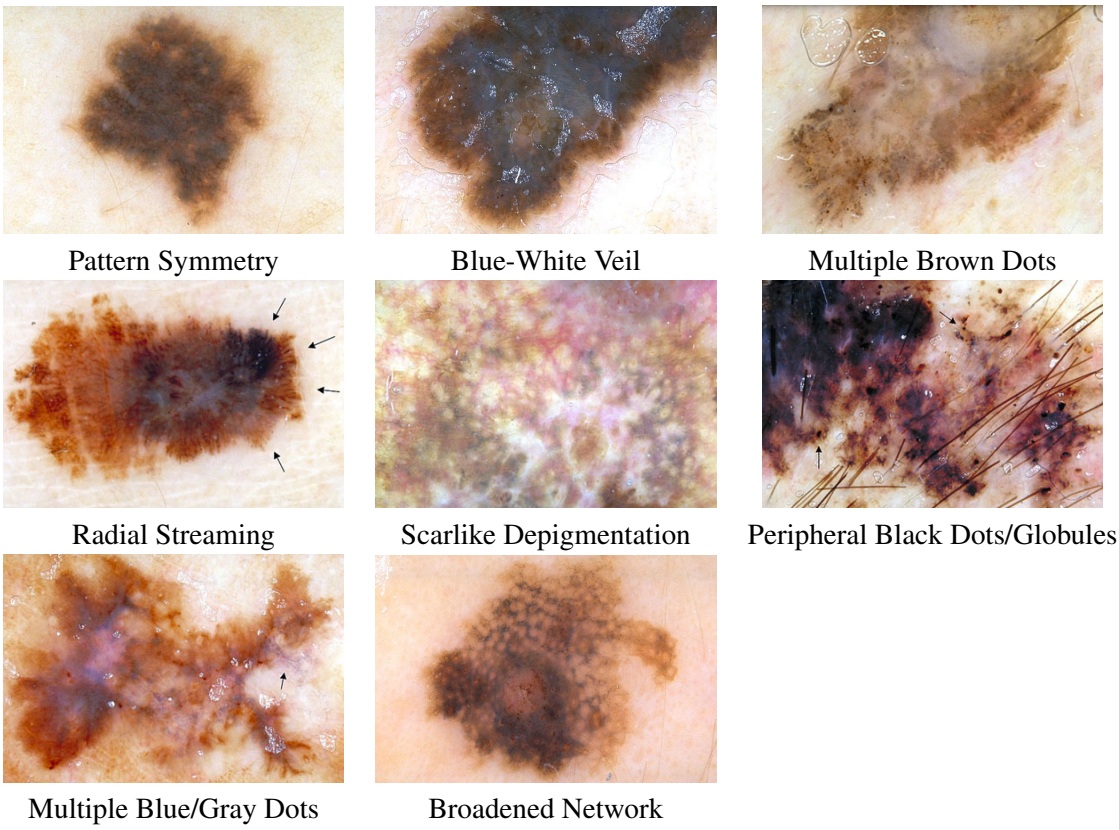


Figure 2.7: Examples of criteria for Menzies method (Adapted from [41]).

ABCD Rule

Described by Stolz et al. [42], it was the first dermoscopy algorithm developed to facilitate differentiation between the types of melanocytic lesions. Its functioning addresses whether the lesion is benign, suspicious or malignant in a quantitative manner. It is based on the criteria: Asymmetry (A), Borders (B), Color (C), Dermoscopic Structures (D) (Table 2.3).

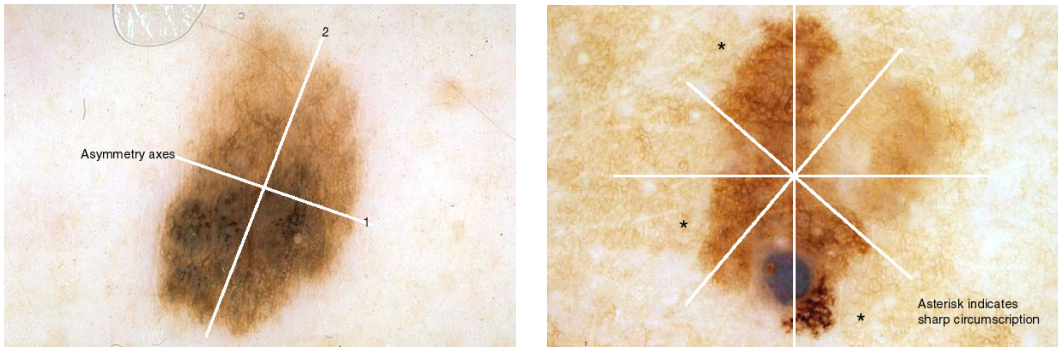
Table 2.3: ABCD rule [37].

Criterion	Description	Score	Weight
(A) Asymmetry	In 0, 1, 2 axes. Assess contour, colors and structures.	0-2	$\times 1.3$
(B) Border	Abrupt ending at the periphery in 0 to 8 segments	0-8	$\times 0.1$
(C) Color	Presence of up to 6 colors (white, red, light brown, dark brown, blue-grey, black)	1-6	$\times 0.5$
(D) Dermoscopic Structures	Presence of network, structureless or homogeneous areas, branched streaks, dots and globules	1-5	$\times 0.5$

A scoring system using these criteria allows to calculate the total dermoscopy score (TDS) using Equation 2.1. TDS represents a grading of the lesions with respect to their malignant potential.

$$TDS = 1.3 \times A_{score} + 0.1 \times B_{score} + 0.5 \times C_{score} + 0.5 \times D_{score} \tag{2.1}$$

For TDS < 4.75, the lesion is classified as benign melanocytic. Values between 4.8 and 5.45 are suspicious and if TDS is higher than 5.45, the lesion is diagnosed as MM.



A = 1 × 1.3 = 1.3	A = 2 × 1.3 = 2.6
B = 4 × 0.1 = 0.4	B = 3 × 0.1 = 0.3
C = 3 × 0.5 = 1.5	C = 4 × 0.5 = 2.0
D = 2 × 0.5 = 1.0	D = 4 × 0.5 = 2. 0
TDS = 4.2	**TDS = 6.9**
Benign Lesion	**Malignant Lesion**

Figure 2.8: Examples of the ABCD rule (Adapted from [41]).

In Figure 2.8, one can visualize the diagnosis performed with this rule. The benign lesion exhibits light-brown, dark-brown and black colors (3×C) as well as networks and dots as dermoscopic structures (2×D). The malignant lesion displays 4 colors (4×C) (light-brown, dark-brown, blue-gray and black) and 4 dermoscopic structures (4×D) (network, homogeneous areas, streaks, globules).

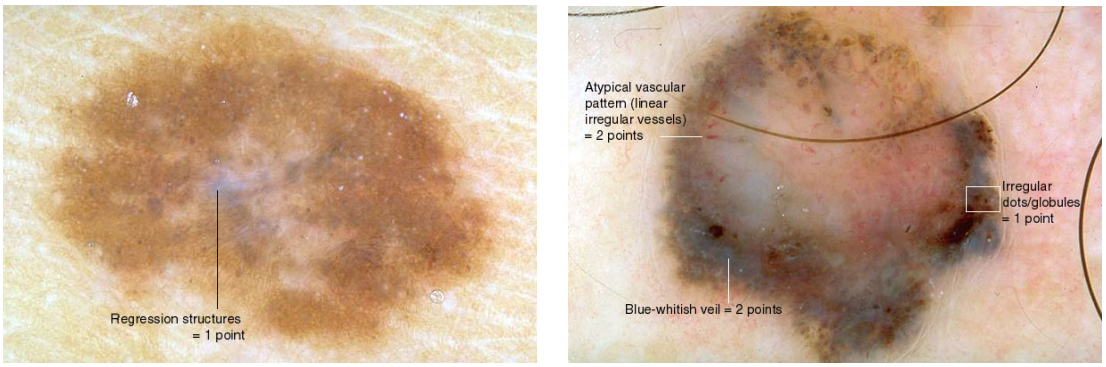
Seven-Point Checklist

Developed by Argenziano et al. [43], this algorithm is a variation of pattern analysis but with a score system. It requires the identification of 7 criteria, usually associated with MM. These are divided in two classes: major (3 features) and minor criteria (4 features), with different scores, respectively.

Table 2.4: Seven-point checklist.

Criterion	Score
Atypical Pigment Network	2
Blue-whiteish Veil	2
Atypical Vascular Pattern	2
Irregular Streaks	1
Irregular Pigmentation	1
Irregular Dots and Globules	1
Regression Structures	1

If any of the criteria is present in the lesion, it will receive a score, as seen in Table 2.4. A total score of greater than or equal to 3 is associated with a high likelihood of MM diagnosis [37]. Two examples of classification are presented in Figure 2.9.



Seven-Point Score = 1
Benign Lesion

Seven-Point Score = 5
Malignant Lesion

Figure 2.9: Examples of the seven-point checklist (adapted from [41]).

Overview

Argenziano et al. (1998) [43] also compared the reliability of the 7-point checklist with the ABCD rule of dermatoscopy and standard pattern analysis. For this study, 342 images of melanocytic lesions were used, with 57 and 60 MMs, and 139 and 86 benign lesions in the train and test sets, respectively. The results are described in Table 2.5.

Table 2.5: Comparison between three dermoscopy algorithms.

Algorithm	SE (%)	SP (%)	Diagnostic ACC (%)
Pattern Analysis	91	90	76
7-Point Checklist	95	75	64
ABCD Rule	85	66	51

Sensitivity (SE) is described as the probability of valid predictions when the lesion is a MM; specificity (SP) is the percentage of correct classifications of benign melanocytic lesions. Accuracy (ACC) is the number of correctly predicted lesions out of all the images. Formally, it is defined as the number of true positives and true negatives divided by the number of true positives, true negatives, false positives, and false negatives.

Pattern analysis is the most accurate algorithm. Nevertheless, this study concludes that all three methods are reliable for diagnosing MMs.

2.4 Teledermatological Screening

Telemedicine relies on technology of communication for exchanging expert medical information. The increasing interest can be explained by the evergrowing development of new technologies. Access, quality and cost-effectiveness are the basic issues of health care delivery and seemingly telemedicine can meet them all [44].

Dermatology is particularly suited for telemedicine, given the importance of its visual component. There is a growing interest in the potential, feasibility and reliability of teledermatology [45].

The literature has reported imaging techniques that can assist in the acquisition of skin lesion images [46]. Imaging techniques worth mentioning include: digital photography, radiography, in vivo confocal laser scanning microscopy, optical coherence tomography, ultrasound imaging, multispectral imaging and thermography. Furthermore, dermoscopy, described in Subsection 2.3.1, is seeing a growing increase in its use. Recently, the usage of dermatoscope coupled to a mobile phone camera has been adopted, hence facilitating acquisition of lesion images.

As rates of skin cancer are increasing, there is a growing concern about the timely delivery of health care both in rural and urban areas. In this regard, teledermatology (TD) could be a valuable tool in triage referrals, reducing time to diagnosis and treatment of malignant lesions, besides its potential benefits in terms of costs and waiting times and the ability to deliver specialised healthcare to more patients.

Teledermatology is often classified by the technology it uses: store-and-forward (SF) consultation, which involves transfer of clinical data to be evaluated at another location and time, and real-time (or interactive) videoconferencing [47]. The former has several advantages including lower costs, use of less complex equipment and less time-consuming consultations, offering the potential to shorten waiting lists. It is particularly suited for patients with poor access to healthcare as there is no need for coordinating scheduled visits, improving healthcare access and delivery [48]. It might be used across different time zones, not interfering with daily activities.

Despite the referred benefits, TD also presents a few limitations, namely clinical, economic, technological, legal and ethical issues [44]. For example, regarding clinical limitations, physical touch is important in diagnosing some skin conditions and it is lost in TD. For technological constraints, the cost of mobile devices equipped with high-quality cameras remains high, not making it accessible to the entire population [49].

An alternative method to frame teledermatology is based on the type of healthcare delivery. TD can be categorized into consultative, direct-care and triage models [50]. Notably, triage prioritizes patients based on the severity and urgency of their skin condition.

The use of teledermatology based on dermoscopy as a triage tool has shown high accuracy [48] and can reduce burden on healthcare systems and waiting times for necessary skin cancer surgery [51]. Automated classification systems can be a tool to help quickly screen a large number of patients, identify those most at risk and ideally detect skin cancers at an early stage.

2.5 Summary

Skin lesions are organized according to their type of source cells: melanocytic arises from melanocytes and non-melanocytic develop from keratinocytes, and is subdivided depending on their location at the basal or squamous layer of the upper skin layer, epidermis. Both types of lesions can then be classified as benign or malignant according to the non-cancerous, pre-cancerous or cancerous nature of the skin cell. Therefore, skin cancer can be NMSC or MM. The former is the most common skin cancer, affecting a much higher number of individuals than MM, however the latter is extremely dangerous due to its rapid pace and capacity to invade tissues and metastasize to other organs.

Early diagnosis acquires extra importance in this issue and is intrinsic to optimal patient health outcomes: if skin cancer is diagnosed in an initial state, there is an estimated 5-year relative survival rate of 98%; if diagnosed in a distant stage, the survival rate drops to 17% [2]. Moreover, treatment expenses can also be heavily reduced.

The difficulty of early clinical diagnosis has led to the development of dermoscopy, which is a non-invasive and effective imaging of potential skin cancer cases. The most commonly used algorithms for lesion inspection using dermoscopy are: pattern analysis, Menzies scoring method, ABCD rule and seven-point checklist.

These algorithms of dermoscopic criteria allow for an increased sensitivity and accuracy of the diagnosis process but the process remains highly dependent on the observer's experience and training. Moreover, considering that dermatology is particularly suited for telemedicine and with the burden in healthcare systems at present, automated systems may be the answer toward a system capable of diagnosing malignant skin lesions at an early stage.

Chapter 3

Automatic Skin Lesion Analysis

3.1 Artificial Intelligence in Healthcare

"AI will not replace doctors but instead will augment them, enabling physicians to practice better medicine with greater accuracy and increased efficiency."

– Benjamin Bell

Artificial intelligence (AI) is popularly known as the property of a computer or machine that mimics human intelligence characterized by behaviours such as cognitive ability, memory, learning and decision making. It is defined as the ability to mimic the capabilities of the human mind—learning from examples and experience, recognizing objects, understanding and responding to language, making decisions, solving problems, and combining these to perform "human" functions [52].

The idea of "machines that think" has been around for a long time, originating in ancient Greece and being relegated to science fiction in the first half of the 20th century. The term Artificial Intelligence was only created in 1956 [53]. Nowadays, AI is part of everybody's daily lives.

The evolution of AI has been empowered by the availability of large amounts of data and development of computer systems that can process data faster, more accurately and efficiently than humans can and with lower expenses [52, 54].

AI is prevalent in business and society and is beginning to be applied to healthcare due to the increasing availability of medical data.

Literature suggest that AI can perform as well or better than humans at various tasks such as diagnosing disease, speech transcription [55] and gaming [56]. Despite providing more accurate medical diagnosis, machines will not replace human physicians in the foreseeable future; in fact, AI must be considered an asset that can assist them to make better clinical decisions [57, 58].

3.2 Evolution of Skin Lesion Classification

Skin lesion classification has not escaped the trend toward AI diagnosis and the first description of using a computer for the analysis of cutaneous MM images was reported in 1987 [59].

Year 1995 marked the advent of dermoscopy, which enhanced the accuracy of both dermatologists and automated systems by allowing a better visualization of skin lesions.

The first survey of automated MM diagnosis was published by Day and Barbour (2000) [60]. The major issues reported were: (a) lack of a standard set of test images, (b) lack of detail in the description the proposed procedures, (c) usage of small datasets for model validation.

The literature was inspected again by Korotkov and Garcia (2012) [61]. The authors organized the overall pipeline of a computer-aided diagnosis system for skin lesion diagnosis. In addition, they reinforced the importance of providing a publicly available benchmark dataset for the proposed algorithms as a way to significantly improve performance and unite the efforts of different research groups. Each pigmented skin image for such a dataset should be accompanied by the ground truth definition of the lesion's borders and its diagnosis with additional dermoscopy reports from several dermatologists [62].

A number of CAD systems turned into commercially available products [61, 63], being used by some dermatologists around the world. However, these solutions are expensive, do not provide completely automated diagnoses and show need for improvement.

The existing CAD methods can be roughly divided into two groups: machine and deep learning. The latter is currently the preeminent option for skin lesion analysis [64]. The typical problems reported in 2000 [60] remained relevant throughout the years and are still observed. However, they do not affect systems as much because of several attempts to mitigate them.

Figure 3.1 compares the stages followed in a conventional ML setting and, simultaneously, demonstrates how DL promoted a step forward, by merging components in a single unit.

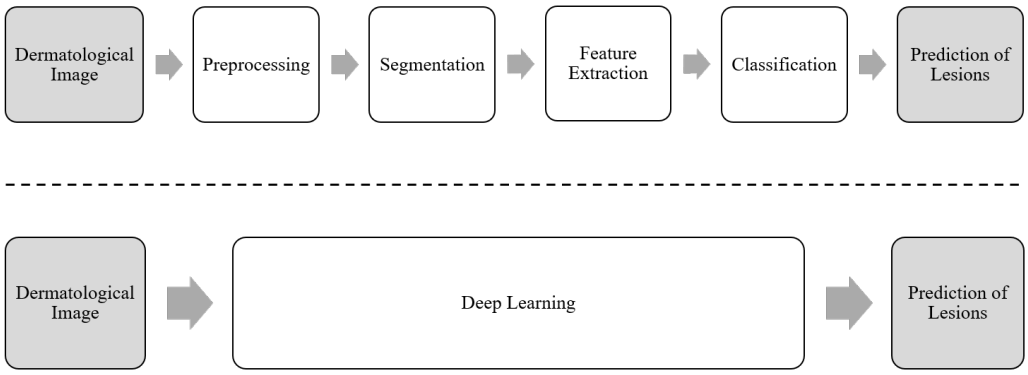


Figure 3.1: Comparison between machine learning and deep learning pipelines for classification of skin lesions.

3.3 Datasets and the Important Role of Challenges

To elaborate a reliable and a robust system for skin lesion classification, it is of extreme importance to have a miscellaneous image dataset. The decision of a CAD system depends heavily on the training set.

There are relatively few datasets in the general field of dermatology and even fewer datasets of skin lesion images. Most existing works on automated skin disease analysis use either private or very small publicly available datasets. Hence, such studies act merely as a proof-of-concept for the efficacy of AI in dermatology. Publicly available datasets are described in the following list and an overview is given in Table 3.1.

- PH² [65] is a dermoscopic image database acquired at the Dermatology Service of Hospital Pedro Hispano (Matosinhos, Portugal), made available in 2013. It consists of 200 melanocytic lesions, including 80 common nevi, 80 atypical nevi, and 40 MMs. Images were acquired using a magnification of $\times 20$ under the same conditions as Tuebinger Mole Analyzer system [65]. In addition, manual segmentation and the clinical diagnosis of the skin lesion as well as the identification of other important dermoscopic criteria are provided. These dermoscopic criteria include assessment of lesion asymmetry, identification of colors and a number of dermoscopic structures: pigment network, dots, globules, streaks, regression areas and blue-whitish veil [66]. The PH² database is freely available for research and educational purposes[1].

- The International Skin Imaging Collaboration (ISIC) is an international effort to improve MM diagnosis [67], whose aim is to aggregate a publicly accessible dataset of dermoscopy images[2]. It is currently the standard source for dermatoscopic image analysis research because of its permissive licensing, and large size but it is biased towards melanocytic lesions.

  - The first ISIC challenge was organized in 2016. A dataset with 900 dermoscopic images in JPEG format, binary masks in PNG format, dermoscopic feature files in JSON format and the gold standard malignancy diagnosis was provided [68], with the goal to support research and development of algorithms for automated diagnosis of MM.
  - In 2017, ISIC organized a new challenge [69] with a dataset of 2000 JPEG dermoscopic images, binary masks (PNG), dermoscopic features (JSON) and gold standard lesion diagnoses, which focused on three specific classes of lesions: MM, SK and benign nevi (NV).
  - The ISIC 2018 challenge was divided into three separate tasks: (1) lesion segmentation, (2) lesion attribute detection, (3) disease classification. Task 1 and Task 2 training data consist of 2594 images and 12 970 ground truth masks (5 for each image) extracted from ISIC 2017 Challenge [69] and HAM100000 datasets [70]. For Part 3, 10 015

[1] http://www.fc.up.pt/addi/
[2] https://www.isic-archive.com/

dermoscopic images [70, 71] divided in 7 classes, strongly imbalanced towards benign lesions, were provided.

  - The ISIC 2019 challenge dataset is a collection of some databases (HAM10000 [70] - Medical University of Vienna, BCN_20000 [72] - Hospital Clínic de Barcelona, MSK [69] - Anonymous) including 25 331 JPEG dermoscopic images and associated metadata, and it is labelled in 8 classes. The images have different resolutions and were corrected using different preprocessing and preparation protocols. Meta information for most images on the patient's age, gender, general anatomical site and common lesion identifier is available.

  - In 2020, the ISIC challenge [73] focused on a new approach for their dataset: multiple lesions from the same patient, because in practice dermatologists base their judgment integrating information from multiple lesions of the same patient. Therefore, such dataset totals a number of 33 126 JPEG or DICOM dermoscopic images, representative of 2056 patients, with an average of 16 lesions per patient; metadata on the patient's ID, sex, age, and lesion anatomic site is also provided. Images were collected in various parts of the globe: Memorial Sloan Kettering Cancer Center, New York (USA); Melanoma Diagnosis Centre, Sydney, Melanoma Institute Australia and University of Queensland, Brisbane (Australia); Medical University of Vienna (Austria); and Hospital Clínic de Barcelona (Spain).

