Luís Manuel Gonçalves Góis

Automatic Classification of Cutaneous Lesions Using Dermoscopic Images

Master's Thesis, in Biomedical Engineering, field of Clinical Informatics and Bioinformatics, oriented by Professor Francisco José Santiago Fernandes Amado Caramelo, co-oriented by Professor Miguel Patrício Dias and supervised by Doctor José Carlos Cardoso

September, 2017

· U C ·

UNIVERSITY OF COIMBRA

• U  C •

FCTUC FACULDADE DE CIÊNCIAS
E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Luís Manuel Gonçalves Góis

Automatic Classification of Cutaneous Lesions Using Dermoscopic Images

Thesis submitted to the
University of Coimbra for the degree of
Master in Biomedical Engineering

Jury:
Américo Manuel Costa Figueiredo (President of the jury)
Luís Alberto da Silva Cruz (Member of the jury)
Francisco José Santiago Fernandes Amado Caramelo (Supervisor)

September, 2017

This work was developped in collaboration with:

Institute for Biomedical Imaging and Life Sciences

Coimbra's University Hospital Centre

Acknowledgments

I would like to express my gratitude and appreciation to my family, for understanding that I couldn't always be available for their needs. Sorry for not paying you the attention you deserve, but I'm sure you're proud of me.

Special thanks to my girlfriend, Rita, who has been accompanying me, not only during this journey, and gives me the strength to go on and push further.

Additionally, I would like to state my acknowledgements to Professor Francisco Caramelo, and Professor Miguel Patricio, my supervisors, who guided me, and provided invaluable help, throughout this thesis development.

The author would also like to thank the Institute of Dermatology from Coimbra's University Hospital Centre, and specially, to Doctor José Carlos Cardoso, for allowing access to the dataset used in this study and furthermore for providing us with an in-depth knowledge about clinical diagnosis of skin lesions.

Lastly, I would like to thank all my other friends, and to address everyone that somewhat contributed to the completion of this work, you have my gratitude.

Abstract

Statistical evidence has revealed that malignant melanoma is the deadliest form of skin cancer, with an increasing incidence rate over the past decades, worldwide. Nevertheless, it's the most treatable one, depending on the stage of the cancer, as further researches have shown that the early detection and intervention of melanoma implicates higher chances of cure. That being said, clinical diagnosis of melanoma is a challenging task for dermatologists, since the processes are prone to misdiagnosis and inaccuracies due to the characteristic similarities of melanoma with other skin lesions.

In the past decades, several computer-aided diagnosis systems have been proposed to increase the specificity and sensitivity of melanoma detection. However, to the best of our knowledge, these systems are still imperfect, which explains why clinical applications have not been created yet. Thus, the hypothesis of this study was to gather the most successful methods in the literature (by performing a systematic review) and combine them with some novel ones, in order to create an effective computer-aided assessment tool that could assist doctors in the categorization of skin lesions as benign or malignant.

In this work, a methodological approach to the automatic classification of skin lesions in dermoscopy images is presented. Noise reduction is the first step we apply, in order to improve the image's illumination parameters, as well as eliminating surrounding hair and additional unwanted artefacts. Secondly, border detection is performed, to differentiate the lesion from the surrounding background skin. Then, a sequence of transformations is applied to each lesion to extract a global set of colour, texture, border and shape attributes, which afterwards are fed into an optimization selection framework, which ranks these attributes according to their importance. To find this optimal feature vector, RelieF and Principal Component Analysis techniques are compared. Lastly, classification is done through the use of three classifiers, namely, Support Vector Machines, Random Forests and Adaptive Boosting.

The proposed method has been evaluated on a set of 100 dermoscopic images, including benign and melanoma cases. Regarding the optimized selection of features, the RelieF method surpasses Principal Component Analysis as the most effective framework for melanoma diagnosis, and in the end, we use 5 of 36 different discriminating parameters to train and test our models. Among the three used classifiers Adaptive Boosting achieves the best average results for 500 runs, obtaining a sensitivity of 99.9%, and a specificity of 97.9%, for the melanoma class, and an overall accuracy of 98.2% for discriminating between malignant and benign classes.

The experimental results show promising signs for a future integration of this system on the clinical level, as a complementary system that could be used to screen images and complement doctors decision on whether or not a biopsy is necessary.

Key words: Melanoma - computer-aided system - dermoscopy - skin cancer - classification

Resumo

Segundo dados estatísticos o melanoma maligno é a forma mais mortífera de cancro da pele, com uma taxa de incidência crescente ao longo das últimas décadas, em todo o mundo. Contudo, é a mais tratável, dependendo do estado do cancro, já que pesquisas mais aprofundadas mostraram que a deteção prematura e o tratamento do melanoma significam maior probabilidade de cura. No entanto, o diagnóstico clínico do melanoma constitui um desafio grande para os dermatologistas, visto que os processos são propensos a diagnósticos errados e imprecisões devido às semelhanças entre as características do melanoma e de outras lesões cutâneas.

Nas últimas décadas vários sistemas de diagnóstico assistidos por computador foram propostos para aumentar a especificidade e a sensibilidade da deteção do melanoma. Todavia, quanto foi possível saber, esses sistemas são ainda imperfeitos, o que explica o porquê de ainda não terem sido criadas aplicações clínicas. Assim sendo, a hipótese deste estudo passou pelo recolhimento dos métodos mais bem-sucedidos da literatura (através de uma revisão sistemática) e da sua combinação com alguns novos, com o objetivo de criar uma ferramenta de avaliação eficiente que possa auxiliar os médicos na categorização de lesões cutâneas, como benignas ou malignas.

Neste trabalho, apresenta-se uma abordagem metodológica para a classificação automática de lesões cutâneas, obtidas através de imagens dermatoscópicas. A redução de ruído é o primeiro passo que aplicamos, a fim de melhorar os parâmetros de iluminação da imagem, bem como eliminar os pelos envolventes e outros artefactos indesejados. Em segundo lugar, é realizada a deteção da fronteira da lesão, para diferenciar a lesão da pele circundante. De seguida, uma sequência de transformações é aplicada a cada lesão, para extrair um conjunto de features globais, de cor, textura, fronteira e forma, que posteriormente são introduzidas numa estrutura de otimização, que as ordena de acordo com o seu grau de importância. Para encontrar o vetor de features ideal, as técnicas de RelieF e Principal Component Analysis são comparadas. Por fim, a classificação é feita através do uso de três classificadores,

a saber: Support Vector Machines, Random Forests e Adaptive Boosting.

O método proposto foi avaliado num conjunto de 100 imagens dermatoscópicas, que incluiam casos benignos e malignos. Em relação à seleção otimizada das features, o método RelieF ultrapassa o Principal Component Analysis como a estrutura mais eficaz para o diagnóstico de melanoma, sendo que no final, acabámos por usar 5 de 36 parâmetros de categorização, para treinar e testar os nossos modelos. Entre os três classificadores utilizados, o método de Adaptive Boosting, globalmente, apresenta as melhores médias de resultados para 500 execuções, obtendo uma sensibilidade de 99,9% e uma especificidade de 97,9%, para a classe dos melanomas, e uma precisão geral de 98,2% para discriminação entre as duas classes.

Os resultados experimentais mostram sinais promissores para uma futura integração deste sistema ao nível clínico, como um sistema complementar que poderia ser usado para visualizar imagens e complementar a decisão dos médicos sobre se uma biópsia é ou não necessária.

Palavras chave: Melanoma – sistema assistido por computador – dermatoscopia – cancro da pele – classificação

List of Acronyms

AdaB adaptive boosting.
ANN artificial neural network.

CAD computer-aided design.
CFS correlation based feature selection.
CLAHE contrast limited adaptive histogram equalization.
CSLM confocal scanning laser microscopy.

DecT decision tree.

GLCM co-occurrence matrix measures.
GUI graphical user interface.
GVF gradient vector flow.