- The Edinburgh Dermofit Image Library [74] is a collection of 1300 macroscopic skin lesion images and corresponding binary segmentation masks collected under standardised conditions with internal colour standards. Images consist of a snapshot of the lesion surrounded by normal skin. The lesions span across 10 different classes. The gold standard diagnosis is based on expert opinion (including dermatologists and dermatopathologists). The Dermofit Image Library is available under an academic licence[3].

- The 7-Point Criteria Evaluation database described by Kawahara et al. [10] includes 2022 clinical and dermoscopic color images (1011 images for each modality), along with corresponding structured patient metadata tailored for training and evaluating computer aided diagnosis (CAD) systems. The lesion cases span 5 diagnosis labels. The 7-point checklist is also provided. This dataset is publicly available online at the website[4].

- Dermnet [75] is a skin disease atlas with website support that contains over 23 000 skin images separated into 23 classes. The ratio between malignant and benign lesions is heavily unbalanced.

[3] https://licensing.edinburgh-innovations.ed.ac.uk/i/software/dermofit-image-library.html

[4] http://derm.cs.sfu.ca/Welcome.html

Table 3.1: Overview of publicly available datasets

DATASET	IMAGES	LESION IMAGES	CLASSES
PH^2 [65]	200	80 - Common Nevi, 40 - Melanomas, 80 - Atypical Nevi	3
ISIC 2016 Challenge [68]	900	273 - Melanoma, 627 - Non-Melanoma	2
ISIC 2017 Challenge [69]	2000	374 - Melanoma, 254 - Seborrheic keratosis, 1372 - Benign Nevi	3
ISIC 2018 Challenge [70, 71]	10 015	1113 - Melanoma, 6705 - Melanocytic Nevus, 514 - Basal Cell Carcinoma, 327 - Actinic Keratosis, 1099 - Benign Keratosis, 115 - Dermatofibroma, 142 - Vascular Lesion	7
ISIC 2019 Challenge [69, 70, 72]	25 331	4522 - Melanoma, 12 875 - Melanocytic Nevus, 3323 - Basal Cell Carcinoma, 867 - Actinic Keratosis, 2624 - Benign Keratosis, 239 - Dermatofibroma, 253 - Vascular Lesion, 628 - Squamous Cell Carcinoma	8
ISIC 2020 Challenge [73]	33 126	26199 - No Melanoma, 6927 - One or more Melanoma	2
Dermofit Image Library [74]	1300	76 - Melanoma, 331 - Melanocytic Nevus, 239 - Basal Cell Carcinoma, 45 - Actinic Keratosis, 257 - Seborrhoeic (Benign) Keratosis, 65 - Dermatofibroma, 88 - Squamous Cell Carcinoma, 78 - Intraepithelial Carcinoma, 24 - Pyogenic Granuloma, 97 - Haemangioma	10
7-Point Criteria Dataset [10, 76]	2022	252 - Melanoma, 575 - Melanocytic Nevus, 42 - Basal Cell Carcinoma, 45 - Seborrheic (Benign) Keratosis, 97 - Miscellaneous (Dermatofibroma, Vascular Lesion) for each image modality	5
Dermnet	23 000	190 - Melanoma	23

As described, ISIC has been organizing annual challenges for “Skin Lesion Analysis Towards Melanoma Detection" since 2016, using photos from their archive. These contribute not only with a public dataset but also a leaderboard, which present a way to benchmark results. They are, undoubtedly, the largest standardized and comparative study in this field to date.

With the emergence of these challenges for skin lesion classification, authors began to report their pipeline performance on pre-established training and testing sets, which allow comparison. Evaluation metrics are also being standardized: area under the curve (AUC) is systematically reported.

The problem of reproducibility is finally being tackled, enabling an eased dialogue between researchers.

3.4 Machine Learning Systems

3.4.1 Background on Machine Learning

Machine learning is a branch of AI, with algorithms that enable computers to learn from data and to improve their accuracy over time without being explicitly programmed. These algorithms extract patterns and features in data, making decisions and predictions in new and non-observed data [77].

ML is divided into two main types of learning:

- Supervised learning - consists of training on a labeled dataset, i.e., the data is labeled with information that the ML model must determine. Its aim to find generalized patterns. This method requires less training data and is used for classification, where the output is a variable, and regression, the result is a real number, tasks.
- Unsupervised learning - the ML model must infer knowledge from unlabeled data, identifying hidden structures or representations. Popular examples include clustering, autoencoders, Generative Adversarial Networks (GANs), etc.

Reinforcement learning is often classified as an additional ML paradigm since it is not exactly supervised nor unsupervised. It does not rely strictly on labels but it also does not explore patterns in data points, respectively. This kind of ML model learns via interaction and feedback, i.e. using a trial and error method. It takes actions in order to maximize a cumulative reward. Each time the algorithm chooses an action, it receives positive or negative feedback on its performance; after this, the algorithm is updated and will avoid penalties in future similar situations.

ML for skin lesion analysis has been attempted for years with its ultimate goal being improvement of the diagnosis process. Several ML approaches are described in the following section.

3.4.2 State-of-the-Art Machine Learning Methods

As reported by Korotkov and Garcia (2012) [61], the overall pipeline of skin lesion analysis follows a generic sequence of steps: image preprocessing, lesion segmentation, feature extraction and classification, as specified in Figure 3.1 of Section 3.2.

Diagnosis is highly dependent on the modality, quality and volume of images used. The inputs to a skin lesion CAD system are either clinical or dermoscopic images. Often, these do not have the optimal quality because of the variations in capturing devices and conditions of acquisition (e.g. contrast, intensity, angle, perspective), therefore affecting the accuracy of the subsequent algorithm.

Preprocessing

The first phase is preprocessing for removal of artifacts such as hair, ruler markings and dark corners, for reduction of noise effects and for image enhancement. Among the most necessary artifact rejection operations is hair removal since it may occlude parts of the lesion, hence making

correct segmentation and texture analysis impossible. The first widely adopted method for this task, system DullRazor, was proposed by Lee at al. [78] in 1997. Color normalization must also be performed as a correction step, to improve the differentiation between the lesion and skin. Median filtering suppresses noise such as small pores on the skin, shines and reflections. Illumination and contrast are also corrected. Principal Component Analysis (PCA) is widely used for edge enhancement during image smoothing.

Segmentation

Segmentation is the primary method for the separation of data into a region of interest (ROI). It is considered one of the most difficult tasks in medical imaging due to the complexity of skin images [79] and great variety of lesion colors, shapes, and sizes. An accurately detected lesion border is crucial for the diagnosis since a number of dermoscopic features (particularly, asymmetry and border sharpness) depend on it. The analysis of these features is, therefore, only as precise as the estimated lesion boundary. Additionally, there is high inter-observer variability in boundary interpretation among dermatologists [80], leading to a lack of definiteness in ground truth. Other difficulties include low-contrast of the lesion border [81], fuzzy border and irregular structures. Operations such as PCA and Karhunen-Loéve Transform (KLT), usually performed in the preprocessing stage, allow for an enhancement of lesion edges, ultimately resulting in better segmentation [82, 83]. A number of low-level segmentation techniques such as edge-based [84], region-based [85] and thresholding [86] approaches have been proposed in the literature [46]. These are conventional approaches because they are computationally simpler and faster, however they require post processing. High-level segmentation techniques include the low-level approaches and build sophisticated algorithms, namely fusion-based techniques, soft-computing based approaches and deformable models. Among their advantages is the fact that they avoid post-processing and deal with low contrast lesion boundaries.

Feature Extraction

For correct diagnosis of a skin lesion, dermatologists rely on the features of the lesion. Feature extraction is an endeavour to mimic clinicians' performance by extracting dermoscopic structures essential to diagnosis.

These features depend on the chosen diagnostic technique, from the ones explained in Section 2.3.2. For example, the border of a lesion and blue-whitish veil are dermoscopic features of the ABCD rule and pattern analysis, respectively. Many studies focus on detecting structures such as pigment network [87, 88, 89], structure-less areas, namely dots [90], globules [87], blotches [91], and asymmetry index [92].

In automated pigmented skin lesion classification, the system aims to extract these features from the images and represent them in a way that can be understood by a computer [93]. These representations will be referred to as feature descriptors. Different feature descriptors are associated with specific methods of diagnosis.

Feature descriptors are mostly classified as either color features [94, 95] or texture based descriptors [95, 96]. The former include island of color [97], color homogeneity [98] and color histograms [99]. The latter can be categorised as spatial frequency, statistical, geometric or model-based [100]. Spatial frequency based texture features are often linked to wavelet transformations; statistical based descriptors include co-occurrence matrices and Fourier properties for describing lesion's local neighbourhood properties; geometric features describe skin lesion characteristics that include shape, border, symmetry, area, diameter, variance, perimeter, circularity and anisotropy; model based textual descriptors are frequently associated with fractals and Markov random fields [46]. The most commonly used texture feature descriptors are the gray-level co-occurrence matrix (GLCM) [96, 99] and wavelet transform [101, 102].

Barata et al. (2013) [95] concluded that the color features outperformed the textual features when used singly but that the combination of the two yielded the most promising results.

The extraction step allows the determination of the malignancy of a skin lesion by a set of finite numerical features. Variables such as body location, age, and imaging parameters greatly influence the resulting values.

Classification

Lesion classification constitutes the final step in the typical framework for automated skin lesion analysis.

After feature extraction, it is often necessary to proceed to the selection of the most relevant characteristics and removal of redundant ones. Reducing the number of features will reduce the computational cost in the later stages. However, this reduction is not trivial as it is may adversely affect feature's discriminatory power.

Depending on the objective of the system, the output can be binary (malignant/benign or suspicious/non-suspicious), ternary (MM/dysplastic NV/common NV) or n-ary for several skin lesion classes [93].

ML methods such as Artificial Neural Networks (ANN), Decision Trees (DT), K-Nearest Neighbors (KNN), Support Vector Machines (SVM) and Logistic Regression are among the most commonly employed.

By using an ANN, Rubegni et al. (2002) [103] proposed a classification system. They reported great results: sensitivity (SE) of 94.3% and a specificity (SP) of 93.8%, on a dataset containing 550 images, with 200 of them being MM.

With a KNN classifier and a dataset consisting of 391 MM and 449 melanocytic nevi images, Burroni et al. (2004) [104] produced mean SE and SP equal to 98% and 79%, respectively.

Celebi et al. (2007) [99] employed an SVM on a dataset of 564 images, 88 of which being MM. They achieved SE of 93.3% and a SP of 92.3%.

Establishing an absolute hierarchy in terms of classifiers' performance is complicated because of the distinct datasets, feature descriptors, classifier parameters and learning procedures. Dreiseitl et al. (2001) [105] investigated the use of the five mentioned ML classifiers on automatic skin lesion classification with ternary output. They found that Logistic Regression, ANN and

SVM perform very well, achieving identical results; KNN has a modest performance; Decision Tree paradigm is not well suited for this problem domain due to the continuous input variables. Nevertheless, even the worst of the five achieves SE and SP values comparable to human experts.

Recently, ensemble methods, whose goal is to combine the strengths of separate classifiers, have also been proposed. They are able to improve the performance of skin lesion classification, outperforming individual classifiers [106].

Supervised ML algorithms are largely preferred to unsupervised approaches [61]. There is a high diversity of ambiguous clinical and dermoscopic features, i.e., they can point to either the malignant or benign nature of a lesion. Thus, there are a number of lesions whose corresponding biopsy-established diagnosis refutes the observed features [107]. In this case, the labelling is extremely necessary to teach a classifier to recognize abnormal manifestations of malignant lesions. Nonetheless, in the past years, a number of studies in which unsupervised learning techniques are introduced have been published [82, 108].

Overview

An example of the pipeline following the aforementioned steps is shown in Figure 3.2.

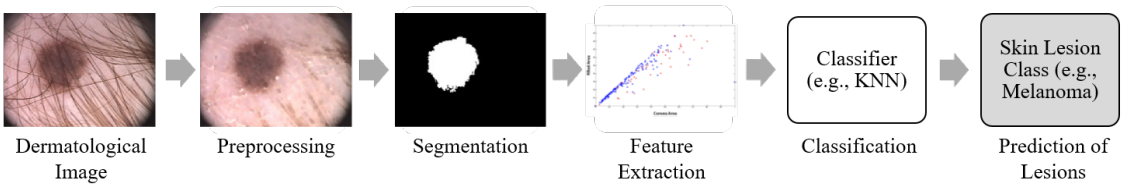


Figure 3.2: Framework of a typical machine learning system (images from [109]).

Results of a CAD system are typically dependent on the dataset, extracted feature descriptors and strength of the classifiers.

Human-engineered features are the main bottleneck in the ML system, as they are generally based on the diagnostic tools used by dermatologists, which are proven to be thoroughly subjective and unreliable. They work well for lesions with well defined and regular features, such as MM and BCC; in other lesions, the features are more complex and this solution becomes infeasible. Thus, hand-designed features extraction requires expertise and may not generalize to larger datasets.

Feature extraction and preprocessing are key tasks for the traditional methods but gruelling operations. Hence, recent literature is distancing from the classical approach and moving toward DL, as neural networks are capable of extracting features that are possibly more representative of the lesion.

3.5 Deep Learning Systems

3.5.1 Background on Deep Learning

Deep learning is a sub-field of ML, which has been highly potentiated by the evolution and empowerment of graphics processing unit (GPU) computing and the ever growing public available datasets. It allows computational models of multiple processing layers to learn and represent data with various levels of abstraction, in a functioning roughly inspired by the human brain [110].

DL incorporates neural networks, hierarchical probabilistic models, and a variety of unsupervised and supervised feature learning algorithms [111].

This field has recently excelled in human visual tasks [112], delivering significantly superior performances when compared with traditional computer vision techniques. However, an acute problem of DL algorithms is that they require massive amounts of data [77].

Deep Neural Networks (DNNs)

DNNs are multi-layered generalized linear models (see Figure 3.3).

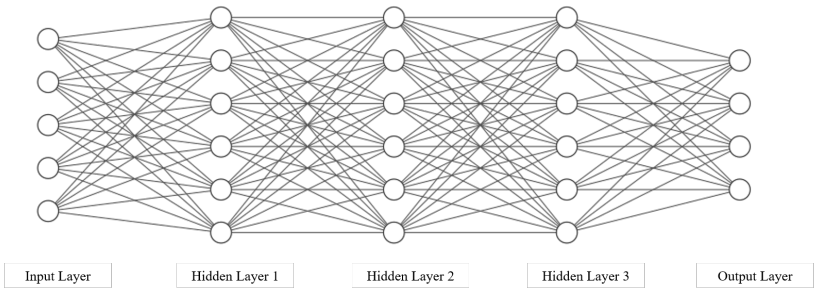


Figure 3.3: Example of a typical DNN architecture

The output of a neuron i localized on layer l is the image of activation function, $\sigma(W_{i,l})$, of a weighted sum of the neuron's input, vector x_{l-1}, with the weights being vector $W_{i,l}$ and scalar bias $b_{i,l}$ [113] (c.f. Equation 3.1).

$$f_{i,l}(x) = \sigma(W_{i,l} \cdot x_{l-1} + b_{i,l}) \tag{3.1}$$

The network is evaluated by calculation of a loss function (distance between the expected output and predicted value). The goal is to minimize the output of said function and such loss depends on the assigned weights; hence, a neural network model is trained using a gradient descent optimization algorithm and weights are updated using the backpropagation of error algorithm.

Updating the weights requires calculating the partial derivatives of the loss concerning each weight. Backpropagation is a technique which uses the chain rule for computing these derivatives. With this procedure, it is possible to find the weights that best adjust the model to the training set. The gradient descent algorithm is one of the most used optimization algorithms and seeks to change the weights in such a way that reduces the error in the next evaluation. Therefore, the optimization algorithm is navigating down the gradient.

Convolutional Neural Networks (CNNs)

Deep CNNs are one of the most important types of DL models, specialized in working with image data. CNNs are a specific type of DNN which use convolution rather than general matrix multiplication and are emerging as a very powerful tool for computer vision tasks, even showing ability to surpass human performance.

The first superlative triumph of CNNs in Computer Vision was achieved when Krizhevsky et al. (2012) [114] won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [115]. They achieved outstanding performance by implementing a large, deep CNN. Whereas traditional ML classifier would result in an approximate error of 30%, DL technology established an error rate under 17%. After this outstanding conquest, the error rate in the competition continued to decrease rapidly and, in 2015, it matched the average human classification error rate, 5% [116]. By 2017, CNNs were only committing half of the mistakes of a human (2.3%) [117].

One of the biggest assets and key of success of DL is the automatic extraction of features from the input data. As stated in Section 3.4.2, the extraction and selection of features were the most challenging and exhausting tasks when performing the traditional methods; by using CNNs, this is no longer necessary. Moreover, the preprocessing required is also greatly reduced.

Unlike ANNs, where every node fully connects to the next layer, each node of a CNN only connects to a number of nodes of the following layer. This pivotal attribute can capture the spatial and temporal dependencies in an image through the application of relevant filters [118].

As mentioned, the main drawback of DL algorithms is that they require enormous amounts of data and computational resources. Medical images and their respective labels are often not publicly available, thus hampering this approach (cf. Section 3.3). As a way to overcome the problem of small datasets, there are various free to use CNN architectures, pre-trained in enormous datasets (e.g., ILSVRC [115]), with application in the medical field. Because of their previously learned ability to extract image features, these can act as feature extractors in new algorithms through a technique known as Transfer Learning [119]. To apply it, the prior classifier of the original architecture is replaced with an untrained classifier fitting for the new task and the system is trained on the medical dataset [120]. A technique designated by data augmentation allows the generation of a multitude of new data by applying distortions: rotations, flips, color changes.

A generic representation of a CNN system for skin lesion classification is shown in Figure 3.4.

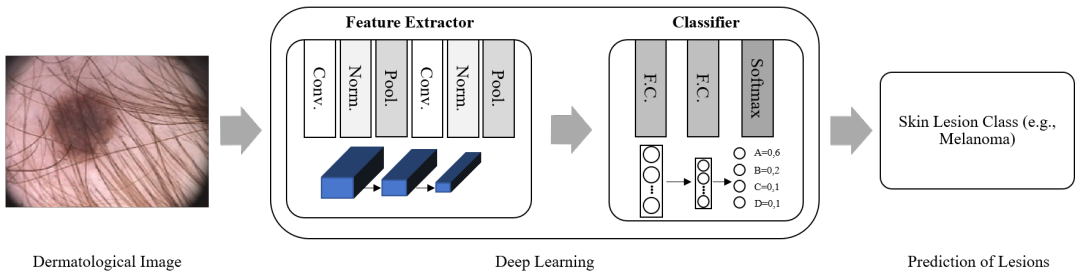


Figure 3.4: Framework of a CNN system (dermatological image from [109]). Conv. refers to Convolution, Norm. to Normalization, Pool. to Pooling and F.C. to Fully Connected.

Multi-task learning (MTL)

Multi-task learning [121] (Figure 3.5) is a learning paradigm whose aim is to leverage useful information contained in multiple related tasks to help improve the generalization performance of all tasks.

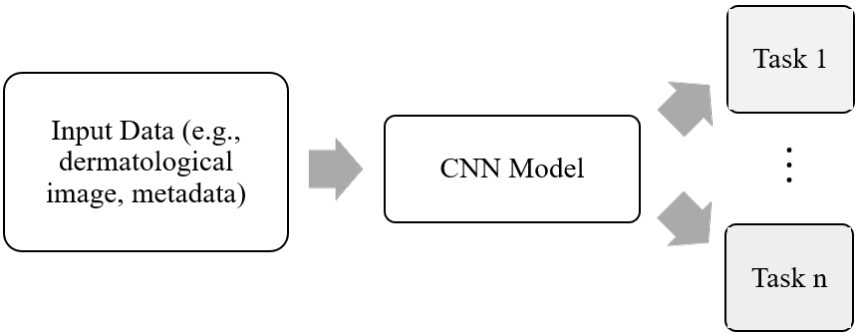


Figure 3.5: Multi-task learning for a deep CNN.

This area of ML can provide several benefits to the model. A number of them occur because the model accesses more data, and if the tasks are closely related, the model can learn beneficial representations. Different kinds of datasets have different noise; thus, by learning multiple tasks, it is easier to distinguish which features are beneficial and detrimental [122].

MTL has proven to be very successful in many computer vision problems [123]. Common approaches often share the convolutional layers, while learning task-specific fully-connected layers.

Deep Multimodal Learning

In medical diagnosis, dermatologists seldom evaluate only one image. These professionals combine dermoscopic or clinical view with external parameters such as medical history and patient personal information, namely age, gender, location of the disease, when analysing each lesion. Thus, it is clear that physicians integrate a myriad of data when making a diagnosis and it would be valuable for a DL network to replicate this behaviour.

Multimodal learning can meet these requirements. This paradigm aims to merge different data modalities with the objective of improving a network's prediction. DNNs offer the flexibility of implementing data fusion from n modalities with techniques such as early (or data-level), late (or decision-level), or intermediate fusion [124]. An illustration of the early and late fusion models is shown in Figure 3.6, as they are the opposite ends of the multimodal learning spectrum. The majority of work in deep multimodal fusion uses intermediate fusion [124], adopting approaches in between the two represented.

Multimodal learning allows for richer representation since different data types can provide complementary information to each other. Consequently, the output is expected to be more accurate than the predictions from individual modalities.

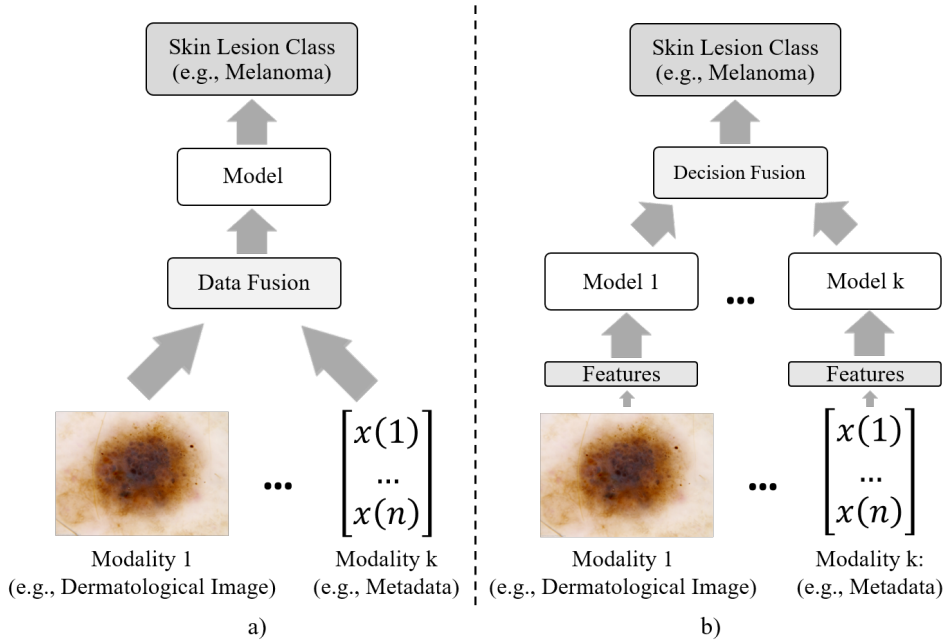


Figure 3.6: Schematic of fusion models for multimodal learning: a) Early fusion, b) Late fusion (dermatological image from [25]).

3.5.2 State-of-the-Art Deep Learning Methods

The blossoming of DL has promoted the development of promising skin lesion classification methodologies. CNNs have showcased promising results, capable of outperforming dermatologists [5].

Opposed to the tasks performed in a conventional setting, in DL all features and classification are learned and performed as a single unit (cf. Figure 3.1 of Section 3.2).

A solution to overcome the problem of limitation of data for skin lesion classification is by exploiting CNN architectures through transfer learning procedures [125, 5, 69]. Another practice to dodge this major issue of DL is a simple procedure: data augmentation [126, 127, 128]. This allows for an expansion of the dataset without all difficulties inherent to image acquisition.

One of the most important implementations of skin lesion classification with CNNs was achieved by Esteva et al. (2017) [5]. A private dataset of 129 450 clinical images, consisting of 2032 different diseases, trained a CNN. For this, a transfer learning procedure was implemented, by using Google's Inception v3 architecture. The final layer was replaced by the skin cancer classification task. All layers of the network were finetuned with RMSProp. The authors used a hierarchical partitioning algorithm using a taxonomy tree for data balancing. Altogether, the network showed performance results on par with 21 experts, indicating a solution capable of classifying skin lesions with a level of competence comparable to dermatologists.

Codella et al. (2015) [125] also applied transfer learning. Two fully-connected layers were taken from the Caffe CNN [129] pre-trained on the ILSVRC 2012 dataset [115, 114] and used as feature extractors; those features were subsequently fed to a SVM classifier. Their dataset was obtained from the ISIC Archive [67], containing 2624 clinical cases. The reported performance

matches the results with hand-designed features, which illustrates the feasibility of using these networks to extract relevant features.

Further work on the use of pre-trained CNNs on the ILSVRC dataset [115] was developed by Kawahara et al. (2016) [126], demonstrating how filters from a pre-trained network generalize to classifying 10 classes of 1300 non-dermoscopic skin images from the Dermofit Image Library [74], with a logistic regressor. This approach outperformed previously published results. Data augmentation with rotations and left-right flips improved results.

Ensemble techniques fuse the results from several classifiers into a final decision and have also been proposed for skin lesion classification [130, 131].

The correlation between skin lesions and their body site distributions was exploited by Liao et al. (2018) [9]. The authors built a deep MTL framework to jointly optimize skin lesion classification with a related auxiliary task, body location classification. The dermatology images used in the study were collected from DermQuest (as of 2021, it is deactivated), an atlas with both skin lesion and body location labels. The dataset categorizes lesions in 25 types and identifies 23 different body locations. Yang et al. (2017) [127] also suggested a MTL model which solves lesion segmentation and lesion classifications tasks at the same time. The model trained on 2000 training samples and 150 evaluation samples from the ISIC 2017 Challenge [69] attained promising results.

Yap et al. (2018) [8] investigated the combination of available data for classification. The Microsoft ResNet-50 CNN architecture with weights pre-trained on ImageNet was used to reduce the overfitting for a database of 2917 cases containing both clinical and dermoscopic images. A CNN trained on dermoscopic images presented higher accuracy than a CNN trained with clinical images. Nonetheless, when training the network on combined feature information from dermoscopic and clinical images, the accuracy outperformed single modal CNN, which indicates that both imaging modalities have dissimilar classification information. This new algorithm could be a step forward in developing a skin lesion classifier with both image modalities.

Deep network architectures can also be used as a way to provide features for the final classifier, as demonstrated in the works of Codella et al. (2015) [125] and Ge et al. (2018) [132]. The latter capture discriminative features of a private dataset (MoleMap[5]), annotated by expert dermatologists with disease labels, composed by more than 30 000 images for both imaging modalities. The authors are able to demonstrate that the proposed multimodal method significantly defeats single-modal ones.

Kawahara et al. (2019) [10] proposed a multi-task deep CNN with a base model Inception v3, pre-trained over ILSVRC [115]. The CNN architecture was then trained on multimodal data (clinical and dermoscopic images, as well as patient metadata), to classify the 7-point checklist criteria and perform skin lesion diagnosis. Their dataset containing the 2022 images and metadata has been made publicly available online (cf. Section 3.3). The network was trained using several multi-task loss functions, where each loss considers different combinations of the input modalities, thus allowing the model to be robust to missing data.

[5] https://www.molemap.co.nz/

3.6 Towards Robust Lesion Classification

Computer-Aided Diagnosis has attracted significant research attention and emerged as a tool to support skin lesion diagnosis. CAD systems can be used as a triage tool or second opinion capable of expanding the precision of analysis and decreasing the quantity of unnecessary biopsies. These systems do not depend on the individual, thus an automated analysis has the potential to empower patients with timely and reproducible diagnoses.

Previous dermatological computer-aided classification techniques require extensive preprocessing, lesion segmentation and extraction of features before classification. Recently, DL algorithms have been arising due to their excellence in computer vision tasks. The first DL pipelines for automatic skin cancer diagnosis [125, 133] emerged in 2015. Esteva et al. (2017) [5] published the landmark paper for skin lesion analysis, proposing a method which outperformed 21 specialists. DL solutions have now become a standard, with Haenssle et al. (2018) [134] also reporting higher performance than dermatologists. Multi-task and multimodal approaches have also been adopted by some authors [8, 10] and show ability to increase the performance.

The lack of standardized data is a dire problem for benchmarking. As a response to this, public datasets of skin lesion images, with structured and reliable information, are arising. The appearance of an open and global skin image analysis challenge in 2016, hosted by the ISIC, was a remarkable accomplishment. This international partnership has allowed the organization of the world's largest public repository of dermoscopic images of skin. Such dataset allows for the standardization of the conditions for evaluation of competing algorithms.

Nevertheless, the latest models still exhibit a number of limitations which need further progress in order to build an overall high-performance diagnosis system. For example, some lesions are falsely labelled as malignant, leading to unnecessary biopsies. Han et al. (2018) [135] tested an algorithm, trained on data from Asian individuals, in images from Europeans and its performance dropped, thereby demonstrating the need for a diverse dataset with all ethnicities. Furthermore, most methods in the literature are focused on detecting only one kind of skin lesion (MM). The multimodal and multi-tasking methods are also limited to the available dataset, which has the skin lesions attributes labelled by medical professionals. Such issues are tackled in this dissertation.

Chapter 4

Methodology

This chapter presents an overview of the training dataset utilized in this work, followed by a detailed description of the two major groups of models and training settings used in the dissertation experiments. The first type of model is a traditional CAD system, with steps of manual feature extraction and a multi-layer perceptron (MLP) architecture; the second is built upon a successful pre-trained CNN architecture, implementing a transfer learning approach which is widely used in image classification problems. The metrics used to assess the models are also presented.

4.1 Dataset Selection

The dataset employed in this work is from the ISIC 2017 challenge, as it provides access to multiple classes of skin lesions as well as ground truth annotations which fit and are useful for the objectives of this dissertation. The database aggregates 2000 training, 150 validation and 600 test JPEG dermoscopic images.

Table 4.1: ISIC 2017 challenge dataset description.

	Benign Nevi	**Melanoma**	**Seborrheic Keratosis**	**Total**
Training	1372	374	254	2000
Validation	78	30	42	150
Test	393	117	90	600

Images are labelled according to expert consensus and pathology report information, as malignant and melanocytic melanoma, benign and non-melanocytic seborrheic keratosis and benign and melanocytic nevus. Additionally, superpixel-mapped annotations (JSON files) of the presence and absence of the dermoscopic features (pigment network, negative network, milia like cyst and streaks), as well as expert manual tracing of the lesion boundaries (segmentation masks in PNG format) are provided (Figure 4.1).

Gold standard diagnosis are required for the training of our supervised models for skin lesion classification; annotations for the presence of dermoscopic features are employed in auxiliary tasks

of the multi-task approach and segmentation masks are used in the extraction of supplementary information.

a) b) c)

Figure 4.1: Examples of: a) Original dermoscopic image, b) Superpixel-mapped annotations of dermoscopic features (Yellow ■ denotes pigment network, blue-green ■ is negative network, green ■ is milia like cyst and dark blue ■ is streaks), c) Segmentation mask.

When analysing random examples from the dataset, represented in Figure 4.2, one can verify the difference in luminosity, contrast, aspect ratio, size and/or position of the lesion in different samples. Hence, with such heterogeneity, the dataset can be considered an accurate representation of real-world images, which allows for the training and validation of robust classification algorithms. Moreover, the quality of the data provided by this database allows researchers to focus on developing reliable models rather than concentrating in extensive pre-processing methods before training.

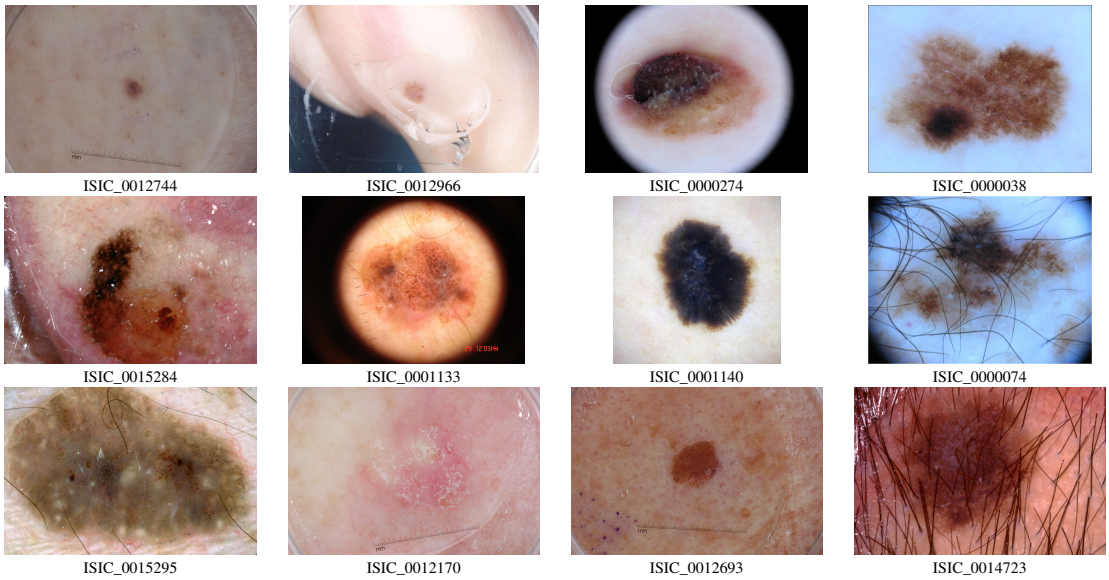


Figure 4.2: ISIC 2017 challenge dataset: samples of nevus in the first row, melanoma in the second and seborrheic keratosis in the third.

The test subset is reserved with the purpose of benchmarking classification results against other submissions presented in the challenge leaderboard.

As most real-world datasets in the health domain, classes are imbalanced. Figure 4.3 represents the distribution of the lesion classes across the train, validation and test sets.

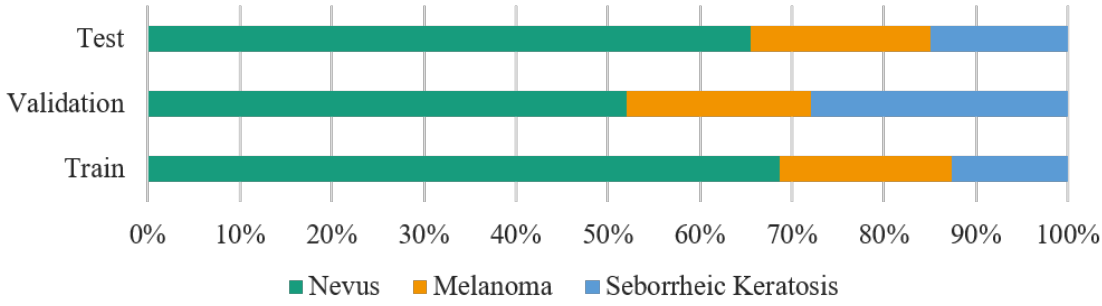


Figure 4.3: Dataset distribution.

4.2 ABCD Rule Feature Extraction

Among the diagnosis algorithms of dermatology, the most popular and responsible for inspiring many CAD systems is the ABCD rule of dermoscopy. According to this method, pigmented skin lesions can be characterized based on four criteria: asymmetry, border, color and number of dermoscopic structures, as described in Section 2.3.2. Thus, hand-crafted features based on this dermoscopy algorithm are designed to replicate this process.

Asymmetry

A pigmented skin lesion is considered asymmetrical when a line across its middle divides it into two halves and one half does not match the other. The overall asymmetry score (A_{score}) of a pigmented lesion is important when evaluating its malignant potential, with the ABCD rule giving it the highest weight out of all criteria (refer to Equation 2.1). This assessment is performed with respect to the shape: benign lesions are usually approximately circular, asymmetrical lesions provide a warning sign of MM.

We start by identifying the major and minor axes of the lesion with regard to the provided segmentation binary mask, calculating the major axis orientation (θ), as described in [99], with μ denoting the central moment:

$$\theta = \frac{1}{2} \cdot tan^{-1}(\frac{2\mu_{11}}{\mu_{20} - \mu_{02}}) \tag{4.1}$$

Secondly, the lesion is rotated θ degrees clockwise to align its major and minor axes with the x and y axes of the image and is centered (Figure 4.4b). For each axis, the mirrored version of one half is overlapped with its correspondent (Figures 4.4c and 4.4d, respectively) and the exclusive OR area between them is computed. A non overlapping area mask is obtained (Figure 4.4e), which allows the estimation of an asymmetry ratio (A) [136, 137], between the preceding (ΔT), and the total lesion area (T): $A = \frac{\Delta T}{T}$, for both axes.

(a) Original Mask (b) Mask Rotated and Centered (c) Top Half Mirrored (d) Bottom Half (e) Non-Overlapping Area Mask

Figure 4.4: Illustration of the asymmetry extraction pipeline.

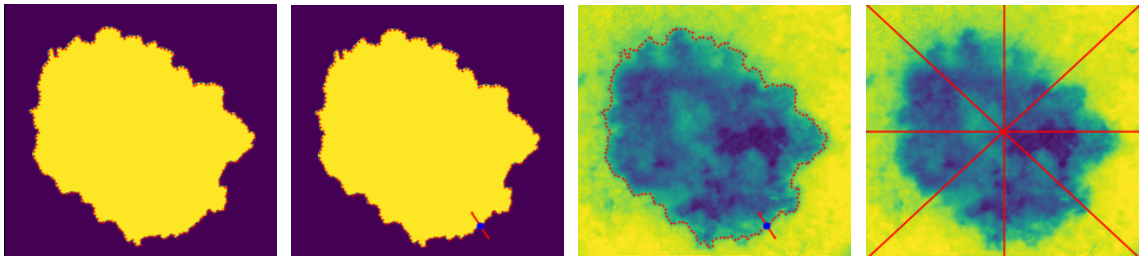
Border

Melanomas are usually associated with irregular and poorly defined borders, while benign nevi present even and smooth borders. In the clinical evaluation of the border, the sharpness of the transition from the lesion to the skin is determined (B_{score}).

Thus, to reproduce this modus operandi, the gradient is computed along the border points, using the blue channel as skin lesions are usually more noticeable in this channel [138]. The first step is to characterize the contour with 200 equidistant points and find the normal direction of each point (Figures 4.5a and 4.5b, respectively). The gradient in each border area is reduced to the mean difference of the pixel intensities along a line, represented in Figure 4.5c, whose length equals 30% of the lesion radius, as described in Equation 4.2 (N refers to the number of periphery points).

$$Gradient = \left| \frac{1}{N} \sum_{n=1}^{N} (inner\ pixels\ intensity - outer\ pixels\ intensity) \right| \tag{4.2}$$

The lesion is subsequently divided into 8 equi-angle slices (Figure 4.5d), and, for each, an average value of the gradient is computed, as in [139].



(a) Equidistant Contour Points (b) Normal Direction at each Point (c) Contour Point and Normal Neighbor Pixels (d) Division of Lesion into Octants

Figure 4.5: Illustration of the border extraction pipeline.

Color

The extraction of color features plays a significant role in distinguishing between MM lesions, which often contain more than two colors, and benign lesions which tend to be generally uniform in color.

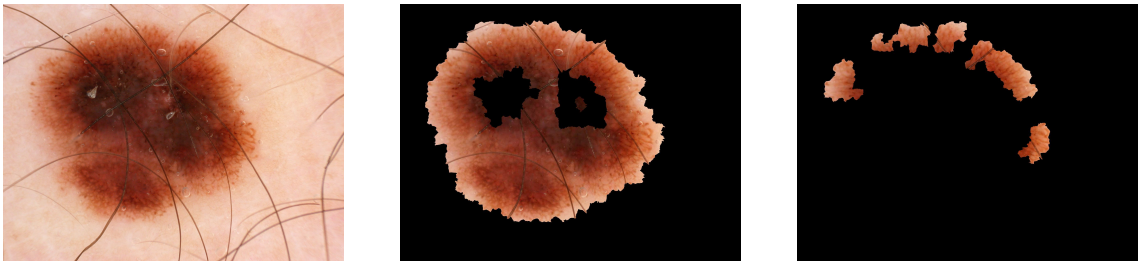
Prevalent statistical measures for characterization of the color distribution are the average, standard deviation, skewness, maximum, minimum in at least one color space [140].

With a methodology based in the Texture-Colour-Geometry Feature Extraction (TCGFE) library [141, 142] developed at Fh-AICOS Portugal, we proceed to the extraction of the average, standard deviation and skewness of the red, green, and blue components of the lesion.

Dermoscopic Structures

There are five dermoscopic structures identified by the ABCD rule: structureless areas, dots, globules, streaks and pigment network, the latter being the most thoroughly analysed.

Relying on manual segmentations of pigment network and streaks regions provided by the ISIC 2017 Challenge (Figures 4.6b and 4.6c), one can extract features such as the percentage of the lesion occupied by each criteria.



(a) Original Image (b) Pigment Network (c) Streaks

Figure 4.6: Example of image pixels with dermoscopic structures.

Furthermore, motivated by the specific visual pattern of each of these structures, descriptors which characterize the texture of a lesion, particularly the existence of repeated visual patterns, are also extracted. The GLCM is computed over the grayscale image for the estimation of the joint probability of two pixels that are separated by a fixed distance [140]. By employing the GLCM descriptor, we estimate the following statistical measures: homogeneity and correlation.

4.3 ABCD Rule-Inspired Neural Network

4.3.1 Model Architecture

With the aim to study a simple ML approach inspired by the ABCD rule of dermoscopy for skin lesion classification, a MLP classifier is initialized with two hidden layers of 16 neurons each activated by a ReLU function and followed by dropout regularization layers (rate of 20%), which results in neurons being randomly omitted at each epoch. From the ABCD features described in Section 4.2, a total of 25 features were generated and used as input for the neural network. The output of the final layer is passed to a softmax function to obtain a distribution over 3 classes. The network is represented in Figure 4.7.

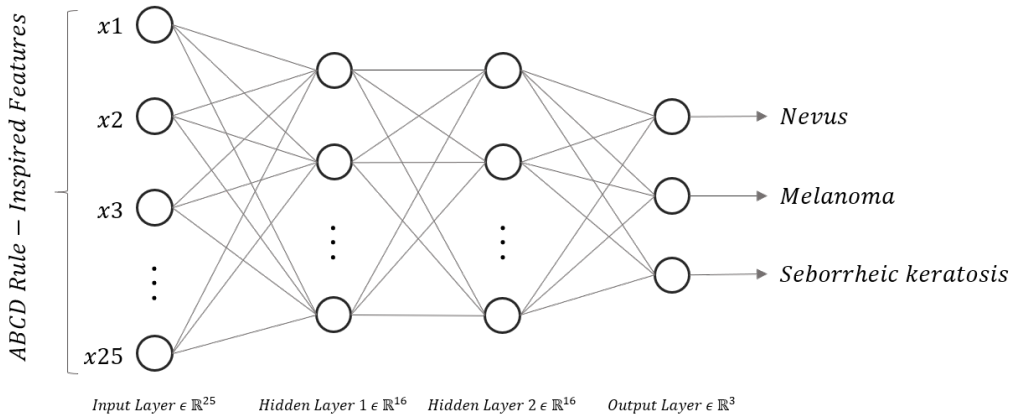


Figure 4.7: Architecture of the ABCD rule inspired neural network for skin lesion classification.