HSL hue-saturation-lightness.
HSV hue-saturation-value.
HTF homomorphic transform filtering.

k-NN k-nearest neighbour.
KLT karhunen-loève transform.

LDA linear discriminant analysis.
LMT logistic model tree.

MIFS mutual information based feature selection.
MRI magnetic resonance imaging.

NBayes naive bayes.

OCT optical coherence tomography.

PCA principal component analysis.
PDE partial differential equations.

RBF radial basis function.
RF random forests.
RGB red-green-blue.
RLM run-length matrix.
ROI region of interest.

SVM support vector machine.

TDS total dermoscopy score.

UV ultraviolet.

List of Figures

List of Tables

Contents

1

Introduction

1.1 Contextualization

1.1.1 Human Skin

The human skin is the largest organ in the human body and consists of two principal layers with distinct function and distinct optical properties: the epidermis and the dermis (Figure 1.1). Below the dermis there's another structure called hypodermies, which as its importance as a subcutaneous tissue.

The epidermis is the outermost layer of the skin, being made up of a stratified-squamous-epithelium[1]. The epidermis contains no blood vessels, and cells in the deepest layers are nourished almost exclusively by diffused oxygen from the surrounding air and to a far lesser degree by blood capillaries extending to the outer layers of the dermis. The epidermis can be further subdivided into the following strata (beginning with the outermost layer): corneum, lucidum (only in palms of hands and bottoms of feet), granulosum, spinosum, basale. Cells are formed through mitosis at the innermost layer and then move up the strata changing shape and composition as they differentiate and become filled with keratin. They eventually reach the top layer (stratum corneum) and are sloughed off, or desquamated[2]. This process forms the keratinized layer of skin, responsible for a waterproof protective wrap over the body's surface which also serves as protection against external aggressions, such as injuries, infections and ultraviolet radiation. The main type of cells which make up the epidermis are the following:

▶ Keratinocytes - These represent the majority (95%) of cells in the epidermis and are the driving force for continuous renewal of the skin [1]. They form the protective

[1] Squamous/flattened epithelial cells arranged in layers.
[2] Lose their cohesion and separate from the surface.

layer consisting of keratin-impregnated cells, serving as the first line of defence as they are a barrier between an organism and its environment. The daughter keratinocytes produced by division in the basal layer are referred to as basal cells and the flat keratinocytes cells in the outer part of the epidermis (that are constantly shed as new ones form) are referred to as Squamous cell.

▶ Melanocytes - These dendritic melanin-producing cells found in the basal layer of the epidermis distribute packages of melanin pigment to the surrounding keratinocytes, allowing the hair and skin to have their characteristic colour. Melanin acts as a filter that protects the deeper layers of the skin from harmful effects of ultraviolet (UV) radiation, strongly absorbing light in the blue part of the visible and the UV spectrum.

▶ Langerhans cells - Dendritically shaped cells similar to the melanocytes, located in the squamous epithelia of the epidermis, their function is to detect antigens that have penetrated the epidermis and deliver them to the local lymph nodes where antibodys will be produced to fight them.

▶ Merkel cells - They exist in the basal layer of the epidermis. They act as mechanosensory receptors in response to touch, relaying touch-related information such as texture and pressure to the brain.

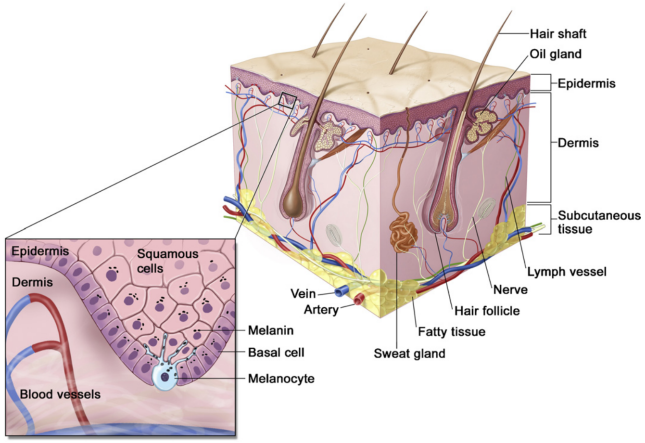

Figure 1.1: Anatomy of the skin, showing the epidermis, the dermis, and subcutaneous (hypodermic) tissue [1].

The other principal skin layer, the dermis, plays an important role in thermoregulation, healing and sense of touch, providing also energy and nutrition to the epidermis, being located beneath the epidermis and is tightly connected to it by the basement membrane. The dermis is made of collagen and elastic fibbers, which confer elasticity to the skin, and is structurally divided into two areas: a superficial area adjacent

to the epidermis, called papillary region (thin layer), and a deep thicker area known as reticular region. While the former serves as a "glue" that holds the epidermis and the dermis together, the latter contains hair follicles, sweat glands, sebaceous glands, apocrine glands, lymphatic vessels and blood vessels. The nerve endings provide the sense of touch and heat, while the blood vessels provide nourishment and waste removal from its own cells as well as from the stratum basale of the epidermis [1].

Last but not least, we have the hypodermis, a structure that is not part of the skin but lies below the dermis acting as a subcutaneous tissue. The hypodermis contains 50% of the body fat, consisting of loose connective tissue, adipose tissue and elastin. Its role is supplying skin with blood vessels and nerves and attaching it to underlying bone and muscle.

1.1.2 Cutaneous Lesion

Skin lesions, depending on their behaviour, may be classified as benign[3] or malignant[4] (responsible for skin cancer), with certain cells being more cancer-prone than others. Benign lesions, such as seborrhoeic keratosis or nevi for example, show a more ordered and controlled growth, and do not proliferate into other tissues. On the contrary, malignant skin cells are generally unlimited in self-growth, and may invade other tissues, with the severity of the cancer raising as the distance from metastasised tissues to the primary initial focus increases. The development of these malignant skin cells is basically based on mutations of oncogenes and tumour suppressor genes predominantly induced by UV radiation.

Skin cancer is generally characterized by an abnormal run-away growth of groups of cells on the skin and is the most common form of cancer, being responsible, globally, for about 40% of all cancer cases [2]. Skin cancer types are named after the skin cell in which the cancer develops, with the great majority of skin carcinomas[5] arising from basal cells, squamous cells and melanocytes (already referenced in the previous subsection). Since most of the skin cancers develop from non-pigmented cells and not from pigmented melanocytes, the two most common types of skin cancers come from basal and squamous keratinocytes, which later develop into basal cell carcinoma and squamous cell carcinoma accordingly [1]. These forms of skin cancer, are generally highly curable, however skin cancers that develop from pigmented melanocytes are

[3]Mass of cells (tumour) that lacks the ability to invade neighboring tissue or metastasise.
[4]Tumour that has the ability to multiply uncontrollably and metastasise.
[5]Carcinoma is another word for cancer.

the most aggressive ones and they are called melanoma. This malignant tumour arises from melanocytic cells and primarily involves the skin, showing the highest metastatic rate among all skin tumours and accounting for more than 90% of skin cancer-related deaths [3],[4]. In addition, it's also important to make a brief reference to melanocytic nevi, lesions originated from melanocytes, and that are one of the most common types of lesion that we will be working with in our dataset. They are benign neoplasms[6] or hamartomas[7] that might transit to a melanoma, and that's where the problem starts, dermatologists knowing if they are in the presence of a benign lesion, somewhat suspect, or highly suspect of being a melanoma lesion. Judgement of whether or not a lesion is suggestive for removal is based on several clinical criteria that refer to the morphology and the changes over time of the individual lesion and to the overall clinical context of the patient. In an attempt to remove all possible melanomas, many benign skin tumours are biopsied because of the presence, to a varying extent, of clinical features associated with melanoma. Removing benign lesions is considered an acceptable price to pay so as not to miss melanoma, thus the current practice is to remove any pigmented or non pigmented lesion that is suggestive of melanoma and to perform histopathologic examination.

1.1.2.1 Melanoma

Melanoma is the deadliest form of skin cancer and its incidence is increasing in the last decades, especially in white populations, with the highest incidence rates worldwide being reported in Australia and New Zealand, rendering malignant melanoma as one of the most common tumours in these Caucasian populations [3]. In individuals with more pigmentation such as Asians and Africans, melanomas are less common and are almost always found on either the acral or mucosal surfaces [5]. These facts regarding the major incidence increase of melanoma among people with less pigmentation can be easily proven when we look at Europe, a continent where the highest incidence rates are in Sweden, Norway, and Denmark, while the lowest are in the Mediterranean countries. This north-south gradient can be justified by the darker pigmentation of Mediterranean populations, a factor that allows them the decreased susceptibility regarding the dangers of sunlight.

On the basis of these stats regarding Europe we present recent data obtained by the Robert Koch-Institute in Berlin referring Germany (a central European country). They estimated values for raw and age-standardized incidence rates of melanoma

[6] An abnormal growth of tissue, which might become a tumour if it forms a mass.
[7] Focal malformation that resembles a neoplasm.

from 1999–2008 (Figure 1.2), and as we can see there has been a rise in incidence rates in men, from 13.7 to 22.1 cases per 100 000 people and year and in women, from 16.5 to 21.2 cases per 100 000 people and year. Therefore, we can conclude that over the past decades, there has been a significant increase in the melanoma incidence rates for men and women [3].

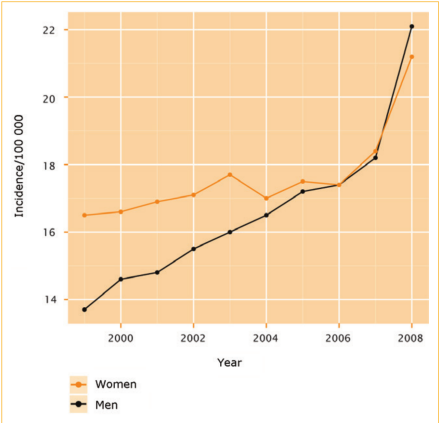

Figure 1.2: Melanoma incidence rates in Germany [3].

This high rise regarding incidence rates in recent years is thought to be related to the thinning and/or depletion of the ozone layer. However, UV exposure is showing an increasing impact on the number of registered skin cancer cases, and that can be explained by changes in leisure and travel habits. There are several other factors regarding genetics (the inheritance of melanoma is polygenic), or constitutional factors as evidenced previously, with people with lighter skin and poor immune function generally more at risk of developing skin cancer, but the most important exogenous factor is exposure to ultraviolet irradiation, as typically experienced during summer vacations [4].

There's one important statistic that is also important to realize, which relates to the relative stabilization of mortality rates, as in fact the increase in mortality rates has been markedly lower than the rise in incidence rates. A possible reason for this reality comes from the awareness caused by the notably rising incidence rates, a problem that eventually leads to enhanced early detection of prognostically more favourable tumours [3].

1.1.3 Dermoscopy

Nowadays exist several non-invasive skin imaging techniques such as confocal scanning laser microscopy (CSLM), optical coherence tomography (OCT), clinical imag-

ing (generally acquired via a still camera), ultrasound, magnetic resonance imaging (MRI) and spectroscopic imaging, but the one used to obtain our database was dermoscopy. Dermoscopy is a portable non-invasive skin imaging technique that uses special optic equipment to grant a magnified visualization of the skin surface and subdermal structures, allowing computer-assisted examination of skin lesions by opening a new dimension of the clinical morphological features of skin lesions. Dermoscopy is a relatively simple technique that can be carried out in a doctor's office, clinic, or hospital, which makes use of a dermatoscope, a device that unlike other techniques such as the previously referred ones does better in avoiding images from being affected by the presence of artefacts, such as hairs, shadows and lines, by poor resolution, and by variable observing conditions, such as distance and illumination.

The introduction of dermoscopy into the clinical practice of dermatology is associated with both a significant increase of the sensitivity for melanoma diagnosis and a significant reduction of numbers of benign skin lesions unnecessarily excised. According to [6], a systematic review has reported that the average sensitivities for melanoma of the naked eye and dermoscopy examinations were 74% and 90%, respectively. This kind of data are similar to other studies which estimated that dermoscopy allowed 10–27% higher sensibility than clinical diagnosis by the unaided eye [6]. These stats prove that this image tool has the capacity of reducing the number of presumptive diagnoses that have to be confirmed histologically after skin biopsy, especially because the results occur without losing specificity, which means that better melanoma detection does not increase the number of unnecessary excision of benign tumours.

In essence, dermoscopy has helped to overcome several problems, being able to provide much more accurate results than using clinical evaluation alone by the naked eye, and also being very valuable for analysing particularly pigmented skin lesions and other characteristics of a lesion, including symmetry, size, borders, presence and distribution of colour features.

1.1.4 Diagnosis rules used by dermatologists

In the world of dermatology its quite a common practise to evaluate the lesion being examined with the help of certain rules. These rules represent diagnostic methods from where the practitioners choose a model that they prefer. The most common of these analysis techniques can be semi-quantitative models (ABCD rule, 7-point rule) or qualitative models (Menzies' method, pattern analysis). They help doctors

diagnose tumours earlier and easier, and have shown high reliability if they are based on quantitative automated scoring systems [7]. In sum, all these algorithms are valid ways of evaluating skin lesions with dermoscopy, thus afterwards we summarize the concept behind each of them:

► ABCD rule - This mathematical approach was created after a meticulous analysis of multiple dermoscopic criteria in order to discover the best crucial parameters to diagnose melanoma. This way a four test criteria was created and gave birth to the ABCD rule, where A stands for asymmetry, B for borders, C for colours and D for diameter (or different structural components). Each criteria is given a score, with asymetry being given the highest weight, and in the end a formula called total dermoscopy score (TDS) is elaborated based on a multiplication of each of these scores by conversion factors. Every one of these criteria is appraised in a different way: asymmetry is measured by dividing the lesion into two perpendicular axes and assessing the form of it and also the disposition of its structures; in a similar way the borders are also evaluated by dividing the lesion into perpendicular axes, plus two more oblique axes, this way eight borders remain and are observed regarding the pigmentation of its net; the colours for instance are evaluated regarding the appearance of six specific possible colours (black, dark brown, light brown, gray-bluish, white, and red), with each of them being given a score if they are present in the lesion; at last we have the D rule, where five different possible structures (pigmented net, clustered globules, ramified streaks, amorphous area, and dots) are scored concerning their presence. With that said, after all parameters being calculated a score is given, with lesions which score below 4.75 being considered benign and if they score above 4.75 they are considered suspect, and an excision should be assessed, specially above 5.45 which means a lesion is highly suspicious of being a melanoma.

► 7-point checklist - Its main goal is to distinguish 7 dermoscopic structures from a lesion and achieve a final score also based on a mathematical approach. These structures can be divided into two categories, since the weight of their characteristics aren't all the same. One of these categories is known as the major criteria, which attributes two points when a dermatologist finds an atypical pigmentary net, a blue-whitish veil, or a atypical vascular pattern. On the other hand, we have the minor criteria, which gives one point if the dermatologist finds radial streaks, pseudopods, irregular pigmentation, globules and irregular spots, or regression patterns. In the end, after the total score is achieved, the lesions are considered benign if they score below three while if they score equal to or higher than three, the lesion has a 95% chance of correctly being considered melanoma [8].