4.3.2 Training Strategy

Normalized ABCD rule-inspired features are used as input to the MLP classifier represented in Figure 4.7. For the training protocol, a batch size of 3 and the Adam optimization algorithm with a learning rate (LR) of 1e-3 are employed.

A loss function is of extreme importance in ML: it determines the distance between the model's current output and its expected output, therefore guiding the training of the model. Categorical cross entropy (CCE) (Equation 4.3) is the loss function employed in this work, since it leads to better generalization models and faster training [143]. By computing the average loss for validation data, one can verify if the model is generalising or overfitting.

$$Cross\ Entropy = -\sum_{i=1}^{K} y_i \cdot log(\hat{y}_i) \tag{4.3}$$

with K equaling to the number of scalar values in the model output, $\hat{y}_i$ being the i-th predicted value and y_i the corresponding true value.

All experiments are trained for a consistent number of epochs: 130; nevertheless, as too many epochs can lead to overfitting of the training dataset, early stopping is utilized to halt training when the validation loss no longer yields an improvement after 20 epochs.

4.4 CNN for Skin Lesion Classification

CNNs are widely used in automatic image classification systems, outperforming classic systems and even surpassing the documented human performance on ImageNet [112]. To overcome the limitation of the size of publicly available skin lesion datasets and following the trend in the field of skin lesion diagnosis, a transfer learning scheme is implemented. By employing a pre-trained model, it is possible to take advantage of its previous knowledge while retraining it for the new task, i.e., fine-tuning. The EfficientNet-B3 architecture with weights pre-trained on ImageNet is used in this work.

4.4.1 Model Architecture

EfficientNet [144] is a family of CNNs created using a scaling ratio of depth (number of layers), width (number of channels per layer) and resolution (input image size), which vary depending on the variant of the CNN selected (EfficientNet-B0 to EfficientNet-B7). This architecture focuses on both accuracy and efficiency, as the name indicates, and is able to achieve state-of-the-art results while being multiple times smaller and faster [144], as represented in Figure 4.8.

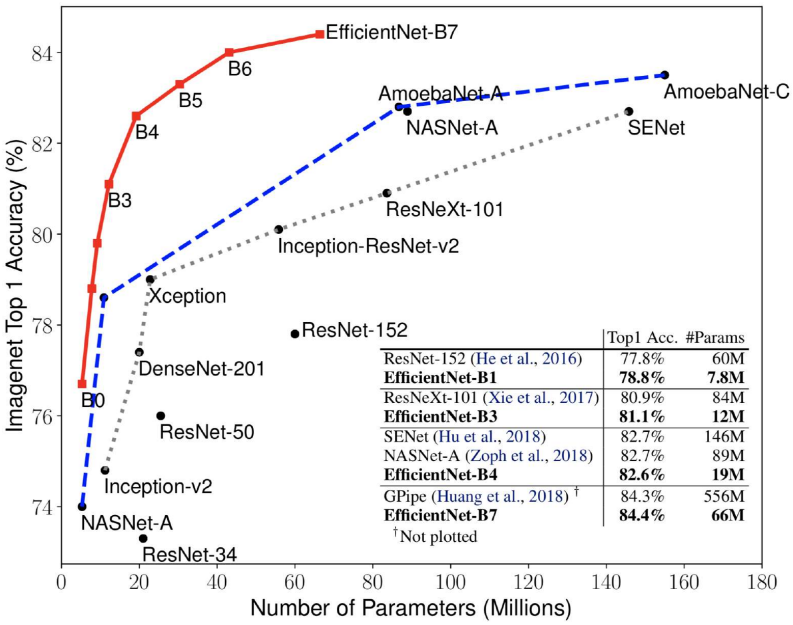


Figure 4.8: EfficientNet VS other CNNs on the ImageNet dataset (taken from [144]).

EfficientNet-B3, designed to receive RGB images with dimensions of (300×300) as input, is used as the backbone of all experiments performed in this dissertation, with the exception of the two-layer neural network presented in Section 4.3.1. Variant B3 is chosen because of the balance when comparing the number of parameters required and accuracy achieved with other CNNs. The input data must range [0, 255] as rescaling and normalization are included as part of the model.

Although other pre-trained CNNs such as VGG [145], ResNet [146], Inception [118], could be considered for this problem, the main focus of the work is the investigation of the multi-tasking and multimodal approaches and their impact on the results, independently of the model structure. Therefore, such methodologies can be applicable with any other architecture.

4.4.2 Training Strategy

To optimise performance, the following training strategy was employed. Each dermoscopic image is resized to (300×300) pixels in order to make it compatible with the original dimensions of the EfficientNet-B3 and leverage the natural-image features learned by the ImageNet pre-trained network.

To perform a fair comparison, all models were set with the same hyperparameters. Each CNN model is compiled using the Adam optimization algorithm [147] and the loss function is specified to be CCE (Equation 4.3).

To accommodate EfficientNet-B3 for the desired tasks, the classification layers are replaced with specific ones for prediction of specific tasks related to skin lesion diagnosis. A frozen layer approach [10, 148] is adopted to ensure the best performances, avoiding destroying any of the information that the pre-trained layers contain. Newly added layers are initialized randomly and the weights associated with them are changed until they converge, with the LR initially set to 1e-3. Because the skin lesion dataset is quite different from ImageNet and to push for better performance, we unfreeze an entire sub-block of the architecture at a time and fit the model for 5 epochs with a very low LR (1e-6), saving it after this number of epochs. This procedure is repeated until all blocks are unfrozen. We have models with batch size of 3 or 6, depending if data balancing techniques are being applied. Each model is trained for 130 epochs.

Data Augmentation

The amount of data being used on any ML process has a significant impact on its success: labeled data is scarce in the field of medical imaging. Publicly available skin lesion datasets are small and this number of samples may not be sufficient for an adequate performance in our DL models. To address such problem, image data augmentation [149, 150, 151] is applied.

Online data augmentation (transformations applied during training) is performed in the DL models, ensuring that these receive new variations of the images at each epoch and therefore being a method to reduce overfitting. The following transformations are employed:

- random clockwise rotation of 5 degrees;
- horizontal flips and shifts up to 20% of the image size;
- zoom between 80% (zoom in) and 130% (zoom out);
- brightness adjustment between 0.2 and 0.8 (note that values below 1 result in darkening the image; above 1, it is the opposite).

Augmentation techniques are only applied in the training set, as it is the learning data; validation and test sets were not augmented for comparison purposes.

4.5 Evaluation Metrics

The evaluation metrics used to rate the performance of the models proposed in this dissertation are in line with the ones used in the ISIC 2017 challenge.

The effectiveness of the algorithm is validated through the computation of the loss on the provided validation set. The performance of the different frameworks presented in this dissertation is evaluated on the test set of 600 samples (considering the distribution provided by ISIC 2017 challenge), which remains unused during training.

For a multi-class classification problem, metrics are applied to each label independently to get a per-class metric or are averaged out across all classes. Moreover, each sample is given a predicted class according to the maximum predicted probability amongst all cases, rather than using a specific threshold.

The classification score consists of True Positives (TP), False Positives (FP), True Negatives (TN) or False Negatives (FN):

- # TP - Number of samples correctly diagnosed as a specific lesion
- # FP - Number of samples wrongly diagnosed as a specific lesion
- # TN - Number of samples correctly diagnosed as not being a specific lesion
- # FN - Number of samples wrongly diagnosed as not being a specific lesion

These measures can be organized under the concept of a confusion matrix. A confusion matrix, used as foundation of other metrics, allows a tabular evaluation of the performance of a classification model, by comparing the actual target values with those predicted by the network. Matrices can be normalized through the division of each entry by the total number of true samples in the class. Given that the dataset used in this dissertation is imbalanced (Figure 4.3), a normalized version of the confusion matrix will be used since the unnormalized matrix does not consider the proportion of the total class size which is predicted correctly, possibly leading to improper conclusions.

Accuracy (ACC) is the most commonly used metric. Nonetheless, it is not advisable to use it as the main metric when there is a high class imbalance, since the model can correctly predict majority class samples while classifying incorrectly samples from minority classes and still have considerably high accuracy.

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{4.4}$$

As such, balanced multi-class accuracy (BMA), the macro-average of the per-class recall (also known as sensitivity or true positive rate - TPR), is the primary metric considered in the ISIC Challenge.

$$BMA = \frac{1}{C} \cdot \sum_{i=0}^{C} TPR_i \tag{4.5}$$

Sensitivity (SE) is the percentage of true positives that are correctly identified:

$$SE = \frac{TP}{TP + FN} \tag{4.6}$$

Specificity (SP), also called true negative rate (TNR), measures the proportion of true negatives classified as negative:

$$SP = \frac{TN}{FP + TN} \tag{4.7}$$

Precision (or positive predictive value - PPV) is the ratio between the number of correctly predicted cases and the total number of positives.

$$PPV = \frac{TP}{TP+FP} \tag{4.8}$$

Contrarily, negative predictive value (NPV) is the percentage of correctly predicted negative cases.

$$NPV = \frac{TN}{TN+FN} \tag{4.9}$$

Receiver Operating Characteristic (ROC) curve and Area Under the Receiver Operating Characteristic (AUC) are obtained for each class considering a binary classification of each vs. all remaining classes. AUC is a representation of the TPR with respect to the false positive rate (FPR), equal to $1 - Specificity$, at various threshold settings, where a score of 0.5 indicates a random classifier and $AUC = 1$ denotes a perfect classification.

Chapter 5

Experiments

This chapter reports the experiments implemented to assess the impact of different ML methodologies in the automatic skin lesion analysis problem.

Firstly, a classical CAD system with hand-crafted features based on the ABCD rule of dermoscopy faces the standard in image classification in current days: a fine-tuned pre-trained CNN. A multi-task model with dermoscopic feature classification as auxiliary tasks is proposed and optimized through a number of class balancing techniques. The inclusion of a segmentation step in the framework before inputting the images to the CNN is also discussed. The hand-crafted features extracted from the dermoscopic images are used as auxiliary data for multi-tasking prediction and, finally, used as additional input for multimodal learning.

5.1 Hand-Crafted versus Deep Learning Generated Features

To assess the relevance of hand-crafted features for experiments described below, an approach to automatically classify skin lesions based on the ABCD rule is firstly proposed. In addition, such ML model is compared to a transfer learning approach in order to evaluate if pre-trained models and their automatically generated features are superior in skin lesion classification problems.

The standard pipeline in automatic dermoscopic image analysis (Figure 3.2) is composed by three main stages: image segmentation, feature extraction and lesion classification. Whereas segmentation masks are provided by the dataset used in this work, annotated features are not provided. In this work, besides the findings regarding multi-tasking and multimodal training, methods to extract features related to the ABCD rule and use them as auxiliary data are proposed.

Asymmetry, border, color, and dermoscopic structures descriptors are extracted through methodologies delineated in Section 4.2 and are grouped together into a single feature vector, with 25 categories. Normalization is then employed through *MinMaxScaler* so that the features values remain in range [0,1], thus preventing characteristics with greater intervals from having a bigger influence in model fitting. The selected normalized characteristics are used as input data for the MLP classifier (Figure 4.7) which classifies skin lesions into three classes: NV, MM or SK. The framework followed is visualized in Figure 5.1.

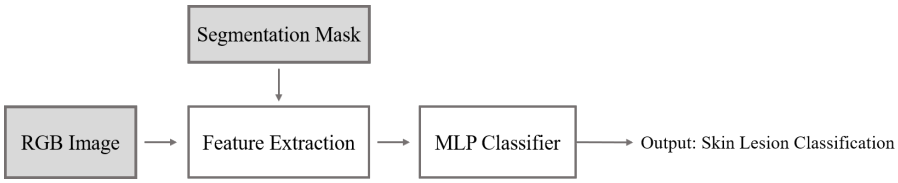


Figure 5.1: Diagram of the neural network model for skin lesion classification.

As reported in Section 3.5.2, current state of the art in image classification show that transfer learning is more effective than training a model from scratch. Accordingly, transfer learning in a DL based approach must also be explored.

An EfficientNet-B3 architecture with the weights and biases of the network pre-trained on the ImageNet dataset is initialized. A detailed explanation of this CNN choice is presented in Section 4.4.1. A global average pooling layer is introduced on top of the frozen base network (EfficientNet-B3) to reduce the number of parameters for the classifier, followed by batch normalization and dropout (rate of 0.2) layers as regularization to prevent overfitting. The original pre-trained model's classifier is replaced by a softmax layer with x neurons (x being 3 which is the number of lesion classes) to translate each of the class's probabilities. A visual representation is provided in Figure 5.2.

This system is trained end-to-end from RGB image pixels and labels annotated by experts.

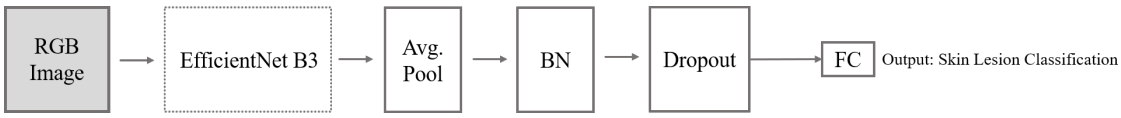


Figure 5.2: Diagram of the baseline model for skin lesion classification. 'Avg. Pool' denotes the average pooling layer, 'BN' is Batch Normalization, 'FC' is fully connected layer.

5.2 Multi-Task Learning

In this experiment, we propose to jointly optimize several tasks: skin lesion classification (the main focus of our work), with one or more related auxiliary tasks. Multi-tasking is performed with the intention of biasing the model towards more meaningful features. With proposed auxiliary tasks closely related to the main task, learning them likely allows the model to learn beneficial representations and focus attention on parts of the image that the baseline network could possibly ignore.

Skin Lesion and Dermoscopic Features Classification

The baseline model is altered after the dropout layer with the inclusion of new softmax layers, one for each structure (Figure 5.3), with the number of neurons equaling the number of classes per task. A hard parameter sharing approach, i.e., sharing the hidden layers between all tasks while keeping separate task-specific output layers, is adopted.

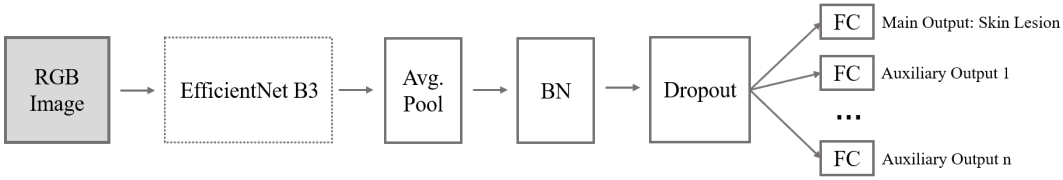


Figure 5.3: Diagram of the multi-task model. 'Avg. Pool' denotes the average pooling layer, 'BN' is Batch Normalization, 'FC' is fully connected layer.

The first auxiliary tasks tested are the classification of the presence of clinical dermoscopic features. The motivation behind this design is to make use of the correlation between the two: for example, milia like cyst structures are usually indicative of SK [152] whereas negative network has high SP for MM [153]. One can avail the superpixel annotations from the dataset to infer the existence of the aforementioned dermoscopic structures and, consequently, label the data in different classes.

Table 5.1: Class distribution of dermoscopic structures in the training set.

(a) Presence Labels

Structure	Class	# Samples
Pigment Network	Absent	869
	Present	1131
Negative Network	Absent	1874
	Present	126
Milia like Cyst	Absent	1429
	Present	571
Streaks	Absent	1884
	Present	116

(b) Multi-class labels (r is the percentage of lesion occupied by the dermoscopic feature)

Structure	Class	Class Interval	# Samples
Pigment Network	0	$r = 0$	869
	1	$0 < r \leq 0.04$	275
	2	$0.04 < r \leq 0.1$	278
	3	$0.1 < r \leq 0.2$	296
	4	$r > 0.2$	282
Negative Network	0	$r = 0$	1874
	1	$0 < r \leq 0.035$	63
	2	$r > 0.035$	63
Milia like Cyst	0	$r = 0$	1429
	1	$0 < r \leq 0.01$	179
	2	$0.01 < r \leq 0.03$	196
	3	$r > 0.03$	196
Streaks	0	$r = 0$	1884
	1	$0 < r \leq 0.025$	55
	2	$r > 0.025$	61

Hence, the first multi-task model includes 5 outputs: the main being skin lesion classification and 4 auxiliary tasks for pigment network, negative network, milia like cyst and streaks classification.

Two strategies are experimented. The first uses binary classification in the auxiliary tasks as we are only interested in predicting if each structure is present or absent; in the second, the percentage of the total lesion that contains each feature is used and, after analysing the distribution of the percentages, the former positive class is divided and labelled in well distributed degrees of presence, i.e, classes with approximately the same number of cases, as a way to help the model to focus (refer to Table 5.1).

5.3 Optimization of Multi-Task Models with Class Balancing

The skin lesion dataset employed in this work is imbalanced between classes. As a result, class imbalance introduces bias towards the most represented class, compromising the performance of the previously described models, which must be tackled.

Resampling Techniques

A prominent technique for handling imbalance is resampling: by performing undersampling, samples from the majority class are removed, whereas in oversampling, examples from the minority class are duplicated. The latter is applied in this dissertation since it is widely used and proven to be robust [154], and, unlike undersampling, it does not discard a portion of available data, which is extremely important given the small size of the dataset.

The first methodology lies in the duplication of MM and SK samples so that there is the same number of disease samples per class. Batches contain 3 random skin lesion cases and are arranged in such a way that there is one copy per lesion class.

Thereafter, we carry out another procedure which ensures data frequency [10], i.e., each batch always holds at least 1 positive sample of every skin lesion and dermoscopic feature, resulting in a batch size of 6. Because the number of cases is not equal for all classes, after picking all possible choices from one of the labels, the set must be restarted. Additionally, since the category labels are not mutually exclusive, it is important that the same sample is not represented twice in the same batch; we secure that by removing each case from the set after picking it.

By guaranteeing there exists one case of each unique label, model weights will be updated based on all the unique labels in each gradient descent step [10]. Nonetheless, while this improves class balance, there is still imbalance as including a case within one category will also include its labels in all other categories. To further address this issue, class weights are used.

Class Weights

Class weighting is also employed for handling the skewed distribution of classes: its purpose is to over-penalize the misclassification of samples from the minority class. The weight W_i of each class i is given by Equation 5.1:

$$W_i = \frac{N}{C \cdot n_i} \tag{5.1}$$

where N is the total number of samples, C is the number of classes, and n_i is the number of samples for class i.

By applying this modification (W_i) in the loss function employed in this work (CCE), a weighted loss function (Equation 5.2) is obtained:

$$Weighted\ Cross\ Entropy = -\sum_{i=1}^{K} W_i \cdot y_i \cdot log(\hat{y}_i) \tag{5.2}$$

The different weight values will influence backpropagation during the training phase. In practical terms, the added loss from a misclassified sample of MM will have a bigger impact than a misclassified benign NV.

The difference between calculating class weights based on the distribution of the training set prior to training and finding the dynamic weights, i.e., computing according to the dispersal of classes in each batch is examined.

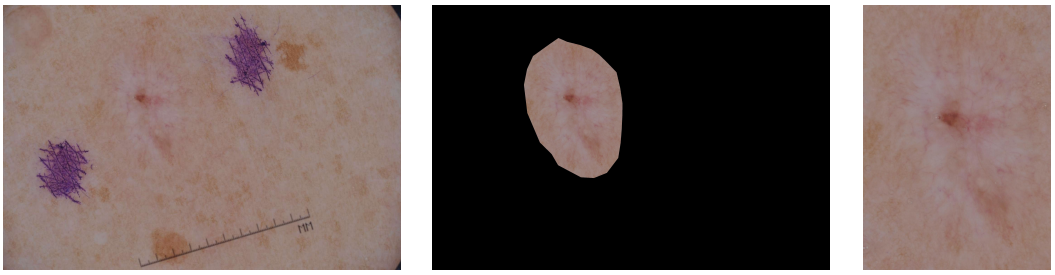
After applying the oversampling and class weighting techniques to the training data, their impact in the multi-task model, described in Section 5.2, is tested and compared against passing raw data, as described in the same section.

5.4 Assessment of Segmentation Impact in Skin Lesion Analysis

Segmentation can be applied as a preprocessing method in a skin lesion classification pipeline, aiming to remove background noise and/or artifacts such as hair, ruler markings and non-target lesions, which could deceive the classifier.

The goal of this investigation is to understand the role that segmentation plays in classification performance and whether removing the pixel intensities outside the target lesion is advantageous or not. The performances of 3 identical models which receive different inputs are assessed. Such models are optimised through data oversampling and computation of dynamic weights. The inputs, which can be observed in Figure 5.4, are:

- unaltered skin lesion images;
- segmented images with no background information, created through a bitwise AND operation using the original image and its corresponding binary mask;
- images cropped around the lesion, with a bounding box obtained from the binary mask dimensions.



(a) Original Image. (b) Segmented Image (c) Cropped Image

Figure 5.4: Examples of the inputs for the multi-task model.

5.5 Multi-Task and ABCD Rule Criteria Classification

We further explore the multi-task paradigm and how parallel tasks can improve the focus of a model by adding more auxiliary classification outputs to the top performing version proposed in previous Section 5.3. For a supervised classification in the new tasks and because the dataset of this work does not contain expert annotations of ABCD rule criteria, we are required to proceed to data labelling.