► Menzies' method - An analysis technique with a different approach to the previous one, since instead of having graded criteria it score characteristics as categorically present or absent, in an attempt to reduce intra- and inter-observer errors. This type of classification considers two types of characteristics, positive and negative. The negative ones are two, the symmetry of pattern within a lesion (not necessarily symmetry of contour) and the presence of a single colour. They are responsible for proving that a lesion is not a melanoma, which means that if one of these is present the lesion is defined as benign. On the other hand, we have the positive characteristics, which are basically nine features, more particularly: blue-whitish veil, multiple brown dots, pseudopods, radiated streaks, areas of scar depigmentation, peripheral black dots/globules, multiple colors (five or six), multiple blue/gray dots, and enlarged pigmentary net. If one or more of these nine positive features can be identified, in addition to the absence of negative characteristics, then the lesion is considered a melanoma.

► Pattern analysis - This approach is one of the most commonly used for providing diagnostic accuracy for cutaneous melanoma, it seeks to identify specific patterns, which may be global and local. Initially, the lesions are analysed with regard to their global features (arrangements of textured patterns covering most of the lesion), which allow a brief preliminary categorization of the skin lesion, and afterwards by their local features, for a more detailed assessment of individual or grouped characteristics . The goal of the dermatologist is to identified if skin lesion is melanocytic or non-melanocytic, and in that sense we present the patterns associated with those identifications in Table 1.1 [7].

	benign melanocytic lesions	**non-melanocytic lesions**
global patterns	reticular; globular; cobblestone; homogeneous; parallel; star-burst;	multicomponent; unspecific;
local patterns	typical pigmented network; regular dots and globules; regular streaks; regular blotches; symmetric area without structure;	atypical pigmented net; irregular dots and globules; irregular streaks; bluish veil; regression areas(white or with blue dots); asymmetric area without structure;

Table 1.1: Pattern analysis of skin lesions

All things considered, these are the most prevalent forms, used in the current practise, for distinguishing skin lesions.

1.2 Motivation/Contribution: The importance of Computer-Aided Algorithms

As previously mentioned in Subsection 1.1.2, skin cancer is one of the most prevalent cancer types, with thousands of patients losing their life every year as incidence increases faster than that of almost all other cancers [9]. Additionally, skin cancer is known for having one of the most deadliest types of cancer, more precisely, malignant melanoma, a less common but far more deadly type of skin cancer. One of the worst scenarios for a clinician dealing with skin lesions is failure to diagnose melanoma, since in its advanced stages (with signs of metastasis) melanoma is almost incurable, and the treatment, being solely palliative, includes surgery, immunotherapy, chemotherapy, and/or radiation therapy. That being said, the early diagnosis and intervention is the main way to achieve higher chances of cure and avoiding later stage treatments, which are normally ineffective.

Clinical diagnosis and prognosis of melanoma are challenging, since the processes are prone to misdiagnosis and inaccuracies even for experienced dermatologists [9]. The number of melanomas that are not clinically suggestive and are consequently left untreated have been reported to range from 1.5% to 15%, and that is probably caused because although the majority of melanomas exhibit a sufficient array of clinical features to justify biopsy, melanoma may occasionally mimic a variety of benign lesions [4].

It's precisely due to these unfortunate statistics that there is an urgent need to develop innovative strategies able to increase the diagnostic accuracy and to help dermatologists. As so, the aim of this work was to create a CAD system capable of handling large amounts of data, and providing, preferably in real-time and in an automated fashion, a likelihood for the diagnosis of skin lesions, more particularly evaluate if we are dealing with a melanocytic lesion somewhat benign or suspect of being a melanoma or whether were being presented with a real melanoma. The intent is to increase the performance of the diagnoses when compared to the ones used by dermatologists, firstly by increasing the sensitivity results, for the sake of don't incorrectly classifying a malignant lesion as benign (which is of vital importance), and secondly, by increasing the specificity and avoiding a benign lesion to be classified as malignant, because despite the excision of benign lesions being tolerable, it should be minimized as possible to reduce morbidity. Indeed, high sensitivity is more important than a high specificity in this case, but one of the biggest problems of the current CAD systems is their high rate of false positive assessments.

Furthermore, there is another problem related to the current systems that we would also like to improve, even if it's only a futuristic idea. It's a methodological problem, that comes as a shortcoming in the design of these CAD systems, since they are intended to diagnose a lesion without any interaction with the dermatologist, and sometimes without sufficient information for diagnosis. Due to the fact that they're not designed to work in support of the doctors, only a few systems are found in routine clinical use, which makes its practical value still unclear, although most patients would accept using computerized analysis for lesion screening.

Moreover, considering the difficulty of standardizing the diagnostic criteria and the wide variability of the encountered structures, the computerized image analysis techniques have become important instruments in this research area. Matter of fact, they have become a powerful tool in the diagnosis of skin lesions, especially in distinguishing between malignant melanomas and benign melanocytic skin lesions. The ideal CAD system should define the type of a lesion and provide dermatologists with comprehensive information regarding the grounds of this decision. The purpose of it will be to provide an important diagnostic cue for the clinician, one that is not subjective, and that will be able to assist the physician in making decisions. That way when our algorithm corroborates a suspicion of skin cancer, the clinician will be more careful when analysing a lesion and will not underestimate it. We hope this will help on timely diagnosis and treatment of melanoma, decreasing unfavourable prognosis and therefore raising patient survival.

In sum, our algorithm aims to meet high performance expectations, but most of all, even if we don't reach the most satisfactory results we want it to be a relevant starting point to further development. Moreover, this work combines the results of research done so far related to all the steps needed for the development of an automatic diagnostic system, and we hope it can continue being a subject of improvement, and an important contribution to the research area of skin lesion classification for several reasons.

Finally, it's important to point out that the refinement of current approaches reported in the most recent literature and the development of new techniques and methods will help to improve the ability to diagnose skin lesions more precisely and to improve the classification accuracy of these lesions, specially if the goal of significantly reducing melanoma mortality rate is achieved.

2

Automatic procedures for the categorization of skin lesions

2.1 Introduction

The main goal of this work was developing a novel method for automatic classification of skin lesions. In order to understand the state of the art in the field we decided to perform a systematic review based on a specific research question. Systematic reviews are known to be an efficient way to provide convenient evidential articles, which then serve as a powerful tool for covering the area, and in some cases, for making informed decisions.

With this kind of approach, our efforts, were to find as much as possible of the relevant articles addressing the review's research question, instead of being over-influenced by studies which are simply the easiest or most accessible to find. The research question we addressed was "Which are the best methods for automatic classification of skin lesions?". In order to tackle the research question, we followed the PICO strategy, which divides the research question into four parts: population (P), intervention (I), comparison (C) and outcome (O). Although the PICO method is not straightforward in our situation, it makes possible to split the question into different parts which can then be referred with keywords. As so, we considered the population as the methods which use automatic procedures to assess images, obtained with a dermatoscope (intervention), aiming at differentiating (outcome) malignant from benign skin lesions (comparison). The articles were obtained by resorting to the PubMed search engine, while using the following keywords combined with boolean symbols (and/or) connectors:

"*(detect* OR automatic OR analyst* OR classification) AND (skin lesions OR melanoma OR skin cancer OR tumours OR border detection OR pigment*) AND*

(dermoscop) AND (machine learning OR processing OR segmentation OR features OR statistical OR clustering)*"

Further limitations of the research were adopted, by imposing rules to the language of the articles – only english writen papers were admited – and the year of publication which was naturally imposed in the PubMed form – only papers after 2002 (15 years) were admited. The systematic review was performed in several steps, firstly by assessing only the title, secondly the abstract and only then the whole article. All the articles that didn't satisfy the imposed constraints or that were not in agreement with the aim of the research question were discarded. In the end, after all articles being analysed, we elaborated a summary "table" as the synthesis of our results, which we present in Figure 2.1.

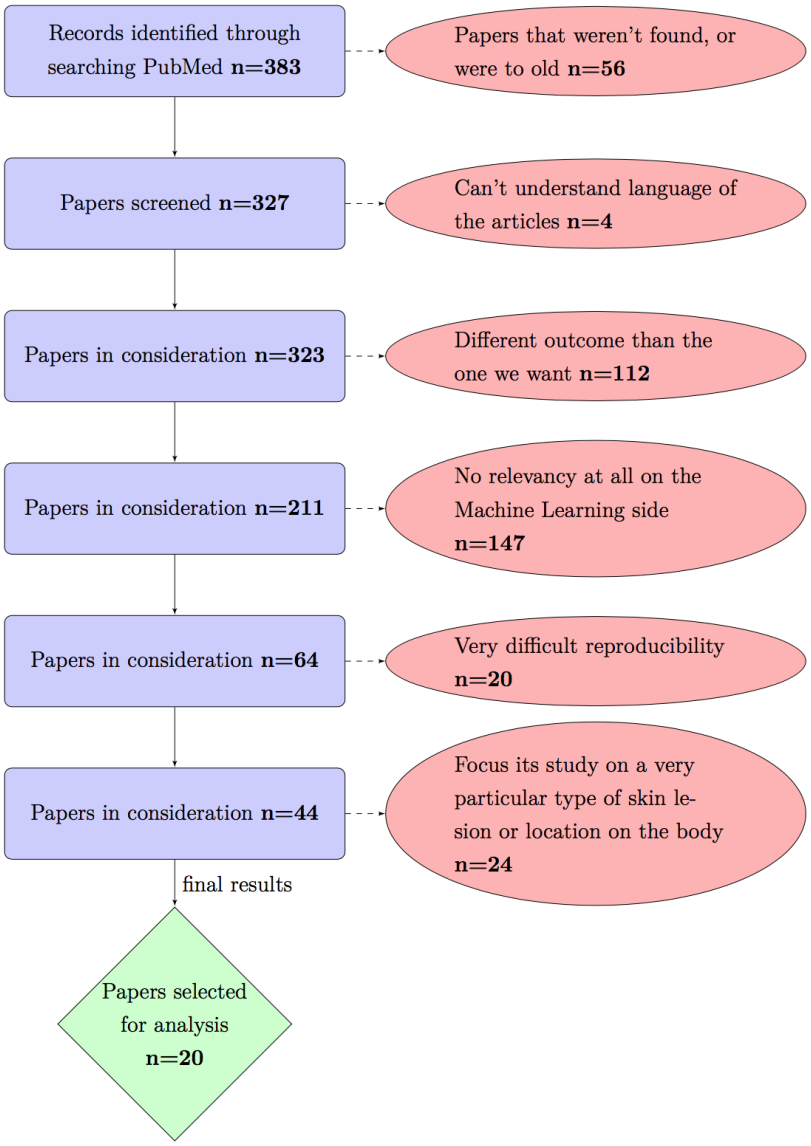

Figure 2.1: Results from the systematic review.

In this figure, statistical information is provided showing how many articles were included and excluded, with the rules related to the dismissal of the articles also being presented. Our attempt was to identify, appraise and synthesize all the empirical evidence that met the pre-specified eligibility criteria that we had defined at the beginning of the review.

2.2 Work-flow of the literature's algorithms

The main focus from the majority of the reviewed articles, follows the steps represented in Figure 2.2, which concern the image's processing and analysis, the segmentation approach, the feature extraction, and the classification methodology.

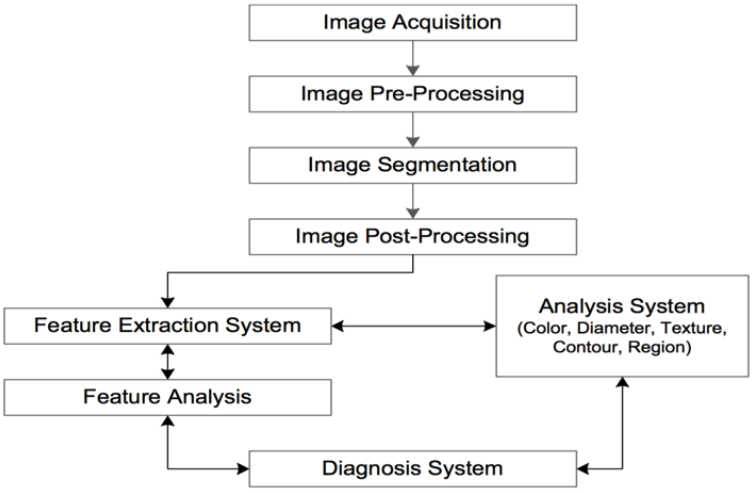

Figure 2.2: Design of a typical skin lesion algorithm

Table 2.1 shows which steps are explored in each article concerning pre-processing, segmentation, features extraction and classification.

As it can be observed, it is clear that the majority of the literature referenced in Table 2.1 follows the steps previously defined. Therefore, we can conclude that computer-aided design (CAD) systems, involved in the context of skin lesion characterization and diagnosis, are no exception to this rule subdivision of work-flow steps. For that reason, the state of the art analysis will be focused on the following four image analysis steps:

► Pre-processing - in order to enhance the visual appearance of the images and improve the manipulation of the dataset.

► Segmentation - to decompose images into features of interest, notably distinguishing them from the background.

► Features Extraction/Selection - to select and extract the most relevant features according to their significance.

► Classification - to train a classifier that in the end will be able to map the input data to a category.

Pre-Processing	Segmentation	Features Extraction	Classification	*References*
✓	✓	✓	✓	[10]
✓	✓	✓	✓	[11]
✓	✓	✓	✓	[12]
✓	✓	✓	✓	[13]
✓	✓	✓	✓	[14]
✓	✓	✓	✓	[15]
✓	✓	✓		[16]
	✓	✓	✓	[17]
		✓	✓	[18]
✓	✓	✓	✓	[19]
	✓	✓	✓	[20]
	✓	✓	✓	[21]
✓	✓	✓	✓	[22]
	✓	✓	✓	[23]
✓	✓	✓	✓	[24]
✓	✓	✓	✓	[25]
✓	✓	✓	✓	[26]
✓	✓	✓		[27]
✓	✓	✓	✓	[28]
✓	✓	✓	✓	[29]

Table 2.1: Analysis of the methods applied by other authors.

3

State of the Art

3.1 Pre-processing

Pre-processing is known as the first stage of detection to enhance the quality of an image, removing irrelevant noise and unwanted parts in the background of it. Its purpose is to perform one of the major challenges in medical data, the ability to distinguish, detect, identify and characterize anomalies in the data that could signal disease, allowing later to proceed into a medical diagnosis or prognosis and treatment or excision, if necessary.

When we focus primarily on the issue of our work, we realize that when we explore a database of dermoscopic images, some of them may not have the optimal quality for subsequent analysis. Hence, the pre-processing step serves to compensate for the imperfections of image acquisition as it tries to improve the quality of images by removing unrelated and extra parts in the background of the image for further processing. Good performance and selection of the pre-processing techniques, not only ensures correct behaviour of the algorithms in the following stages of analysis, but can also have a great influence on the results if it can contribute to improve the accuracy of the system. In Table 3.1, we present references to the articles which have implemented the most common pre-processing techniques for skin lesion images. Furthermore, we briefly summarize these techniques together with the explanation of their methods.

Pre-processing has come a long way regarding skin lesions, and there already exist a series of pre-processing steps that are being applied currently, namely colour space transformation, contrast enhancement, and artefact removal. Hair removal is among the most common and necessary artefact rejection operations. Dermoscopy images often contain hairs present on the skin, that may occlude parts of the lesion, making correct segmentation and texture analysis quite difficult or even impossible. In order

to avoid this problem, several methods have been developed in the last decade. A typical hair-removal algorithm comprises two steps: hair detection and hair repair. Hair repair, is also known as in-painting and it basically consists in filling the image space occupied by hair with proper intensity values.