Dermatologists assess semi-quantitatively the ABCD rule traits. Considering asymmetry, skin lesions are divided in three levels: fully symmetric (0), asymmetric on one axis (1) and asymmetric on two axes (2). For the definition of such classes given the asymmetry ratio calculated in Section 4.2, a threshold, T_0, is set in such a way that more than half of MMs have $A_{score} = 2$ and approximately 60% of benign lesions are scored 0 or 1 (note that MMs are typically asymmetrical whereas both halves of benign lesions usually match) . Hence, if the non overlapping area exceeds 6% of the lesion area, the lesion is considered asymmetrical in that axis.

$$A_{score} = \begin{cases} 0 \rightarrow \text{fully symmetric}, & \text{if } A_x \leqslant T_0 \textit{ and } A_y \leqslant T_0 \\ 1 \rightarrow \text{asymmetric on one axis}, & \text{if } A_x \leqslant T_0 \textit{ or } A_y \leqslant T_0 \\ 2 \rightarrow \text{asymmetric on two axes}, & \text{otherwise} \end{cases} \tag{5.3}$$

Regarding the border, an abrupt cut-off of the peripheral region of the lesion sets a particular octant to a score of 1; otherwise, it is scored as 0. Therefore, we set a threshold value of 50 for the gradient which classifies the transition into either soft (0) or abrupt (1). This procedure is executed for all the divisions, resulting in a final B_{score} between 0 and 8.

Only asymmetry and border information are considered in three separate trials:

- Addition of one task: asymmetry classification - 3 classes;
- Addition of one task: border classification - 9 classes;
- Addition of the above tasks.

5.6 Multimodal Multi-Tasking

The objective of this technique is to provide the network with readily available information tailored to what physicians use (ABCD rule) and thus are characteristics proven to be relevant for malignancy classification. Through the usage of these appropriate features, we aim to decrease the amount of overfitting, therefore enhancing the performance of CNN-based algorithms. Furthermore, they could potentially be useful to help explain the diagnosis prediction.

Early Fusion

We explore pixel-level image fusion, whose goal is to generate a composite image from multiple inputs containing complementary information [155], by combining the original dermoscopic images with the corresponding expert traced segmentation masks prior to introducing this information into the best performing multi-task model, as represented in Figure 5.5.

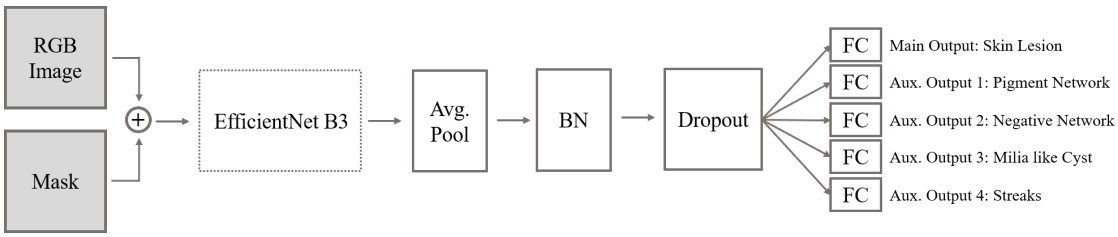


Figure 5.5: Diagram of the multimodal multi-tasking model with pixel-level image fusion. 'BN' is Batch Normalization, 'FC' is fully connected layer.

As in previous experiments, the backbone model chosen is EfficientNet-B3; this pre-trained model expects the input dimensions of the new problem to be equal to the dimensions of the previous of the old task: (300×300) for the height and width, and 3 channels for RGB components, as detailed in Section 4.4.1. Therefore, a challenge arises: the network must be modified to take an image with 4 channels as the third dimension.

Changing the number of channels affects the dimensions of the pre-trained weights of the CNN: the preprocessing normalization layer only holds the mean and variance of each RGB channel and the weight dimensions of a convolutional layer are determined by the input and output depths. To deal with the first, we set the state of the fourth channel of the layer by exposing it to the masks of the training data. For the latter, the weight dimensions of the first convolutional layer are expanded and the fourth value is set to be the mean of the pre-trained RGB weights.

Late Fusion

The purpose of this investigation is to analyse the impact of features extracted using computer vision or ML techniques in the prediction of the classifier.

To perform late feature fusion, the multi-task models are set to receive two inputs: the dermoscopic images and the characteristics of skin lesion that doctors look for when diagnosing and

classifying MMs. Hence, the aforementioned hand-crafted ABCD rule features are directly concatenated with the feature vector obtained by the EfficientNet-B3. Figure 5.6 provides a representation of this framework.

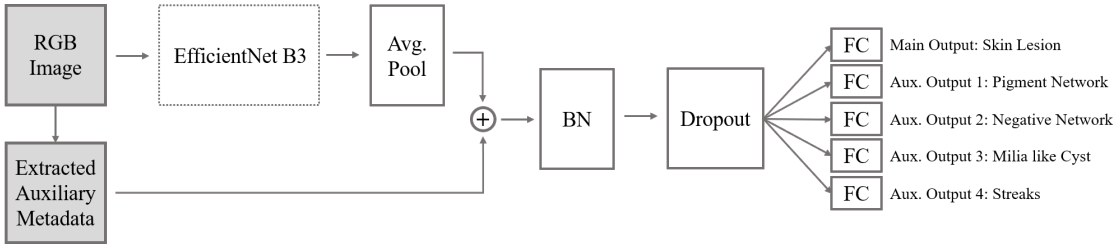


Figure 5.6: Diagram of the multimodal multi-tasking model with late feature fusion. 'BN' is Batch Normalization, 'FC' is fully connected layer.

Not all features might be beneficial to the classification task, some can possibly weaken the performance of the classifier. Consequently, different cases are tested: including only asymmetry or border information, or using both simultaneously.

Chapter 6

Results and Discussion

The results of the aforementioned experiments are evaluated, critically analysed and discussed throughout this chapter. The best performing ISIC challenge submissions are presented and compared to the solutions proposed in this work. A reflection about relevant information found in the results is also provided, in addition to the clinical applicability, limitations and opportunities for further development of this study.

For the sake of length, this chapter's result tables are limited to AUC, SE, SP and BMA as evaluation metrics; additional metrics for each experiment are found in Appendix A.

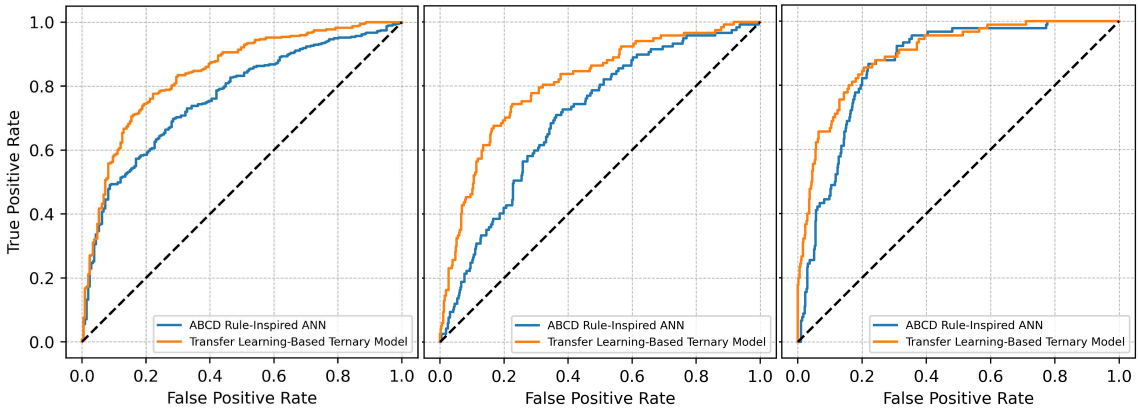
6.1 Hand-Crafted versus Deep Learning Generated Features

The proposed ML approach, based on the classical ML pipeline, to evaluate the success of ABCD rule inspired features in skin lesion diagnosis is tested on the ISIC 2017 Challenge database. Extracted features from all images are used as an input to the MLP classifier, predicting three possible outcomes: NV, MM and SK.

CNNs are a specific type of neural networks, as MLPs, particularly suited for image classification problems. These networks act as automatic feature extractors and preserve spatial interaction. Given the requirements for big amounts of labeled data in DL classification problems opposed to the small size of medical datasets, transfer learning procedures are the current start of the art in automatic skin lesion analysis. To make use of the advantages of this methodology, a model for this problematic built on EfficientNet-B3 is tested and compared to the preceding CAD system.

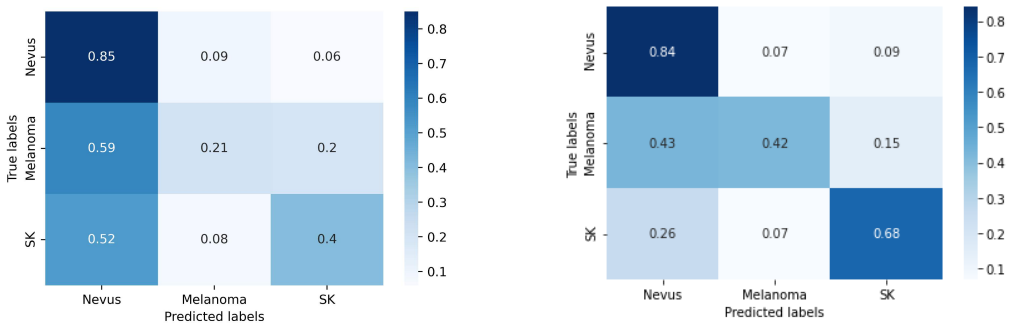
ROC curves comparing the performance of the above-mentioned models in each skin lesion class and normalized confusion matrices are shown in Figures 6.1 and 6.2, respectively.

Through analysis of the ROC curves, one can observe the improvement in MM classification with DL generated features. The normalized confusion matrix of the ML model denotes a strong bias of the algorithm towards NV. Around 70% of the dataset images are classified as NV, when only 55% are in fact, and more than 80% of MM are being treated as a benign lesion which is potentially dangerous.



(a) Nevus. (b) Melanoma. (c) Seborrheic keratosis.

Figure 6.1: Comparison between ROC curves of the models with hand-crafted and deep learned features.



(a) ABCD rule-inspired NN (b) Baseline transfer-learning

Figure 6.2: Normalized confusion matrices for the models with hand-crafted and deep learned features.

Table 6.1: Results of the models with hand-crafted and deep learned features (bold values highlight the best result for each metric).

Model	**AUC**			**SE**			**SP**			**BMA**
	Nv	MM	SK	Nv	MM	SK	Nv	MM	SK	
ABCD rule neural network	0,75	0,72	0,87	**0,85**	0,21	0,40	0,44	0,92	**0,91**	0,49
Transfer learning model	**0,84**	**0,81**	**0,90**	0,84	**0,42**	**0,68**	**0,65**	**0,93**	0,90	**0,65**

Table 6.1 confirms the intuition: pre-trained CNNs significantly outperform traditional ML approaches. MM SE of the transfer-learning model is twice the value obtained for the ML model and SK SE also denotes a major improvement (from 0.40 to 0.68). In terms of SP, results are similar except for NV where the CNN model outperforms the other once again.

This experiment is comparing the hand-crafted features against the ones generated automatically by a DNN, allowing to confirm the superiority of the latter. Over the last decade DL, particularly CNNs, have become the standard technique in most computer vision systems. CNNs combine the benefits obtained by ANNs, such as the MLP of this work, and, additionally, take

advantage of spatial information. Thus, although the hand-crafted features used as input in the MLP classifier attempted to mimic clinical diagnosis procedure as they are inspired by the ABCD rule, deep learned features outperform them and are more effective at solving the problem of skin lesion classification.

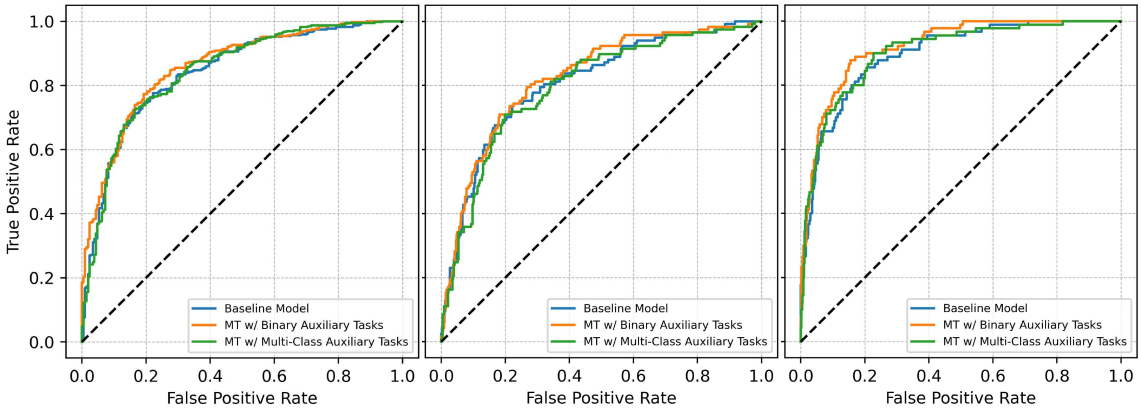
This experiment establishes a baseline that the following models are built upon on.

6.2 Multi-Task Learning

Skin Lesion and Dermoscopic Features Classification

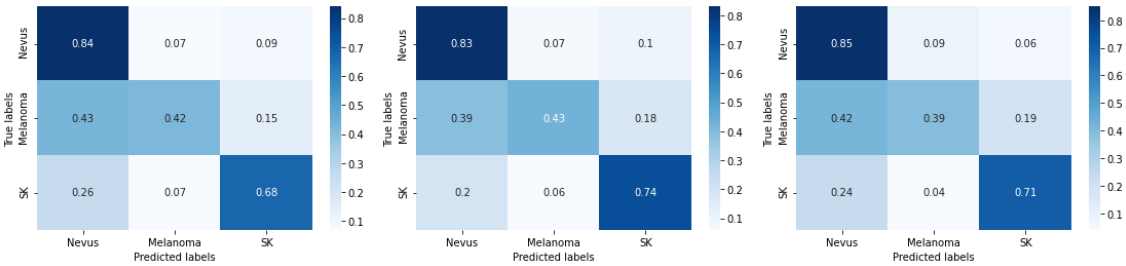
A multi-task network is compared against the baseline lesion classification model. By sharing representations between related tasks, the intention of this paradigm is to enable the model to generalize better on the main task [122].

Figure 6.3 compares the ROC curves of the baseline and the multi-task models for each skin lesion class. Normalized confusion matrices for the same models are represented in Figure 6.4.



(a) Nevus. (b) Melanoma. (c) Seborrheic keratosis.

Figure 6.3: Comparison between ROC curves of multi-task models with dermoscopic features classification as auxiliary tasks.



(a) Baseline model (b) Multi-task with binary auxiliary tasks (c) Multi-task with multi-class auxiliary tasks

Figure 6.4: Normalized confusion matrices for the multi-task models with dermoscopic features classification as auxiliary tasks.

The ROC curves (Figure 6.3) suggest that the preceding models achieve very similar results, with a slight advantage for the multi-task model with binary classification of the dermoscopic features in all lesion classes. Matrices 6.4b and 6.4c show that the addition of binary auxiliary tasks increases the number of correctly predicted MM and SK. On the other hand, multi-class prediction of dermoscopic features is worsening MM classification.

Table 6.2: Results of the multi-task models, along with the results from the baseline multi-class model (bold values highlight the best result for each metric).

Model	**AUC**			**SE**			**SP**			**BMA**
	Nv	MM	SK	Nv	MM	SK	Nv	MM	SK	
Baseline transfer-learning	0,84	0,81	0,90	0,84	0,42	0,68	0,65	**0.93**	0,90	0,65
Multi-task with binary dermoscopic features classification as auxiliary tasks	**0,86**	**0,82**	**0,92**	0,83	**0,43**	**0,74**	**0,69**	**0,93**	0.88	**0,67**
Multi-task with multi-class dermoscopic features classification as auxiliary tasks	0,84	0,80	0,90	**0,85**	0,39	0,71	0,66	0,92	**0,91**	0,65

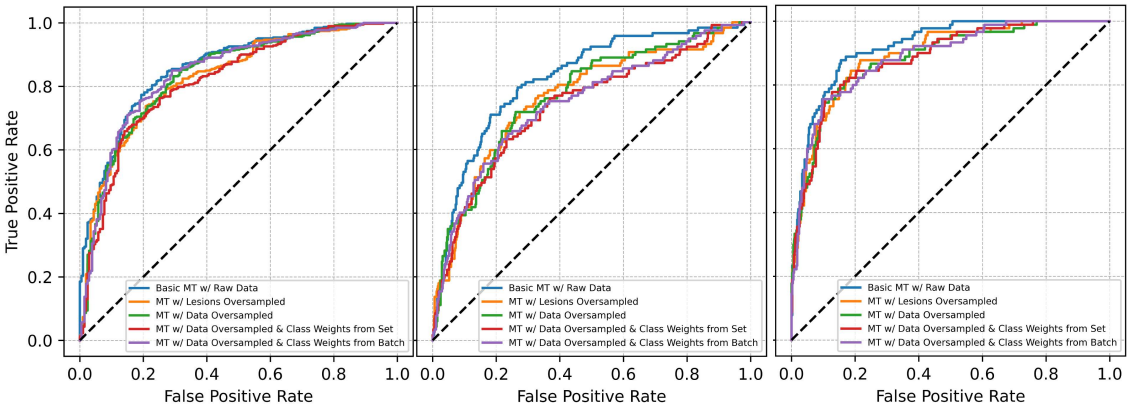
Through an attentive analysis of Table 6.2, one can verify that the MTL algorithm which predicted absence or presence of the dermoscopic structures exceeded the performance of the baseline model in all metrics of interest, therefore being considered in further steps of the framework. Contrarily, the multi-task model which graded the presence of features in several levels had a poorer performance, only marginally improving NV SE and SK SP.

6.3 Optimization of Multi-Task Models with Class Balancing

Methodologies presented thus far did not address the problem of the imbalanced classes: inevitably, such models usually present a superior prediction in the majority class in comparison with the two other classes. To overcome this issue, data balancing techniques are applied to the best performing model at this point.

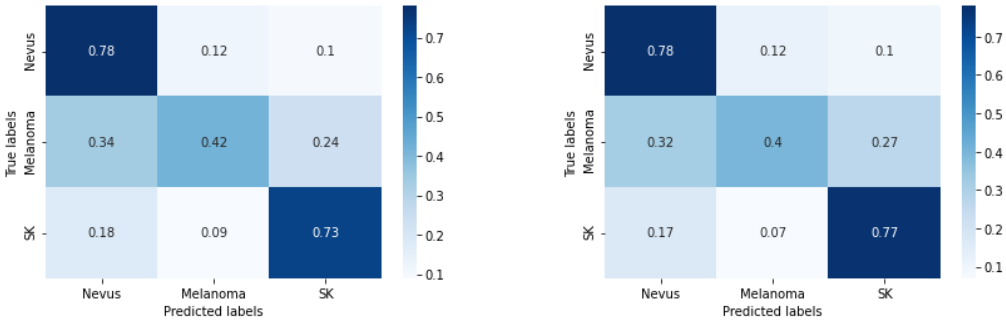
Figure 6.5 shows that when doing the one-vs-rest method, the basic multi-task model with no class balancing strategies outperforms the others.

Contrarily, when considering the three classes, one can verify the positive impact of imbalanced learning techniques in the normalized confusion matrices of Figure 6.6. By assigning higher weight to the minority classes (MM and SK), the model grants more attention to these misclassified samples therefore learning to classify them better. While in the basic multi-task model, most cases are classified as NV, there are many more correct predictions of the outnumbered classes in the optimized multi-task models.



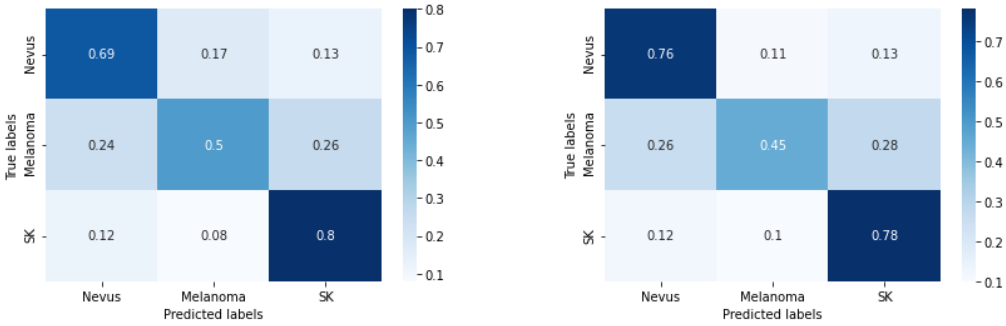
(a) Nevus. (b) Melanoma. (c) Seborrheic keratosis.

Figure 6.5: Comparison between ROC curves of the multi-task models with class balancing techniques.



(a) Lesion oversampling and no weights.

(b) Oversampling and no weights.



(c) Oversampling and weights from set.

(d) Oversampling and weights from batch.

Figure 6.6: Normalized confusion matrices for the multi-task models with class balancing techniques.

The best overall performance, according to BMA scores, is achieved by the baseline multi-task with no class balancing techniques (Table 6.3). Nevertheless, SE of MM and SK as well as SP of NV are significantly improved with the introduction of class weights. When considering the distribution of the training set for the calculation of class weights, half of MMs are correctly classified, which is the best result thus far.