Processes	*References*
Artefact removal:	
Hair	[28],[24],[26],[25],[15],[19],[27],[16],[13],[10]
Various artefacts	[13],[14],[24], [28],[29],[16],[12],[15]
† Black frames	[16],[25]
Image Enhancement:	
Contrast improvement	[24],[25],[26],[14]
Edge improvement	[28]
Correction of uneven ilumination	[27],[25],[26],[14]

Table 3.1: Pre-processing operations

The most widely adopted method and the one seen as the pioneer of hair removal is called DullRazor, a method proposed in 1997 that was used by [24] and [28]. Recently, some of the existing methods have been reviewed and new algorithms have emerged, for example both [26] and [25] proposed the use of derivative of Gaussian for hair detection and the use of morphological edge-based techniques for the refinement of detected lines. In addition, both these two articles employed fast marching image in-painting techniques for hair repair. This non-iterative in-painting method proved to be more effective than others in repairing the hair-occluded information, for instance like the exemplar-based one. While some of the approaches use generalized methods of supervised learning to detect hairs, others use more specific algorithms like [15] and [19], who decided to detect black hair by filtering their images using Gabor directional filters, a process also used for the extraction of features. Moreover, and short while ago, a number of methods have been developed mostly based on morphological operations and adaptive thresholding. For example [27] used grey scale morphological closing operation to detect black hair, while [16] decided to detect black hair using the white top-hat transform followed by in-painting, based on the replacement of hair line pixels with values calculated on the basis of the neighbourhood pixels. These latter author in [13] decided to perform a similar algorithm, but this time using black top-hat transform for the detection of hair, a morphological image processing technique that returns an image, containing elements that are darker than their surroundings and smaller than the structur-

ing element. [10], presented a new alternative in-painting method using non-linear partial differential equations (PDE).

Despite hair removal being a crucial pre-processing step, there are several other artefacts that images contain such as black frames, small pores, shines, reflections, thin hairs, air bubbles, as well as intrinsic cutaneous features that can affect border detection, such as blood vessels, hairs, and skin lines. Thus, these external objects greatly affect the quality of the lesion's border and texture in a negative way, which results in loss of accuracy as well as an increase in computational time. The most straightforward way to remove these artefacts is to smooth the image using a general purpose filter such as the median filter. This filter reduces the intensity gradients inside the lesion and in the surrounding healthy skin, and generally allows to adequately remove artefacts while preserving, and sometimes enhancing region boundaries, even the ragged edges. This approach to suppress the noise was applied by [13], [14] , [24], [28] and [29]. An alternative filter that can be used to reduce the influence of skin lines, air bubbles and light, thin hairs is the appliance of a Gaussian filter like [16] did. Furthermore, other methods not so divulged have been created such as morphological closing applied by [12] to remove outliers from the image, or even a simple threshold algorithm followed by in-painting operation as [15] developed. For the removal of a more specific artefact such as the black frames, [16] proposed an iterative algorithm based on the lightness component of the hue-saturation-lightness (HSL) colour space, and [25] provided a simple scanning method based on a linear search method across the boundaries of the image.

Another very important step of the pre-processing work-flow is the colour transformation phase, where several approaches have been employed. Both [19], [13] and [24] decided to convert the red-green-blue (RGB) colour space to grey-scale image while [22], [26] and [25] for instance converted RGB skin images to the L*a*b* colour space, a perceptually adaptive space that seems to provide more accurate results than RGB or hue-saturation-value (HSV). Some articles like [15] and [27] decided to go for a more simple approach, sticking to the RGB plane and using the blue component. One article in particularly, in this case [11], decided to transform the RGB colour space into different colour spaces such as grey scale, YCbCr, and HSV, and examined the efficiency of each one for texture analysis, realizing that malignant lesions were more distinctive in the blue colour channel.

In order to improve the behaviour of the algorithms in the following stages of analysis, image enhancement operations were made to soften the constraints of image acquisition. In [25] and [26] a homomorphic transform filtering (HTF) technique was

employed to correct uneven illumination and contrast enhancement, while in [28] karhunen-loève transform (KLT) was used in order to perform edge enhancement and facilitate segmentation. Other alternatives in this category were also tested, for example [24] used a Gaussian filter to enhance the contrast of the image, while [14] applied contrast limited adaptive histogram equalization (CLAHE), a method that processes each tile of the image and allows the output to be more precise than enhancing the contrast of an entire image. In [14], local–global block analysis is used to normalize the filtered image in order to remove variable illumination, and [27] applied local adaptive threshold to different segments of the image willing to compensate for non-uniform lighting changes throughout the image.

3.2 Segmentation

The role of segmentation is to simplify or change the representation of an image into something that is more meaningful and easier to analyse, being crucial in most tasks requiring image analysis. In computer vision, image segmentation is the process that consists in the partitioning of an image into disjoint regions that are homogeneous while keeping track of all the important properties of the image. As a matter of fact the resulting segments cover the entire image, or a set of contours extracted from the image. More precisely, it can also be regarded as the process of grouping together pixels that have similar attributes with respect to some characteristic or computed property, such as colour, intensity, or texture for instance.

In our case image segmentation is used to locate boundaries in skin lesion images, a process also called edge detection. The goal is to recognize patterns or regularities in the image as a means to extract the region of interest (ROI), in this case, the skin lesion. This operation generally requires the detection of discontinuities in the image, in order to identify the region of interest. Nevertheless, the level to which the subdivision is carried has to be very precise since there is no point in performing the segmentation past the level of detail required to identify those elements.

A lot of effort has been made to improve skin lesion segmentation algorithms and come up with adequate measures of evaluating their performance, however, in general, its very difficult to achieve a reliable and accurate border detection method by purely automatic means. Firstly, since dermatologists do not usually delineate lesion borders for diagnosis, there exists a ground truth problem, but in addition there's also high inter and intra-observer variability in boundary perception which makes very hard for multiple persons to agree on the discrimination between subtle

variations in contrast or blur, when perceiving the boundaries of a lesion. Secondly, the morphological structure of a lesion itself can act as a confusion factor for both manual and automatic segmentation. Moreover, various conditions, such as type of lesion, location, colour conditions or angle of view, add to the diverse difficulties in segmenting using the same imaging modality. These problems have led to the development of a wide variety of segmentation algorithms, thus we provide information regarding our selected articles and try to emphasize the role of their methods. In Table 3.2, we present references to some of the most common available techniques that aim to provide robustness in segmentation, adapting to specific conditions of the image type.

Methods	*References*
Common:	
Threshold-based	[16],[23],[27],[14],[19],[15],[24],[21]
Edge-based	[25],[26]
Region-based	[16],[13],[29]
Artificial intelligence	[22],[10]
Active contours	[28]
Geometric deformable models	[11]
Rare:	
Manual segmentation	[20]
Delaunay Triangulation	[12]
Probability distribution	[17]

Table 3.2: Segmentation operations

One of the classic approaches to segment skin lesions is to use threshold-based methods. The threshold levels can be either manually or automatically selected and are applied to generate a binary map, where the skin lesion is extracted from the background. Several articles like [15] or [24] decided to opt for simple thresholding algorithms, while [21] for instance decided to use adaptive thresholding with the purpose of not using the same threshold throughout the whole image. Some algorithms are more sophisticated in comparison with relatively simple methods such as the previously referred. Among these new methods there's [23] who applied a hybrid border detection method which used global thresholding to detect an initial boundary of the lesion, followed by adaptive histogram thresholding on optimized colour channels to refine the border. It's also important to refer Otsu's thresholding method, an automatic algorithm which employs the normalization histogram to find the optimum threshold level to be applied for lesion segmentation. Several articles like [27] used

this method to segment the affected skin lesion from the normal skin. Some of them decided to test it in fusion with other methods, like [14], who did it by applying Otsu's Segmentation and then connecting the non-zero pixels to the neighbourhood non-zero pixels to draw the border for later extraction while [19] computed a global threshold using Otsu's method and followed it with an active contour algorithm that segments the image into lesion and background regions, using Sparse-Field level-set method. Most of these articles which employed threshold-based algorithms ended up performing morphological operations, not only for erosion but also for dilation, in order to remove unwanted noise particles, especially the ones touching the border of the image, and to smooth the edges and expand the border.