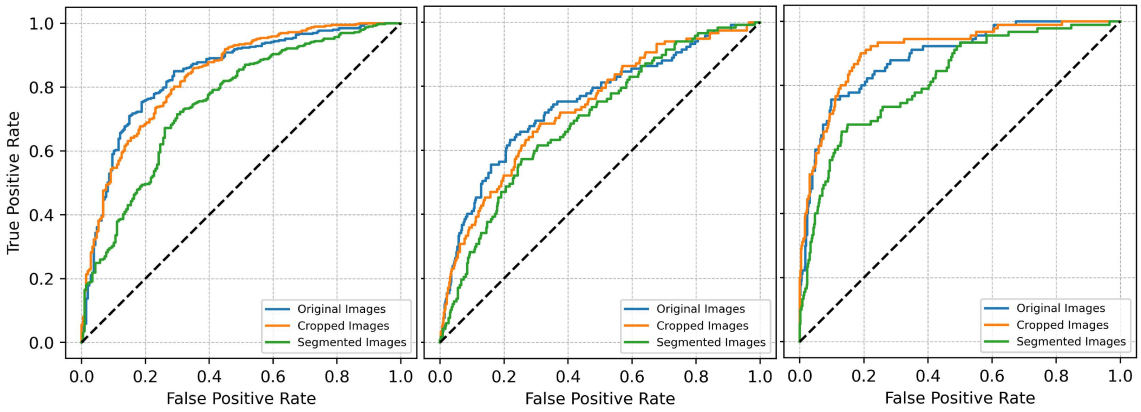
Table 6.3: Results of the multi-task models with data balancing techniques (bold values highlight the best result for each metric).

Model	**AUC**			**SE**			**SP**			**BMA**
	Nv	MM	SK	Nv	MM	SK	Nv	MM	SK	
Basic multi-task with no class balancing techniques	**0,86**	**0,82**	**0,92**	**0,83**	0,43	0,74	0,69	**0,93**	**0,88**	**0,67**
Skin lesions oversampled + no class weights	0,83	0,76	0,89	0,78	0,42	0,73	0,73	0,88	0,87	0,64
Data oversampled + no class weights	0,84	0,77	0,89	0,78	0,40	0,77	0,74	0,89	0,86	0,65
Data oversampled + class weights from training set	0,82	0,74	0,88	0,69	**0,50**	**0,80**	**0,81**	0,84	0,84	0,66
Data oversampled + class weights from batch	0,84	0,75	0,89	0,76	0,45	0,78	0,80	0,89	0,84	0,66

It is important to note that in clinical situations, early diagnosis is of great importance and incorrect classification of a malignant lesion as benign can have dire consequences; hence it could make clinical sense to raise a false positive instead of creating a false negative.

6.4 Assessment of Segmentation Impact in Skin Lesion Analysis

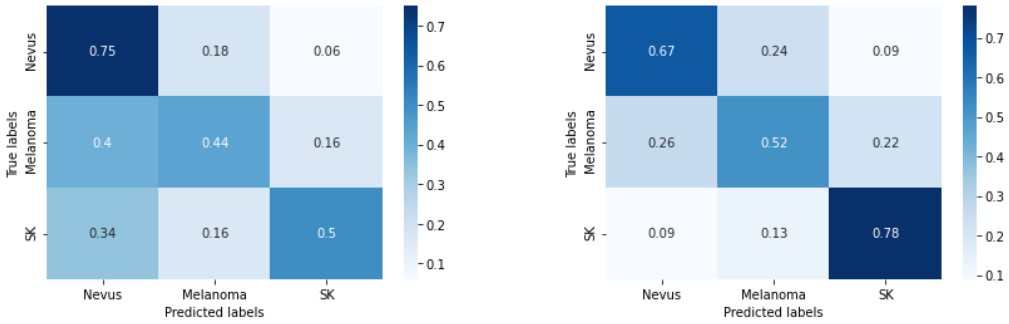
The impact of background information is tested in this experiment, by comparing the performance of the multi-task model optimized with data oversampling and class weights computed according to the composition of each batch, when different versions of the images are used as input.



(a) Nevus. (b) Melanoma. (c) Seborrheic keratosis.

Figure 6.7: Comparison between ROC curves of multi-task model when segmented images and cropped around the lesion images are used as input.

Figure 6.7 demonstrates the poor performance of segmented images (images whose background was removed). The original dermoscopic images surpass the cropped versions in NV and MM classification and the opposite occurs for SK. Matrix of Figure 6.8b shows that more than half of MMs are being correctly predicted.



(a) Segmented images.

(b) Cropped images.

Figure 6.8: Normalized confusion matrices for the multi-task model when segmented images and cropped around the lesion images are used as input.

Using a manually segmented mask to remove background information and pass only the lesion as input to the multi-task model significantly decreases performance, when comparing it with the unaltered input. It is possible that segmented images removed contextual information which could be relevant for the classification task. Results of the original dataset and the crop bounding box images are extremely similar, with the original images providing advantage in the AUC of MM, as seen in Table 6.4, but the cropped images achieving higher MM SE, which could be preferable in a clinical setting.

Table 6.4: Results of the optimized through data balancing multi-task models with modified images as inputs (bold values highlight the best result for each metric).

Model	**AUC**			**SE**			**SP**			**BMA**
	Nv	MM	SK	Nv	MM	SK	Nv	MM	SK	
Original images	**0,84**	**0,75**	0,89	**0,76**	0,45	**0,78**	0,80	**0,89**	0,84	**0,66**
Segmented images	0,75	0,69	0,82	0,75	0,44	0,50	0,62	0,82	**0,91**	0,56
Cropped images	**0,84**	0,73	**0,91**	0,67	**0,52**	**0,78**	**0,82**	0,78	0,88	**0,66**

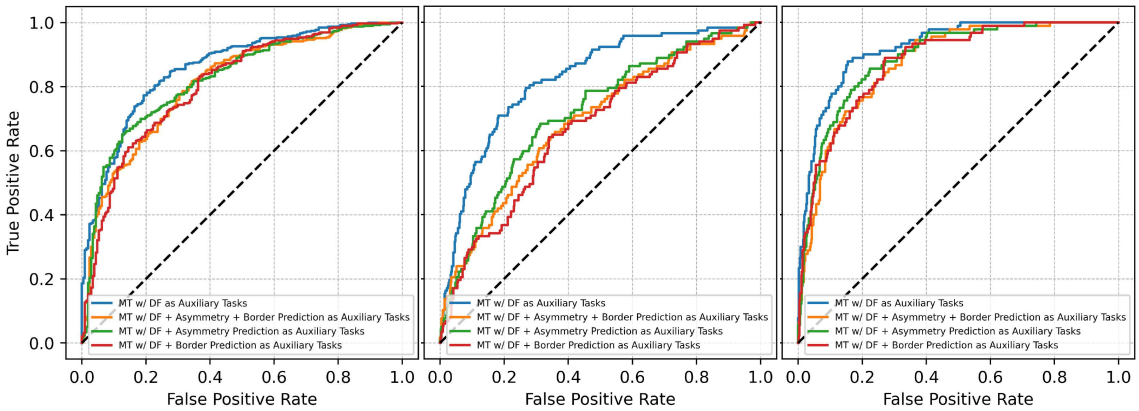
The modification of the input images through the crop of a lesion bounding box results in a bigger number of MM predictions, increasing the number of true positives (SE is 52%) but also doing a worse job at identifying true negatives (SP reduces from 89% in original images to 78% in cropped images). Nonetheless, it is more detrimental not knowing that an individual has cancer than to refer them for additional exams by a dermatologist.

The main conclusion drawn from this experiment is that the network appears to be coarsely focusing on the region of interest, thus meaning background noise does not seem to affect classification; in fact, it may be beneficial. Neighbor pixels surrounding the lesion are important as the difference between background (skin) and foreground (lesion) intensities provides relevant information regarding color and texture variations.

6.5 Multi-Task and ABCD Rule Criteria Classification

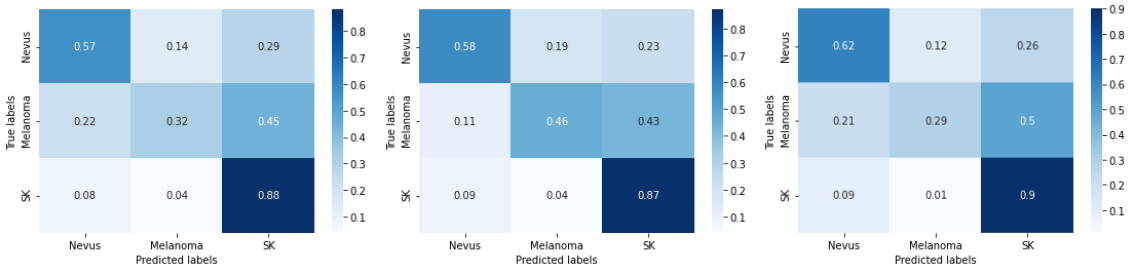
This experiment, besides predicting the five tasks described above, includes the addition of more auxiliary tasks related to the scoring system utilized by dermatologists when differentiating between benign and malignant melanocytic lesions. Labelled training data with asymmetry and border scores is introduced to the multi-task model with oversampled data and class weights (calculated according to the composition of the training set) applied in the loss function. Three versions are studied concerning the insertion of both tasks concurrently or each separately.

By inspecting ROC curves in Figure 6.9, one concludes there is no overall benefit in the inclusion of auxiliary tasks for scoring of the first two ABCD criteria. Normalized confusion matrices of Figure 6.10 present an increase in the percentage of correctly predicted SK. Moreover, they also expose the bias in these models towards SK as well as a deterioration in MM-related metrics.



(a) Nevus. (b) Melanoma. (c) Seborrheic keratosis.

Figure 6.9: Comparison between ROC curves of the multi-task models with ABCD rule criteria classification tasks.



(a) Asymmetry and border as outputs. (b) Asymmetry as output. (c) Border as output.

Figure 6.10: Normalized confusion matrices for the multi-task models with ABCD rule criteria classification tasks.

All metrics of interest are compiled in Table 6.5. Although SK exhibits high SE (87%, 90% and 88% opposed to a value of 80% in the model with no ABCD related auxiliary tasks), its SP is

Table 6.5: Results of the optimized through data balancing multi-task models with ABCD rule related auxiliary tasks (bold values highlight the best result for each metric).

Model	**AUC**			**SE**			**SP**			**BMA**
	Nv	MM	SK	Nv	MM	SK	Nv	MM	SK	
Multi-task with binary dermoscopic features classification as auxiliary tasks	**0,82**	**0,74**	0,88	**0,69**	**0,50**	0,80	0,81	0,84	**0,84**	**0,66**
Multi-task with binary dermoscopic features & multi-class asymmetry prediction as auxiliary tasks	**0,82**	0,71	**0,89**	0,58	0,46	0,87	**0,90**	0,84	0,73	0,64
Multi-task with binary dermoscopic features & multi-class border prediction as auxiliary tasks	0,81	0,67	0,88	0,62	0,29	**0,90**	0,84	**0,90**	0,69	0,60
Multi-task with binary dermoscopic features, multi-class asymmetry & border prediction as auxiliary tasks	0,81	0,69	0,87	0,57	0,32	0,88	0,84	0,88	0,67	0,59

diminished (73%, 69% and 67%, respectively). On the other hand, SE of MM is severely affected by border prediction but it allows to obtain the highest SP value.

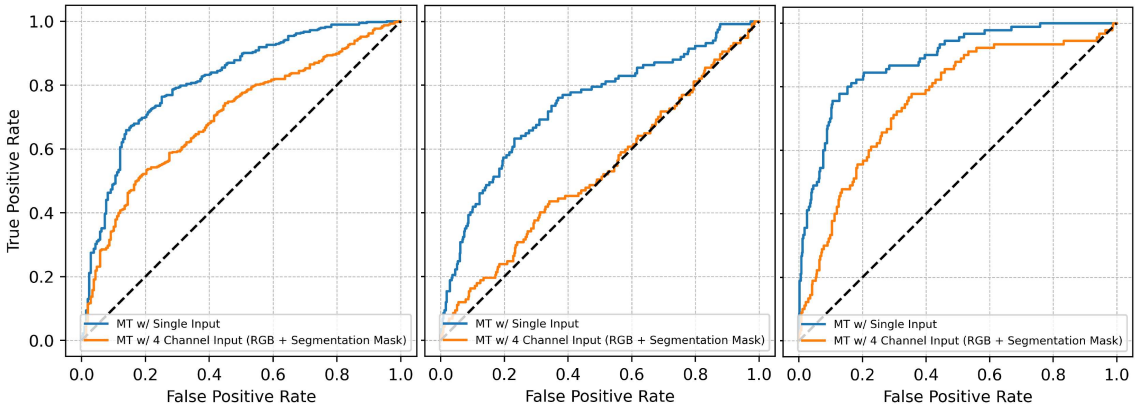
The features extracted automatically for the asymmetry and border classification tasks were expected to be correlated with MM detection and improve the prediction of this class but this does not occur. One can argue that these hand-crafted features are mere surrogates for the true ABCD features and, consequently, the labels that we defined may be imperfect. It is possible that using actual ABCD ratings annotated by dermatologists could yield different results.

6.6 Multimodal Multi-Tasking

Early Fusion

The EfficientNet-B3 architecture was modified to accept a 4-channel input, consisting of the red, green and blue channels of the dermoscopic image and corresponding segmentation mask. The goal of this experiment is to provide the lesion location and investigate if this can enhance the features extracted and, consequently, the prediction.

Through the examination of the ROC curves in Figure 6.11 and normalized matrix in Figure 6.12, one can conclude that having the composite image formed by the dermoscopic snapshot and corresponding expert traced mask as input of the multi-task model reduces considerably its overall performance. Prediction of MM is particularly influenced, decreasing its AUC to 53% which in practice means that the model is making random guesses and has no capacity to distinguish between MM and non-MM samples. As shown in Section 6.4, the CNN already appears to understand which portion of each image corresponds to the lesion; therefore the introduction of the corresponding segmentation mask would not generate beneficial information and a major improvement was not expected. Nevertheless, such a degradation of the performance was also unforeseen. Furthermore, the early fusion approach contributes to the identification of the majority of samples as SK, i.e., it is biased towards this class.



(a) Nevus. (b) Melanoma. (c) Seborrheic keratosis.

Figure 6.11: Comparison between ROC curves of the multimodal multi-task model with pixel-level image fusion.

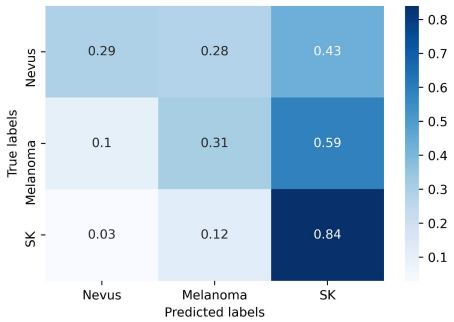


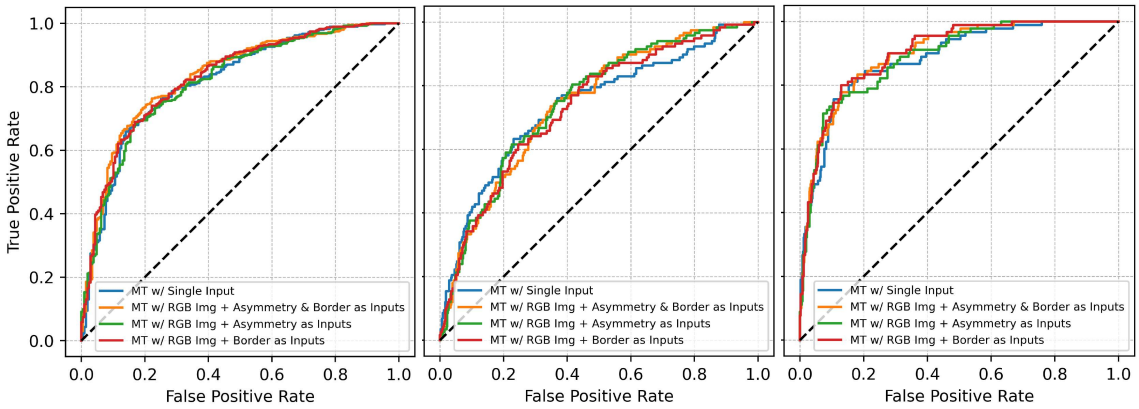
Figure 6.12: Normalized confusion matrix for the multimodal multi-task model with pixel-level image fusion.

Late Fusion

Through late fusion, the objective is to evaluate if the combination of the traditional features inspired by the ABCD rule and CNN features has the ability to boost the classifier performance.

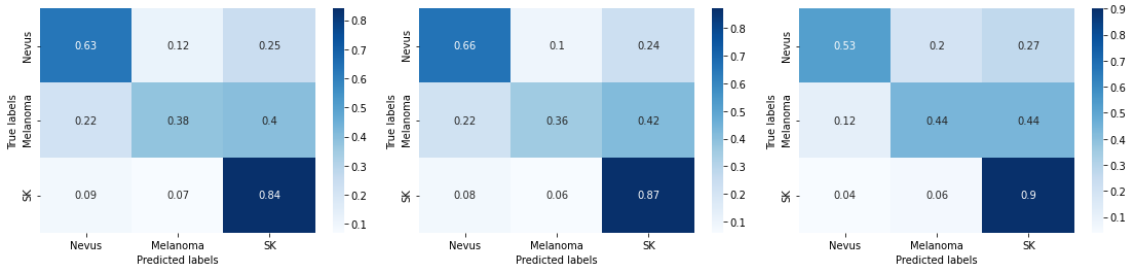
The ROC curves generated for the usage of both criteria simultaneously and one at a time are shown in Figure 6.13. One can infer that the fusion of image and extracted auxiliary metadata reveals a small improvement over the single image modalities. Confusion matrices demonstrated in Figure 6.14 reveal that the addition of both asymmetry and border information positively affects SK diagnosis.

A summary of experimental results of multimodal models is shown in Table 6.6. Pixel-level image fusion only aids the identification of non-NV lesions (SP of 93%). The late-fusion of hand-extracted ABCD rule criteria from the dermoscopic images with DL features extracted by the CNN provides similar results which confirms that the network is able to automatically learn good image representations by itself. However, the AUC values of the multimodal models with late fusion are slightly enhanced: there is an improvement for NV and SK when using the border gradient and the addition of asymmetry benefits MM. Similarly to what we stated in the previous section (6.5),



(a) Nevus. (b) Melanoma. (c) Seborrheic keratosis.

Figure 6.13: Comparison between ROC curves of the multimodal multi-task models with late feature fusion.



(a) RGB image + asymmetry ratio. (b) RGB image + border gradient. (c) RGB image + asymmetry ratio & border gradient.

Figure 6.14: Normalized confusion matrix for the multimodal multi-task models with late feature fusion.

a bias towards SK is detected, hence confirming the strong correlation between this lesion and the first two ABCD criteria. Again, these traits raise the number of MMs missed (SE decreases from 50% to values below 44%) but improve SP and the opposite occurs for SK.

Table 6.6: Results of the optimized through data balancing multi-task models with multiple inputs (bold values highlight the best result for each metric).

Model	**AUC**			**SE**			**SP**			**BMA**
	Nv	MM	SK	Nv	MM	SK	Nv	MM	SK	
Multi-task with single input (RGB image)	0,82	0,74	0,88	**0,69**	**0,50**	0,80	0,81	0,84	**0,84**	**0,66**
Multi-task with 4 channel input (RGB image + segmentation mask)	0,70	0,53	0,76	0,29	0,31	0,84	**0,93**	0,75	0,53	0,48
Multi-task with RGB image + asymmetry ratio as inputs	0,82	**0,75**	0,89	0,63	0,38	0,84	0,83	0,89	0,73	0,62
Multi-task with RGB image + border gradient as inputs	**0,83**	0,73	**0,90**	0,66	0,36	0,87	0,84	**0,91**	0,72	0,63
Multi-task with RGB image + asymmetry ratio and border gradient as inputs	**0,83**	0,74	**0,90**	0,53	0,44	**0,90**	0,91	0,83	0,69	0,62

6.7 Comparison with Benchmark Performances

The aforementioned results are compared to the 'ISIC 2017 Part 3: Lesion Classification' leaderboard[1]. The top 10 performing models and our model with the best performance in terms of AUC and BMA (multi-task model with dermoscopic features classification as auxiliary tasks and no class balancing techniques) are confronted are Table 6.7. For a fair comparison, direct differentiation must be performed with models which did not use additional data sources to train and did not implement ensemble modelling.

Table 6.7: Comparison between the top performing solution proposed in this work and the best challenge submissions (Avg. denotes Average).

Model	**Ensemble Models**	**External Data**	**AUC** MM	SK	AVG.	**BMA**
Matsunaga et al. [6]	✓	✓	0,868	0,953	0.911	0,831
Diaz [156]	-	✓	0,856	0,965	0.911	0,883
Menegola et al. [157]	-	-	0,874	0,943	0.901	0,844
Bi et al. [130]	✓	✓	0,870	0,921	0.896	0,843
Yang et al. [127]	-	-	0,830	0,942	0.886	0,847
DeVries et al. [158]	✓	✓	0,836	0,935	0.886	0,809
Vasconcelos et al. [159]	-	-	0,791	0,911	0.851	0,738
Jia et al. [160]	-	-	0,804	0,855	0.830	0,729
Harangi [161]	✓	-	0,783	0,867	0.825	0,829
Galdran et al. [162]	-	-	0,765	0,881	0.823	0,772
Top performing dissertation approach	-	-	**0,820**	**0,920**	**0.870**	**0,668**

Note: there are 20 submissions in the challenge. Only the top 10 entries ranked according to BMA are shown here, as well as the top performing model presented in the dissertation.

Thus, the leading solution presented in this work reached an average AUC of 87% placing it among the top 35% of the ISIC 2017 challenge submissions. Nonetheless, the fact that we have limited ourselves to a a single model and limited data also has an impact and must be emphasized. Even so note that the main focus of this work was not to achieve the best possible classification performance but rather to investigate the potential of multi-task and multimodality learning and act as a proof-of-concept study.