Segmentation algorithms may also be edge-based, a process that searches for discontinuities in the intensity of pixels, when compared to pixels in the neighbouring regions. However, such algorithms are known to achieve, in many instances, only partial segmentation, and must be applied in combination with other segmentation methods to obtain a adequate result. [25] and [26] are examples of edge-based segmentation algorithms. They both performed rough tumour segmentation by using minimum error thresholding techniques followed by a local cost function adopted to get closed boundaries and to reduce the complexity of the algorithm. Next, they applied bezier-spline curve fitting technique to draw the smooth boundary from the local cost function.

Furthermore, region-based segmentation establish that the data may be initially subdivided into regions based on a grouping criteria, which are then merged together, or regions may be grown through the inclusion of additional data into the region.[16], [13] and [29] are three such examples of seeded region-growing for skin lesion segmentation, with the last one also combining Otsu's method in their segmentation algorithm. However, frequently these techniques have problems when handling boundaries of low contrast.

Another type of segmentation can also be achieved through the application of artificial intelligence principles, which are anchored by human-based learning and reasoning. As an example, [22] employed supervised learning together with random walker algorithm while [10] used a self-generating neural network. Sometimes, this kind of approach may provide faster computation and better performances.

Segmentation may also be achieved through the application of active contours, a technique involving the detection of object contours using curve evolution techniques. In parametric deformable models, the curve deformation is driven by energy forces. The gradient vector flow (GVF) used by [28] is an example of this model,

which applies an external energy based on the extrapolation of the gradient vectors to create a binary mask containing only the tumour. However, these models generally do not adequately handle the presence of large curvatures and topological changes.

Additionally, geometric deformable models aim to track the topological changes of the curve during segmentation and are less dependent on the choice of the initial curve, and allow the estimation of geometrical features of the curve. [11] for instance used a rectangular-grid based on the so called Freeman Chain code to create a geometric curve able to detect boundary pixels and allow to perform segmentation.

Finally, there exist several other algorithms that deviate from the standard ones, which is the case of [17], who found the global minimum between two Gaussian probability distributions in order to allow separation of the lesion from the skin or even [12] who created an image using Delaunay Triangulation and merged it with a filtered image of the skin in order to obtain their final binary image. As for [20], they decided to perform full manual segmentation, an approach that many other authors sometimes apply along with automatic segmentations methods since some of the experimental images are not accurately segmented and they want to obtain more accurate borders.

Without doubt, all these approaches have their advantages and drawbacks. However, it should be noted that most of the algorithms are tested on various fairly small and different datasets. In fact, ground-truth definitions and evaluation metrics differ from study to study, which makes it very difficult to provide unified results for all the tested algorithms. The performance assessment for these algorithms is not trivial, especially based only on the results reported by the authors, a really uncertain way of defining their strengths and weaknesses.

3.3 Features Extraction

Many works can be found on skin lesion feature extraction in the literature. This process represents a crucial step in most CAD systems since the results obtained show that the performance of classifiers is greatly dependent on the selected features. In fact, we have to rely on these so-called features of a lesion to correctly diagnose a lesion, which only emphasizes the importance of an extended analysis on feature categorization. In essence, features extraction is a process that aims to transform the input data into a set of distinctive properties. They have to be very carefully

chosen, in order to perform the desired task using the reduced representation instead of the full size input. Most automated systems, in order to classify a lesion, aim to extract such features from the images and represent them in a way that can be understood by a computer.

In this thesis, our goal is to develop a computer program able to automatically extract and analyse skin lesion features. Thus, our intent is to gain perspective regarding the existing approaches in feature extraction and to obtain a complete source of references on the descriptors of interest. We decided to start by evaluating the categories of features that the authors extracted, which can be seen in Table 3.3. However it must be noted that this table does not contain a complete list of articles in all categories, but only those that appeared in the scope of our survey.

Colour	*Texture*	*Border*	*Geometry/Shape*	*References*
✗	✗	✗		[10]
	✗			[11]
			✗	[12]
✗	✗		✗	[13]
	✗		✗	[14]
	✗			[15]
		✗		[16]
✗	✗		✗	[17]
✗	✗			[18]
✗		✗	✗	[19]
			✗	[20]
✗	✗			[21]
✗	✗		✗	[22]
	✗	✗	✗	[23]
✗	✗		✗	[24]
✗		✗		[25]
✗	✗			[26]
✗		✗	✗	[27]
✗		✗		[28]
✗	✗		✗	[29]

Table 3.3: Types of features extracted by other authors.

These categories of features are among the anatomical and quantitative attributes that dermatologists acknowledge are important for diagnosing melanomas. The majority of the papers referenced in Table 3.3 dedicated efforts to threes categories in particular, among them, colour, texture and shape features, which turned out to be on 65%, 65% and 55% of the articles, respectively. Other approaches like border

feature extraction have also been used but not so often as the previous ones, in fact this type was present in 35% of the articles.

Overall, we get an indication of the distribution of research efforts in relation to specific categories of features extraction, and further forward this will be taken into account. After performing this overview on the existing approaches we have to analyse them and provide specific information regarding the methods and properties selected for each of these types of extraction.

We start by describing the extraction of colour features, which play a vital rule in the early diagnosis of skin lesions. Colours are generally device-dependent, and since we are using the dermatoscope to obtain all our data set of images we can rely on the use of colour features for the characterization of skin lesions, as it allows this process of characterization to be repeatable, most of all invariant under varying viewing conditions such as surface orientation, illumination direction and illumination intensity. In Table 3.4 we present the references for some of the its most common techniques.

Methods	*References*
Colour spaces:	
RGB	[10],[17],[24],[27],[21],[29]
L*u*v	[10]
L*a*b	[19],[13],[21],[29]
HSV	[18],[24],[21],[29]
HSL	[27]
Colour variance:	
Minimum & maximum	[17],[24]
Mean	[17],[24],[27],[29]
Standard Deviation	Idem[1]
Other features:	
Histogram distances	[10],[17],[24], [27],[19],[13],[18],[21]
Centroidal distances	[10],[19],[13]
Concentricity	[19],[13]
K-means	[25],[26],[28]

Table 3.4: Colour features

It's important to refer the use and study of several colour spaces, more particularly RGB, HSV, HSL, L*u*v* and L*a*b colour spaces. [10], [17], [24] and [27] decided

[1]Something mentioned previously, in this case the previous references.

to extract features from the RGB colour space, as they used its histogram distances. [17] and [24] also measured colour variance using the minimum, maximum, mean and variance of some of its RGB channels and [10] calculated centroidal distances using this colour space. Meanwhile the L*u*v* colour space was used by [10] to obtain histogram distances while the l*a*b colour space was used by [19] and [13] for the same purpose of obtaining colour feature set histograms of the components of this colour model, with the last also extracting centroidal distances and concentricity features. Unlike the RGB model, L*u*v* and L*a*b colour spaces are designed to approximate human vision, matching perceptual colour differences with euclidean distances. It aspires to perceptual uniformity, and its L component closely matches human perception of lightness. Thus, it can be used to make accurate colour balance corrections by modifying output curves in the colour opponent components, or to adjust the lightness contrast using the L component.

Additionally, some authors like [18] used the HSV colour space histograms and some like [24] obtained the minimum, maximum, average and standard deviations of the Hue and Value channels in the HSV colour space. Others [27], opted to calculated the mean value and standard deviation histogram from the luminance colour plane of the HSL colour model. The advantages of using HSV or HSL, is that unlike RGB, they allow separation of the image intensity from the colour information, which is often useful in computer vision for various reasons, such as robustness to lighting changes, or removing shadows. Otherwise, if we decide to perform an histogram equalization of a colour image for example we will get very strange colours if we do not take into account the colour components alone.