The most successful ISIC 2017 challenge submissions implemented ensembles of DL networks and extended the provided dataset by using additional data sources to train [163]. Future derivations of this work can be inspired by those approaches.

As explained in Section 4.4.1, EfficientNet-B3 was the only pre-trained CNN architecture employed in this study. The studied techniques can be reproduced in other CNNs to investigate if they lead to superior performance.

[1] https://challenge.isic-archive.com/leaderboards/2017

6.8 Lessons Learned, Limitations and Future Work

Valuable lessons can be extracted after analysis of the proposed models and corresponding results. The best performance solution proposed in this work with respect to the AUC scores and BMA was the multi-task model with no imbalanced learning techniques, which confirms the multi-task network technique is more robust and efficient as compared to the conventional CNN technique. However, regarding three-class prediction, the highest results were obtained in the multi-task model with duplicated samples for each unique label and weighted loss functions (weights computed based on the distribution of the training set) which highly penalize misclassifications of this malignant lesion. Evaluating MM SE is important because the higher it is, the fewer false negative results, and thus fewer cases of cancer are missed. When using both imbalanced learning techniques, more than half of MMs were correctly predicted.

It appears to be an easier task to correctly classify SK than MM and the hand-crafted ABCD rule-inspired features seem to strongly correlate to the first. It can be considered that too many auxiliary tasks might be detrimental to performance. Multimodal fusion of data has the potential to improve the classifier's prediction if more competent characteristics are extracted and/or provided.

Providing segmented images to the CNN model does not add value; the network is capable of detecting the lesion and extracting meaningful features. Hence, when inputting segmentation masks along with the original images, no significant differences were expected; however, MM prediction was heavily degraded and the reasons behind this behavior are unknown.

Regarding the dataset used, while it allows for robust training/validation and comparison to other state of the art methods, it also has significant limitations that must be addressed in future work. Firstly, the methodologies presented were only tested on one dataset, meaning that the results will vary when moving to a different dataset. Moreover, a single train-val-test was used, since it is the ISIC division, allowing for direct comparison to other state of the art methods, and this random division of the data, as well as the ratio of each class can play a significant role. Cross validation tests would be important to ensure that these results are not a consequence of random data allocation for such a dataset split.

To bypass the problem of the small size of the database and broaden the availability of data for research, methods such as GANs, particularly Auxiliary Classifier GANs as they provide stability [164], can be employed to generate artificial and realistic skin lesion images [165, 166].

There is a real concern about the interpretation and explainability of decisions made by DL methods in medical diagnostic systems. Deep learned features are optimal, as they are optimized to achieve the best classification performance. However, DL algorithms are frequently considered as "black box solutions"; the opacity of these algorithms is an obstacle to the trustworthiness of their outcomes [167], specially in medical applications where an incorrect diagnosis can incur in high costs for both the patient and the physician. Hence, it can be important to address this question, moving into the research field of 'Explainable AI' as done by Barata et al. [168].

In terms of clinical applicability, a question that can be raised is the possibility to correctly predict and triage skin cancer based on a single image per case. In a clinical setting, dermatologists

usually combine the visualization of the pigmented skin lesion with external parameters such as medical history and patient personal information, namely age, gender and body part in their assessment, which gives an insight beyond the imaging features used by DL algorithms. The models presented in this dissertation can be further developed to incorporate patient information and/or clinical images as additional inputs, since combining complementary information from multiple modalities has the potential to improve performance. Metadata provided by the database of this work can be used for this purpose, but there is sometimes missing data so such networks would need to be robust to this issue. Moreover, as the ISIC 2017 challenge dataset does not contain macroscopic images, other datasets would also need to be gathered.

The ABCD rule has been further expanded to ABCDE, including a criterion for "Evolving". For this specific trait, a change in shape, size, color and elevation is evaluated. It has become the most important factor to consider in diagnosis since a changing lesion is a warning sign of MM. The development of datasets which track the same individual at different points in time (longitudinal data), thus reflecting the evolution in characteristics of the lesion, could represent an interesting challenge for the community. Additionally, a number of researchers recommend including the "ugly duckling sign", besides the ABCDE rule. It states that MM lesions deviate from the remaining lesions of an individual, and exhibit very different properties. ISIC 2020 challenge [73] presented the first dataset of MM and comparative benign lesions within the same patient. Hence, the inclusion of ugly duckling method in our proposed solution can be beneficial.

This work was developed within the scope of Fh-AICOS Derm.AI project[2], which aims to use AI to power teledermatological screening through the integration of a mobile application to acquire macroscopic skin lesion images with RSE-SIGA[3] and the development of AI-powered risk triage and decision support platform.

Given Derm.AI's framework for risk prioritization and the convenience of a smartphone application for early and autonomous diagnosis, future endeavours can focus on the deployment of the proposed models into real-world scenarios. For this integration, the CNN architecture of this work (EfficientNet-B3) would not be suitable, with a more lightweight model being required: MobileNet [169] could be a solution. Mobile acquired macroscopic images also pose a challenge because they are often subject to various types of distortion [170], hence it would be important to add a block focused on quality classification, i.e., evaluating if an image has sufficient quality for the system.

Altogether, even if decisions made by DL models still have to be corroborated by human experts, automated systems can be a valuable help in the reduction of the workload of physicians as they can assist in decision making processes. Thus, resources of health systems can be more efficiently utilized.

[2] http://dermai.projects.fraunhofer.pt/
[3] https://rse-siga.spms.min-saude.pt/

Chapter 7

Conclusions

Skin cancer is a major global health problem. With its ever-growing incidence and importance of an early diagnosis for a positive prognosis, computer assisted diagnosis systems can play an important role in reducing the burden of physicians. The first approaches reported in the literature followed a process consisting of pre-processing, segmentation also regarded as border detection, feature extraction and classification steps. DL for computer vision is an emerging technology and there have been implementations with DNN architectures for skin lesion classification capable of outperforming human experts performance. MTL has also been explored as a way to improve the predictions of a task by jointly training it with auxiliary related tasks, which helps in the distinction between beneficial and prejudicial features. With multimodal learning, the goal is to help the model to focus on features which are known to be relevant for malignancy classification. To our knowledge there is a small number of comprehensive studies in the literature related to multimodal and multitasking learning applied on skin lesion diagnosis, therefore being the line of research explored throughout the work.

The ultimate objective of this dissertation was to achieve a robust, reliable and competent algorithm for multi-class skin lesion prediction, unlike most methods in the literature which are focused on detecting one class - melanoma. Innovative aspects were introduced: we investigated if extracting auxiliary metadata from the dermoscopic images and fusing it with deep learned features could improve the prediction of the classifier and if MTL could help improve the generalization performance of the main task: skin lesion classification.

A first methodology consisted of comparing the performance of low-level hand-crafted features inspired by the ABCD rule of dermoscopy to deep learned ones extracted by the EfficientNet-B3 architecture pre-trained on ImageNet in a multi-class prediction system. A transfer learning procedure was adopted in the latter to overcome the limitation of data and make use of knowledge learned during training on a general dataset (ImageNet). Best performance was achieved by this pre-trained model (average AUC of 85% against a value of 78% for the manually extracted features), confirming the current trend in this area. This was thus considered the baseline, meaning that more complex solutions and techniques were then applied in this model.

A multi-task model was proposed, with the main focus and principal output still being skin

lesion classification but with the addition of auxiliary related tasks which consisted on prediction of dermoscopic structures correlated with the lesions. The goal of these auxiliary tasks was to provide inductive bias and allow the model to learn representations which can be beneficial for the main task. The application of this learning paradigm produced superior results, achieving an average AUC equal to 87%.

In order to decrease the effect of class imbalance and to prevent infrequent labels from having little contribution to the parameter updates, we proceeded to the implementation of two important class balancing techniques: oversampling and weighted loss functions, to duplicate and over-penalize misclassifications of minority class samples. Additionally, we also ensured that each mini-batch contained at least 1 positive case of each unique label, to update model weights based on all the unique labels in every training step. These methods decreased the overall performance of the model, nonetheless they lead to significantly higher melanoma and SK sensitivities (50% and 80%, respectively), which is desirable for screening purposes.

The role of background information was also evaluated in this work, through the comparison of the same model with different inputs: original dermoscopic view, images cropped around the lesion and images containing only lesion information. The latter led to poor results, explained by the removal of contextual information. The other two resulted in similar results with a slight advantage in melanoma SE (52%) for the cropped images, which suggests that the CNN is able to automatically identify which pixels belong to the lesion.

The MTL technique was further studied with the insertion of more auxiliary tasks closely related to the main task: ABCD rule criteria. By creating features for the asymmetry and border classification tasks, it was expected that those would be shared with skin lesion prediction, hence improving it. However, the results suggested that border strongly correlates with SK as a bias was added towards this class and melanoma classification was jeopardised.

Furthermore, adding the segmentation mask as an extra channel to the RGB dermoscopic image and combining hand-crafted ABCD features with deep learning generated ones through concatenation was investigated. The first approach was not expected to generate a major improvement given that as concluded earlier, the network can perceive what the lesion is and extract meaningful features; nevertheless, MM AUC has a big decrease (from 74% to 53%) which means that the model had no discriminating ability regarding melanoma and was making random guesses. For the late feature fusion approach, asymmetry ratio appeared to improve MM AUC and border gradient benefited NV and SK AUC.

Overall, the main objectives of the dissertation were accomplished. The results reported prove that MTL allows different tasks to share meaningful features, making it more robust and efficient as compared to the conventional CNN technique. Auxiliary classification of ABCD rule criteria did not translate into enhanced performance, which can possibly be explained by the fact that this information was manually extracted, divided in classes and labelled, therefore being imperfect and introducing bias. Even so, the multimodal late fusion strategies with such descriptors increased AUCs. The learning paradigms approached in this work are active areas for improvement and can lead to reliable skin lesion classification systems.

Bibliography

[1] *Cancer*. URL: https://www.who.int/news-room/fact-sheets/detail/cancer (visited on 20/01/2021).

[2] Rebecca L Siegel, Kimberly D Miller and Ahmedin Jemal. "Cancer statistics, 2019". In: *CA: a cancer journal for clinicians* 69.1 (2019), pp. 7–34.

[3] Juliana Berk-Krauss and Mary E Laird. "What's in a Name—Dermoscopy vs Dermatoscopy". In: *JAMA dermatology* 153.12 (2017), pp. 1235–1235.

[4] Michael Binder et al. "Epiluminescence microscopy: a useful tool for the diagnosis of pigmented skin lesions for formally trained dermatologists". In: *Archives of dermatology* 131.3 (1995), pp. 286–291.

[5] Andre Esteva et al. "Dermatologist-level classification of skin cancer with deep neural networks". In: *nature* 542.7639 (2017), pp. 115–118.

[6] Kazuhisa Matsunaga et al. "Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble". In: *arXiv preprint arXiv:1703.03108* (2017).

[7] Nils Gessert et al. "Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data". In: *MethodsX* 7 (2020), p. 100864.

[8] Jordan Yap, William Yolland and Philipp Tschandl. "Multimodal skin lesion classification using deep learning". In: *Experimental dermatology* 27.11 (2018), pp. 1261–1267.

[9] Haofu Liao and Jiebo Luo. "A deep multi-task learning approach to skin lesion classification". In: *arXiv preprint arXiv:1812.03527* (2018).

[10] Jeremy Kawahara et al. "Seven-point checklist and skin lesion classification using multi-task multimodal neural nets". In: *IEEE journal of biomedical and health informatics* 23.2 (2018), pp. 538–546.

[11] *Derm.AI*. URL: https://www.aicos.fraunhofer.pt/en/our_work/projects/dermai.html (visited on 20/06/2021).

[12] Hani Yousef, Mandy Alhajj and Sandeep Sharma. "Anatomy, skin (integument), epidermis". In: (2017).

[13] *What Are Basal and Squamous Cell Skin Cancers? | Types of Skin Cancer*. URL: https://www.cancer.org/cancer/basal-and-squamous-cell-skin-cancer/about/what-is-basal-and-squamous-cell.html (visited on 09/11/2020).

[14] Robert Lanza et al. "Essentials of Stem Cell Biology". In: (2009).

[15] Jennifer Y Lin and David E Fisher. "Melanocyte biology and skin pigmentation". In: *Nature* 445.7130 (2007), pp. 843–850.

[16] *Epidermis | Biology for Majors II*. URL: https://courses.lumenlearning.com/wm-biology2/chapter/epidermis/ (visited on 09/11/2020).

[17] R Shayini et al. "Classification of Skin Lesions in Digital Images for the Diagnosis of Skin Cancer". In: *2020 International Conference on Smart Electronics and Communication (ICOSEC)*. IEEE. 2020, pp. 162–166.

[18] Randy Gordon. "Skin cancer: an overview of epidemiology and risk factors". In: *Seminars in oncology nursing*. Vol. 29. 3. Elsevier. 2013, pp. 160–169.

[19] Melody J Eide et al. "Identification of patients with nonmelanoma skin cancer using health maintenance organization claims data". In: *American journal of epidemiology* 171.1 (2010), pp. 123–128.

[20] Eleni Linos et al. "Increasing burden of melanoma in the United States". In: *Journal of Investigative Dermatology* 129.7 (2009), pp. 1666–1674.

[21] Natalie H Matthews et al. "Epidemiology of melanoma". In: *Exon Publications* (2017), pp. 3–22.

[22] Sertan Kaymak, Parvaneh Esmaili and Ali Serener. "Deep learning for two-step classification of malignant pigmented skin lesions". In: *2018 14th Symposium on Neural Networks and Applications (NEUREL)*. IEEE. 2018, pp. 1–6.

[23] Ana Catarina Fidalgo Barata. *Automatic detection of melanomas using dermoscopy images*. Tech. rep. Technical report, Instituto Superior Tecnico Lisboa, 2017.

[24] Ravindhra G Elluru. "Cutaneous vascular lesions". In: *Facial Plastic Surgery Clinics* 21.1 (2013), pp. 111–126.

[25] *Dermoscopy Atlas*. URL: http://www.dermoscopyatlas.com/ (visited on 17/01/2021).

[26] Rebecca L Siegel, Kimberly D Miller and Ahmedin Jemal. "Cancer statistics, 2017". In: *CA: a cancer journal for clinicians* 67.1 (2017), pp. 7–30.

[27] *Cancer burden statistics and trends across Europe | ECIS*. URL: https://ecis.jrc.ec.europa.eu/ (visited on 28/11/2020).

[28] A. F. Duarte et al. "Skin cancer healthcare impact: A nation-wide assessment of an administrative database". In: *Cancer Epidemiology* 56 (2018), pp. 154–160.

[29] Gery P Guy Jr et al. "Prevalence and costs of skin cancer treatment in the US, 2002- 2006 and 2007- 2011". In: *American journal of preventive medicine* 48.2 (2015), pp. 183–187.

[30] Jenny T Chen, Steven J Kempton and Venkat K Rao. "The economics of skin cancer: an analysis of Medicare payment data". In: *Plastic and Reconstructive Surgery Global Open* 4.9 (2016).

[31] William Gillen and Brett Coldiron. *Skin Cancer Chapter 4 Burden of Disease*. Tech. rep.

[32] Esther Erdei and Salina M Torres. "A new understanding in the epidemiology of melanoma". In: *Expert review of anticancer therapy* 10.11 (2010), pp. 1811–1823.

[33] Zachary J Wolner et al. "Enhancing skin cancer diagnosis with dermoscopy". In: *Dermatologic clinics* 35.4 (2017), pp. 417–437.

[34] *Dermoscopy: Overview, Technical Procedures and Equipment, Color*. URL: https://emedicine.medscape.com/article/1130783-overview (visited on 03/11/2020).

[35] Robert H Johr. "Dermoscopy: alternative melanocytic algorithms—the ABCD rule of dermatoscopy, menzies scoring method, and 7-point checklist". In: *Clinics in dermatology* 20.3 (2002), pp. 240–247.

[36] Hubert Pehamberger, Andreas Steiner and Klaus Wolff. "In vivo epiluminescence microscopy of pigmented skin lesions. I. Pattern analysis of pigmented skin lesions". In: *Journal of the American Academy of Dermatology* 17.4 (1987), pp. 571–583.

[37] *Melanoma Molecular Map Project*. URL: http://www.mmmp.org/MMMP/import.mmmp?page=dermoscopy.mmmp (visited on 03/11/2020).

[38] Giuseppe Argenziano et al. "Dermoscopy of pigmented skin lesions: results of a consensus meeting via the Internet". In: *Journal of the American Academy of Dermatology* 48.5 (2003), pp. 679–693.

[39] *Menzies Method - dermoscopedia*. URL: https://dermoscopedia.org/Menzies%7B%5C_%7DMethod (visited on 03/11/2020).

[40] *Using Dermoscopy to Identify Melanoma and Improve Diagnostic Discrimination (FULL) | AVAHO*. URL: https://www.mdedge.com/fedprac/avaho/article/165262/melanoma/using-dermoscopy-identify-melanoma-and-improve-diagnostic (visited on 03/11/2020).

[41] *Dermoscopy_tutorial*. URL: http://www.dermoscopy.org/consensus/2b.asp (visited on 09/11/2020).

[42] Wilhelm Stolz. "ABCD rule of dermatoscopy: a new practical method for early recognition of malignant melanoma". In: *Eur. J. Dermatol.* 4 (1994), pp. 521–527.

[43] Giuseppe Argenziano et al. "Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions: comparison of the ABCD rule of dermatoscopy and a new 7-point checklist based on pattern analysis". In: *Archives of dermatology* 134.12 (1998), pp. 1563–1570.

[44] Paulina Pala, Beata S Bergler-Czop and Jakub M Gwiżdż. "Teledermatology: idea, benefits and risks of modern age–a systematic review based on melanoma". In: *Advances in Dermatology and Allergology/Postpy Dermatologii i Alergologii* 37.2 (2020), p. 159.

[45] Adelina Costin and Constança Furtado. "Experience of a Pilot Project of Teledermatology Screening at the Department of Dermatology and Venereology of Hospital Garcia de Orta". In: *Journal of the Portuguese Society of Dermatology and Venereology* 77.4 (2019), pp. 311–314.

[46] Damilola A Okuboyejo and Oludayo O Olugbara. "A review of prevalent methods for automatic skin lesion diagnosis". In: *The Open Dermatology Journal* 12.1 (2018).

[47] DJ Eedy and R Wootton. "Teledermatology: a review". In: *British Journal of Dermatology* 144.4 (2001), pp. 696–707.

[48] E Tan et al. "Successful triage of patients referred to a skin lesion clinic using teledermoscopy (IMAGE IT trial)". In: *British Journal of Dermatology* 162.4 (2010), pp. 803–811.

[49] Sarah J Coates, Joseph Kvedar and Richard D Granstein. "Teledermatology: from historical perspective to emerging techniques of the modern era: part II: emerging technologies in teledermatology, limitations and future directions". In: *Journal of the American Academy of Dermatology* 72.4 (2015), pp. 577–586.

[50] Akhilesh S Pathipati, Luke Lee and April W Armstrong. "Health-care delivery methods in teledermatology: consultative, triage and direct-care models". In: *Journal of telemedicine and telecare* 17.4 (2011), pp. 214–216.

[51] Alexander BöRVE et al. "Smartphone teledermoscopy referrals: a novel process for improved triage of skin cancer patients". In: *Acta dermato-venereologica* 95.2 (2015), pp. 186–190.

[52] *What is Artificial Intelligence (AI)? | IBM*. URL: https://www.ibm.com/cloud/learn/what-is-artificial-intelligence (visited on 04/12/2020).

[53] *The History of Artificial Intelligence - Science in the News*. URL: http://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/ (visited on 04/12/2020).

[54] *AI and robotics are transforming healthcare: Why AI and robotics will define New Health: Publications: Healthcare: Industries: PwC*. URL: https://www.pwc.com/gx/en/industries/healthcare/publications/ai-robotics-new-health/transforming-healthcare.html (visited on 04/12/2020).

[55] Wayne Xiong et al. "Achieving human parity in conversational speech recognition". In: *arXiv preprint arXiv:1610.05256* (2016).

[56] David Silver et al. "Mastering the game of Go with deep neural networks and tree search". In: *nature* 529.7587 (2016), pp. 484–489.

[57] Fei Jiang et al. "Artificial intelligence in healthcare: past, present and future". In: *Stroke and vascular neurology* 2.4 (2017).

[58] Thomas Davenport and Ravi Kalakota. "The potential for artificial intelligence in healthcare". In: *Future healthcare journal* 6.2 (2019), p. 94.

[59] Natale Cascinelli et al. "A possible new tool for clinical diagnosis of melanoma: the computer". In: *Journal of the American Academy of Dermatology* 16.2 (1987), pp. 361–367.

[60] Greg R Day and Robert H Barbour. "Automated melanoma diagnosis: where are we at?" In: *Skin Research and Technology* 6.1 (2000), pp. 1–5.

[61] Konstantin Korotkov and Rafael Garcia. "Computerized analysis of pigmented skin lesions: a review". In: *Artificial intelligence in medicine* 56.2 (2012), pp. 69–90.

[62] Josep Malvehy et al. "Dermoscopy report: proposal for standardization: results of a consensus meeting of the International Dermoscopy Society". In: *Journal of the American Academy of Dermatology* 57.1 (2007), pp. 84–95.

[63] M Hand, A Chien and D Grossman. "Screening and Non-Invasive Evaluative Devices for Melanoma Detection: A Comparison of Commercially Available Devices and Dermoscopic Evaluation". In: *J Clin Dermatol Ther* 1.005 (2015).

[64] Titus Josef Brinker et al. "Skin cancer classification using convolutional neural networks: systematic review". In: *Journal of medical Internet research* 20.10 (2018), e11936.

[65] *ADDI - Automatic computer-based Diagnosis system for Dermoscopy Images*. URL: https://www.fc.up.pt/addi/index.html (visited on 20/11/2020).

[66] TF Mendonca et al. "Ph2: A public database for the analysis of dermoscopic images". In: *Dermoscopy image analysis* (2015).

[67] *ISIC Project – ISDIS*. URL: https://isdis.isic-archive.com/isic-project/ (visited on 20/11/2020).

[68] David Gutman et al. "Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC)". In: *arXiv preprint arXiv:1605.01397* (2016).

[69] Noel CF Codella et al. "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)". In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE. 2018, pp. 168–172.

[70] Philipp Tschandl, Cliff Rosendahl and Harald Kittler. "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions". In: *Scientific data* 5.1 (2018), pp. 1–9.

[71] Noel Codella et al. "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic)". In: *arXiv preprint arXiv:1902.03368* (2019).

[72] Marc Combalia et al. "BCN20000: Dermoscopic lesions in the wild". In: *arXiv preprint arXiv:1908.02288* (2019).

[73] Veronica Rotemberg et al. "A patient-centric dataset of images and metadata for identifying melanomas using clinical context". In: *Scientific Data* 8.1 (2021), pp. 1–8.

[74] *Dermofit Image Library - Edinburgh Innovations*. URL: https://licensing.edinburgh-innovations.ed.ac.uk/i/software/dermofit-image-library.html (visited on 23/11/2020).

[75] *About Us | Dermatology Education*. URL: http://www.dermnet.com/about-us/ (visited on 23/11/2020).

[76] *7-point criteria evaluation Database*. URL: http://derm.cs.sfu.ca/Welcome.html (visited on 22/11/2020).

[77] *What is Machine Learning? | IBM*. URL: https://www.ibm.com/cloud/learn/machine-learning (visited on 04/12/2020).

[78] Tim Lee et al. "Dullrazor®: A software approach to hair removal from images". In: *Computers in biology and medicine* 27.6 (1997), pp. 533–543.

[79] Wen-Yu Chang et al. "The feasibility of using manual segmentation in a multifeature computer-aided diagnosis system for classification of skin lesions: a retrospective comparative study". In: *BMJ open* 5.4 (2015), e007823.

[80] Guillod Joel et al. "Validation of segmentation techniques for digital dermoscopy". In: *Skin Research and Technology* 8.4 (2002), pp. 240–249.

[81] Ezzeddine Zagrouba and Walid Barhoumi. "A prelimary approach for the automated recognition of malignant melanoma". In: *Image Analysis & Stereology* 23.2 (2004), pp. 121–135.

[82] David Delgado Gomez et al. "Independent histogram pursuit for segmentation of skin lesions". In: *IEEE transactions on biomedical engineering* 55.1 (2007), pp. 157–161.

[83] Ezzeddine Zagrouba and Walid Barhoumi. "An accelerated system for melanoma diagnosis based on subset feature selection". In: *Journal of Computing and Information Technology* 13.1 (2005), pp. 69–82.

[84] Célia A Zorzo Barcelos and VB Pires. "An automatic based nonlinear diffusion equations scheme for skin lesion segmentation". In: *Applied Mathematics and Computation* 215.1 (2009), pp. 251–261.

[85] M Emre Celebi, Y Alp Aslandogan and Paul R Bergstresser. "Unsupervised border detection of skin lesion images". In: *International Conference on Information Technology: Coding and Computing (ITCC'05)-Volume II*. Vol. 2. IEEE. 2005, pp. 123–128.

[86] Kerri-Ann Norton et al. "Three-phase general border detection method for dermoscopy images using non-uniform illumination correction". In: *Skin Research and Technology* 18.3 (2012), pp. 290–300.

[87] Matthew G Fleming et al. "Techniques for a structural analysis of dermatoscopic imagery". In: *Computerized medical imaging and graphics* 22.5 (1998), pp. 375–389.

[88] Giuseppe Di Leo et al. "An improved procedure for the automatic detection of dermoscopic structures in digital ELM images of skin lesions". In: *2008 IEEE Conference on Virtual Environments, Human-Computer Interfaces and Measurement Systems*. IEEE. 2008, pp. 190–194.

[89] Catarina Barata, Jorge S Marques and Jorge Rozeira. "Detecting the pigment network in dermoscopy images: a directional approach". In: *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE. 2011, pp. 5120–5123.

[90] Sumi Yoshino et al. "Application of morphology for detection of dots in tumor". In: *SICE 2004 Annual Conference*. Vol. 1. IEEE. 2004, pp. 591–594.

[91] William V Stoecker et al. "Detection of asymmetric blotches (asymmetric structureless areas) in dermoscopy images of malignant melanoma using relative color". In: *Skin Research and Technology* 11.3 (2005), pp. 179–184.

[92] William V Stoecker, William Weiling Li and Randy H Moss. "Automatic detection of asymmetry in skin tumors". In: *Computerized Medical Imaging and Graphics* 16.3 (1992), pp. 191–197.

[93] Sungjoon Cho. "Dermal Radiomics: a new approach for computer-aided melanoma screening system". In: (2016).

[94] R Joe Stanley, William V Stoecker and Randy H Moss. "A relative color approach to color discrimination for malignant melanoma detection in dermoscopy images". In: *Skin Research and Technology* 13.1 (2007), pp. 62–72.

[95] Catarina Barata et al. "Two systems for the detection of melanomas in dermoscopy images using texture and color features". In: *IEEE Systems Journal* 8.3 (2013), pp. 965–979.

[96] Mariam A Sheha, Mai S Mabrouk, Amr Sharawy et al. "Automatic detection of melanoma skin cancer using texture analysis". In: *International Journal of Computer Applications* 42.20 (2012), pp. 22–26.

[97] Lucio Andreassi et al. "Digital dermoscopy analysis for the differentiation of atypical nevi and early melanoma: a new quantitative semiology". In: *Archives of dermatology* 135.12 (1999), pp. 1459–1465.

[98] William V Stoecker et al. "Detection of granularity in dermoscopy images of malignant melanoma using color and texture features". In: *Computerized Medical Imaging and Graphics* 35.2 (2011), pp. 144–147.

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

[100] Mihran Tuceryan and Anil K Jain. "Texture analysis". In: *Handbook of pattern recognition and computer vision*. World Scientific, 1993, pp. 235–276.

[101] Ronn P Walvick et al. "Classification of melanoma using wavelet transform-based optimal feature set". In: *Medical Imaging 2004: Image Processing*. Vol. 5370. International Society for Optics and Photonics. 2004, pp. 944–951.

[102] Rahil Garnavi, Mohammad Aldeen and James Bailey. "Computer-aided diagnosis of melanoma using border-and wavelet-based texture analysis". In: *IEEE Transactions on Information Technology in Biomedicine* 16.6 (2012), pp. 1239–1252.

[103] Pietro Rubegni et al. "Automated diagnosis of pigmented skin lesions". In: *International Journal of Cancer* 101.6 (2002), pp. 576–580.

[104] Marco Burroni et al. "Melanoma computer-aided diagnosis: reliability and feasibility study". In: *Clinical cancer research* 10.6 (2004), pp. 1881–1886.

[105] Stephan Dreiseitl et al. "A comparison of machine learning methods for the diagnosis of pigmented skin lesions". In: *Journal of biomedical informatics* 34.1 (2001), pp. 28–36.

[106] Gerald Schaefer et al. "An ensemble classification approach for melanoma diagnosis". In: *Memetic Computing* 6.4 (2014), pp. 233–240.

[107] M Burroni et al. "Dysplastic naevus vs. in situ melanoma: digital dermoscopy analysis". In: *British Journal of Dermatology* 152.4 (2005), pp. 679–684.

[108] Kajsa Møllersen et al. "Unsupervised segmentation for digital dermoscopic images". In: *Skin Research and Technology* 16.4 (2010), pp. 401–407.

[109] Andrea Pennisi et al. "Skin lesion image segmentation using Delaunay Triangulation for melanoma detection". In: *Computerized Medical Imaging and Graphics* 52 (2016), pp. 89–103.

[110] Niall O'Mahony et al. "Deep learning vs. traditional computer vision". In: *Science and Information Conference*. Springer. 2019, pp. 128–144.

[111] Athanasios Voulodimos et al. "Deep learning for computer vision: A brief review". In: *Computational intelligence and neuroscience* 2018 (2018).

[112] Alex Krizhevsky, Ilya Sutskever and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Ed. by F Pereira et al. Vol. 25. Curran Associates, Inc., 2012, pp. 1097–1105.

[113] Michael A Nielsen. *Neural networks and deep learning*. Vol. 25.

[114] Alex Krizhevsky, Ilya Sutskever and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Communications of the ACM* 60.6 (2017), pp. 84–90.

[115] Olga Russakovsky et al. "Imagenet large scale visual recognition challenge". In: *International journal of computer vision* 115.3 (2015), pp. 211–252.

[116] *ILSVRC2015 Results*. URL: http://image-net.org/challenges/LSVRC/2015/results (visited on 21/01/2021).

[117] *ILSVRC2017*. URL: http://image-net.org/challenges/LSVRC/2017/results (visited on 21/01/2021).

[118] Christian Szegedy et al. "Going deeper with convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[119] Karl Weiss, Taghi M Khoshgoftaar and DingDing Wang. "A survey of transfer learning". In: *Journal of Big data* 3.1 (2016), p. 9.

[120] Yasuhiro Fujisawa, Sae Inoue and Yoshiyuki Nakamura. "The possibility of deep learning-based, computer-aided skin tumor classifiers". In: *Frontiers in medicine* 6 (2019), p. 191.

[121] Rich Caruana. "Multitask learning". In: *Machine learning* 28.1 (1997), pp. 41–75.

[122] Sebastian Ruder. "An overview of multi-task learning in deep neural networks". In: *arXiv preprint arXiv:1706.05098* (2017).

[123] Ross Girshick. "Fast r-cnn". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1440–1448.

[124] Dhanesh Ramachandram and Graham W Taylor. "Deep multimodal learning: A survey on recent advances and trends". In: *IEEE Signal Processing Magazine* 34.6 (2017), pp. 96–108.

[125] Noel Codella et al. "Deep learning, sparse coding, and SVM for melanoma recognition in dermoscopy images". In: *International workshop on machine learning in medical imaging*. Springer. 2015, pp. 118–126.

[126] Jeremy Kawahara, Aicha BenTaieb and Ghassan Hamarneh. "Deep features to classify skin lesions". In: *2016 IEEE 13th international symposium on biomedical imaging (ISBI)*. IEEE. 2016, pp. 1397–1400.

[127] Xulei Yang et al. "A novel multi-task deep learning model for skin lesion segmentation and classification". In: *arXiv preprint arXiv:1703.01025* (2017).

[128] Fábio Perez et al. "Data augmentation for skin lesion analysis". In: *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*. Springer, 2018, pp. 303–311.

[129] Yangqing Jia et al. "Caffe: Convolutional architecture for fast feature embedding". In: *Proceedings of the 22nd ACM international conference on Multimedia*. 2014, pp. 675–678.

[130] Lei Bi et al. "Automatic skin lesion analysis using large-scale dermoscopy images and deep residual networks". In: *arXiv preprint arXiv:1703.04197* (2017).

[131] Balazs Harangi. "Skin lesion classification with ensembles of deep convolutional neural networks". In: *Journal of biomedical informatics* 86 (2018), pp. 25–32.

[132] Zongyuan Ge et al. "Skin disease recognition using deep saliency features and multimodal learning of dermoscopy and clinical images". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pp. 250–258.

[133] Ammara Masood, Adel Al-Jumaily and Khairul Anam. "Self-supervised learning model for skin cancer diagnosis". In: *2015 7th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE. 2015, pp. 1012–1015.

[134] Holger A Haenssle et al. "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* 29.8 (2018), pp. 1836–1842.

[135] Seung Seog Han et al. "Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm". In: *Journal of Investigative Dermatology* 138.7 (2018), pp. 1529–1538.

[136] Zhishun She, Y Liu and A Damatoa. "Combination of features from skin pattern and ABCD analysis for lesion classification". In: *Skin Research and Technology* 13.1 (2007), pp. 25–33.

[137] Ihab Zaqout. "Diagnosis of skin lesions based on dermoscopic images using image processing techniques". In: *Pattern Recognition-Selected Methods and Applications* (2019).

[138] M Emre Celebi et al. "Lesion border detection in dermoscopy images". In: *Computerized medical imaging and graphics* 33.2 (2009), pp. 148–153.

[139] Hitoshi Iyatomi et al. "Parameterization of dermoscopic findings for the internet-based melanoma screening system". In: *2007 IEEE Symposium on Computational Intelligence in Image and Signal Processing*. IEEE. 2007, pp. 189–193.

[140] Catarina Barata, M Emre Celebi and Jorge S Marques. "A survey of feature extraction in dermoscopy image analysis of skin cancer". In: *IEEE journal of biomedical and health informatics* 23.3 (2018), pp. 1096–1109.

[141] Luıs Rosado, Maria João Vasconcelos and Márcia Ferreira. "A mobile-based prototype for skin lesion analysis: Towards a patient-oriented design approach". In: *International Journal of Online Engineering (iJOE)* 9.S8 (2013), pp. 27–29.

[142] Maria João M Vasconcelos, Luıs Rosado and Márcia Ferreira. "A new risk assessment methodology for dermoscopic skin lesion images". In: *2015 IEEE International Symposium on Medical Measurements and Applications (MeMeA) Proceedings*. IEEE. 2015, pp. 570–575.

[143] Ping Zhou and Jim Austin. "Learning criteria for training neural network classifiers". In: *Neural computing & applications* 7.4 (1998), pp. 334–342.

[144] Mingxing Tan and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 6105–6114.

[145] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[146] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[147] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[148] Adria Romero Lopez et al. "Skin lesion classification from dermoscopic images using deep learning techniques". In: *2017 13th IASTED international conference on biomedical engineering (BioMed)*. IEEE. 2017, pp. 49–54.

[149] Luis Perez and Jason Wang. "The effectiveness of data augmentation in image classification using deep learning". In: *arXiv preprint arXiv:1712.04621* (2017).

[150] Agnieszka Mikołajczyk and Michał Grochowski. "Data augmentation for improving deep learning in image classification problem". In: *2018 international interdisciplinary PhD workshop (IIPhDW)*. IEEE. 2018, pp. 117–122.

[151] Connor Shorten and Taghi M Khoshgoftaar. "A survey on image data augmentation for deep learning". In: *Journal of Big Data* 6.1 (2019), pp. 1–48.

[152] SM Stricklin et al. "Cloudy and starry milia-like cysts: how well do they distinguish seborrheic keratoses from malignant melanomas?" In: *Journal of the European Academy of Dermatology and Venereology* 25.10 (2011), pp. 1222–1224.

[153] Maria A Pizzichetta et al. "Negative pigment network: an additional dermoscopic feature for the diagnosis of melanoma". In: *Journal of the American Academy of Dermatology* 68.4 (2013), pp. 552–559.

[154] Mateusz Buda, Atsuto Maki and Maciej A Mazurowski. "A systematic study of the class imbalance problem in convolutional neural networks". In: *Neural Networks* 106 (2018), pp. 249–259.

[155] A Ardeshir Goshtasby and Stavri Nikolov. "Image fusion: advances in the state of the art". In: *Information fusion* 2.8 (2007), pp. 114–118.

[156] Iván González Dıaz. "Incorporating the knowledge of dermatologists to convolutional neural networks for the diagnosis of skin lesions". In: *arXiv preprint arXiv:1703.01976* (2017).

[157] Afonso Menegola et al. "RECOD titans at ISIC challenge 2017". In: *arXiv preprint arXiv:1703.04819* (2017).

[158] Terrance DeVries and Dhanesh Ramachandram. "Skin lesion classification using deep multi-scale convolutional neural networks". In: *arXiv preprint arXiv:1703.01402* (2017).

[159] Cristina Nader Vasconcelos and Bárbara Nader Vasconcelos. "Convolutional neural network committees for melanoma classification with classical and expert knowledge based image transforms data augmentation". In: *arXiv preprint arXiv:1702.07025* (2017).

[160] Xi Jia and Linlin Shen. "Skin lesion classification using class activation map". In: *arXiv preprint arXiv:1703.01053* (2017).

[161] Balazs Harangi. “Skin lesion detection based on an ensemble of deep convolutional neural network”. In: *arXiv preprint arXiv:1705.03360* (2017).

[162] Adrian Galdran et al. “Data-driven color augmentation techniques for deep skin image analysis”. In: *arXiv preprint arXiv:1703.03702* (2017).

[163] Noel CF Codella et al. “Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)”. In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE. 2018, pp. 168–172.

[164] Augustus Odena, Christopher Olah and Jonathon Shlens. “Conditional image synthesis with auxiliary classifier gans”. In: *International conference on machine learning*. PMLR. 2017, pp. 2642–2651.

[165] Alceu Bissoto et al. “Skin lesion synthesis with generative adversarial networks”. In: *OR 2.0 context-aware operating theaters, computer assisted robotic endoscopy, clinical image-based procedures, and skin image analysis*. Springer, 2018, pp. 294–302.

[166] Alec Xiang and Fei Wang. “Towards interpretable skin lesion classification with deep learning models”. In: *AMIA annual symposium proceedings*. Vol. 2019. American Medical Informatics Association. 2019, p. 1246.

[167] Juan Manuel Durán and Karin Rolanda Jongsma. “Who is afraid of black box algorithms? on the epistemological and ethical basis of trust in medical AI”. In: *Journal of Medical Ethics* 47.5 (2021), pp. 329–335.

[168] Catarina Barata, M Emre Celebi and Jorge S Marques. “Explainable skin lesion diagnosis using taxonomies”. In: *Pattern Recognition* 110 (2021), p. 107413.

[169] Andrew G Howard et al. “Mobilenets: Efficient convolutional neural networks for mobile vision applications”. In: *arXiv preprint arXiv:1704.04861* (2017).

[170] Thanh-Toan Do et al. “Accessible melanoma detection using smartphones and mobile image analysis”. In: *IEEE Transactions on Multimedia* 20.10 (2018), pp. 2849–2864.

Appendix A

Supplementary Tables of Chapter 6 - Results

Table A.1: Additional results of the models with hand-crafted and deep learned features (bold values highlight the best result for each metric).

Model	**PPV**			**NPV**		
	Nv	MM	SK	Nv	MM	SK
ABCD rule neural network	0,74	0,38	0,43	0,61	0,83	0,90
Baseline transfer-learning ternary	**0,82**	**0,59**	**0,48**	**0,68**	**0,87**	**0,94**

Table A.2: Additional results of the multi-task models, along with the results from the baseline multi-class model (bold values highlight the best result for each metric).

Model	**PPV**			**NPV**		
	Nv	MM	SK	Nv	MM	SK
Baseline transfer-learning ternary	0,82	0,59	0,48	0,68	**0,87**	0,94
Multi-task with binary dermoscopic features classification as auxiliary tasks	**0,84**	**0,61**	0,53	0,68	**0,87**	**0,95**
Multi-task with multi-class dermoscopic features classification as auxiliary tasks	0,82	0,53	**0,59**	**0,69**	0,86	**0,95**

Table A.3: Additional results of the multi-task models with data balancing techniques (bold values highlight the best result for each metric).

Model	**PPV**			**NPV**		
	Nv	MM	SK	Nv	MM	SK
Baseline with raw data	0,84	**0,61**	**0,53**	**0,68**	**0,87**	0,95
Skin lesions oversampled + no class weights	0,85	0,46	0,50	0,63	0,86	0,95
Data oversampled + no class weights	0,85	0,47	0,49	0,64	0,86	0,95
Data oversampled + class weights from training set	**0,88**	0,44	0,46	0,58	**0,87**	**0,96**
Data oversampled + class weights from batch	**0,88**	0,50	0,45	0,63	**0,87**	**0,96**

Table A.4: Additional results of the optimized through data balancing multi-task models with modified images as input (bold values highlight the best result for each metric).

Model	**PPV**			**NPV**		
	Nv	MM	SK	Nv	MM	SK
Original Images	**0,88**	**0,50**	0,45	**0,63**	**0,87**	**0,96**
Segmented Images	0,79	0,37	0,51	0,57	0,86	0,91
Cropped Images	0,87	0,37	**0,53**	0,57	**0,87**	**0,96**

Table A.5: Additional results of the optimized through data balancing multi-task models with ABCD rule related auxiliary tasks (bold values highlight the best result for each metric).

Model	**PPV**			**NPV**		
	Nv	MM	SK	Nv	MM	SK
Multi-task with binary dermoscopic features classification as auxiliary tasks	**0,88**	**0,44**	**0,46**	**0,58**	**0,87**	0,96
Multi-task with binary dermoscopic features & multi-class asymmetry prediction as auxiliary tasks	0,87	0,39	0,32	0,51	0,84	**0,97**
Multi-task with binary dermoscopic features & multi-class border prediction as auxiliary tasks	0,87	0,39	0,32	0,51	0,84	**0,97**
Multi-task with binary dermoscopic features, multi-class asymmetry & border prediction as auxiliary tasks	**0,88**	0,42	0,34	0,54	0,84	**0,97**

Table A.6: Additional results of the optimized through data balancing multi-task models with multiple inputs (bold values highlight the best result for each metric).

Model	**PPV**			**NPV**		
	Nv	MM	SK	Nv	MM	SK
Multi-task with single input (RGB image)	0,88	0,44	**0,46**	**0,58**	**0,87**	0,96
Multi-task with 4 channel input (RGB image + segmentation mask)	0,88	0,23	0,24	0,41	0,82	0,95
Multi-task with RGB image + asymmetry ratio as inputs	0,88	0,45	0,35	0,55	0,85	0,96
Multi-task with RGB image + border gradient as inputs	**0,89**	**0,49**	0,35	0,57	0,85	0,97
Multi-task with RGB image + asymmetry ratio and border gradient as inputs	0,92	0,38	0,34	0,51	0,86	**0,98**