Furthermore, it's worth mentioning some authors who combined and tried to use multiple colour spaces, [21] for instance calculate one histogram per component of the RGB, HSV and L*a*b* colour spaces, in order to approximate the probability distribution of each colour component, whereas [29] used the same colour spaces to extract the mean and standard deviation values over particular channels while also calculating centroidal distances, and finally [22] calculated the mean, standard deviation and reciprocal of coefficient variation of the values in S and V from HSV and L* of L*a*b*. Some articles decided to base colour-feature extraction by focusing on the shades of colour that dermatologists usually identify and analyse, like light-brown, dark-brown, white, red, blue, and black shades of colour for example. [25], [26] and [28] utilized the K-means algorithm to provide locally adapted dominant colors and the corresponding percentage of occurrence of each colour within a certain neighbourhood. In addition they also used a technique based on perceived color diferences to calculate the colour symmetry of the lesion area. [13] for instance

decided to extract information on the number of colors present within a lesion area, together with information about the presence of two specific colours: white and black.

Next in line, we have the extraction of another type of features, more particularly texture features, with Table 3.5 showing the references for some of the most common extracted texture features. Some authors like [27], [23] decided to quantify texture through the analysis of several statistical moments from the intensity histogram of the data. Moments such as the variance to analyse the visual perception of roughness, skewness to measure asymmetry or kurtosis to correlate uniformity were all taken into care, as they are crucial for the detection of malignant skin lesions.

Another approach was the use of grey-level co-occurrence matrix measures (GLCM) to obtain some of these features. This matrix was used by [10], [14], [22], [24], [17], [13] and [29], and basically it characterizes the texture of an image by calculating how often pairs of pixel with specific values and in a specified spatial relationship occur in an image, and then allowing to extract statistical measures from this matrix. The majority of the features extracted by these authors, using this matrix, were measures like entropy, contrast/inertia, energy, inverse difference moment, homogeneity, maximum probability, dissimilarity or even correlation. Means and standard deviations of marginal distributions derived from the co-occurrence matrix may also be used to construct other texture metrics.

Methods	*References*
Statistical moments used:	
Variance & skewness & kurtosis	[27],[23]
Co-occurrence matrix measures:	
Entropy & contrast/inertia & energy	[10],[14],[22],[24],[17],[13],[29]
Inverse difference moment & homogeneity	Idem
Maximum probability & dissimilarity	Idem
Run-length statistics:	
Run percentage	[13]
Mean & standard deviation	[14],[11],[15], [19]
Rare Approaches:	
Wavelets	[23],[26]
Local binary patterns	[18],[26],[17]
Histograms of gradient magnitude and phase	[21]

Table 3.5: Texture features

Furthermore, there are another type of features called run-length statistics that also provide useful information, via consecutive pixels that have the same value. For every angle, the number of runs of a certain length at a certain grey-scale is used to construct a run-length histogram. The run percentage, and its mean and standard deviation, may be used for texture quantification like [13] did, based on a grey level run-length matrix (RLM). In most of the articles authors used the previous referred matrices to extract mean and standard deviation features which was the case of [14], [11], [15] and [19]. Additionally, some authors applied some unusual methods to extract structural and statistic information. [23] and [26] for instance decided to resort to wavelets, unlike [18] who decided to use a visual descriptor known as Local Binary Patterns or [21] who computed histograms of gradient magnitude and phase for all the lesion pixels.

Besides the methods described, there is the third type of attribute for feature extraction, the border one. In Table 3.6, references related to some of the most common extracted border features are presented. In this case, there are no common approaches, there's a lot of diversity on the methods applied.

Methods	*References*
Border quantification features:	
Spatial & frequency domains	[23],[16]
Distance map	[19]
Fractal dimension & edge abruptness	[14],[25],[28]
Pigmentation transition	[25]
Sample statistics:	
Equivalent circular diameter & mean intensity variance	[25]
Mean & standard deviation of span, depth and thickness	[27],[10]

Table 3.6: Border features

[23] and [16] tried to analyse the border's spatial and frequency domains to extract the main characteristics of the lesion boundary, while [19] used distance map to capture the ondulation and the angular characteristics of the lesion margin. [25] and [28] exploited the extraction of several parameters that might relate to border irregularity fractal dimension, edge abruptness, and pigmentation transition while also extracting important border quantification features such as equivalent circular diameter, mean intensity variance, and the centroid of each lesion's border region. Others took advantage of sample statistics to employ this extraction like [27] and [10] who employed mean and standard deviation of span, depth or even average

thickness.

Finally, shape features may also be extract from the image. Some images in our dataset, especially malignant melanoma ones (sometimes too big to the dermatoscope lowest zoom), may contain incomplete lesions, and despite these cases being rare, shape features may not be the best approach. Therefore, we should be careful in the way we use it. In the following, we describe a set of widely used shape features effective on skin lesions, also presenting them in Table 3.7 along with the respective references.

Methods	*References*
Shape measurements:	
Bounding area & convex area & filled area & solidity	[14],[12],[13],[19],[29]
Asymmetry	[17],[29],[14]
Diameter & Perimeter	[14],[13]
Geometrical parameters:	
Compactness	[13],[14],[29]
Aspect ratio	[29]
Eccentricity & sphericity	[13]
Variance of the radial distance distribution	[13],[17]
Frequency information:	
Wavelet/Fourier domains	[20],[23]

Table 3.7: Shape features

Geometrical features have been used mainly to describe lesion's outline, as its irregularity usually indicates malignancy. Thus, the ROI shape may be analysed according to some common geometrical parameters like [14], [12], [25], [19], [24], [29],[27], [13] and [17] decided to do. Those features are based mostly on such properties of an object such as convex area, filled area, solidity, bounding rectangle area, perimeter, diameter, asymmetry, or geometric moments: aspect ratio, variance of the radial distance distribution, compactness, elongation, eccentricity or even sphericity. Otherwise, shape can also be described in terms of frequency information like [20] and [23] did, either by wavelet formalism or fourier transformation. Important if we are interested in analysing the radial distance to detect highly irregular borders with many notches. Also this approach, according to [25], can also help in the detection of pigment networks and dots type differential structures. Finally, [22] decided to extract shape features but for a more particular study, to analyse uniformity and smoothness of the orientation change in streaks.

3.3.1 Feature Selection

Once the features have been determined feature selection is the next important step, a procedure to be carried out prior to lesion classification for dimensionality reduction purposes. Feature selection consists in reducing data dimensionality by rejecting redundant, unimportant, or noisy features, thus resulting in increased prediction accuracy, less complex classifier models, and better computation efficiency. However, this reduction is not trivial because eliminating redundancy among feature descriptors may adversely affect their discriminatory power.

Feature selection algorithms may be divided into two main categories: filters and wrappers. Filter methods rely on general characteristics of the data to select a subset of features without involving any learning algorithm. They are usually fast, which allows to compare several alternative methods within an optimization framework. Out of numerous available filters, three are worth noticing due to their satisfactory performance on various data sets: ReliefF, mutual information based feature selection (MIFS) and correlation based feature selection (CFS). On the other hand, wrapper methods use the prediction performance of a predetermined learning algorithm to evaluate the goodness of feature subsets. This means that if we want to use wrappers like recursive feature elimination on a given data set, the target learning algorithm should demonstrate satisfactory results for the original data set, as wrappers are based on feedback principle. However, since some features extracted in this study might be irrelevant or redundant as well as due to class imbalance, wrappers will not likely fulfil those restrictions. For that reason, most of the authors adopted the filter methodology for this kind of study. As numerous features extracted are strongly mutually correlated, many authors decided to use the CFS filter for feature selection as it takes into account not only relationships between features and the decision class, but also relationships between features themselves.

The number of selected features is a parameter which requires a careful tuning. As a matter of fact, we should take into care that too few features may prevent classifiers from distinguishing between various classes whereas too many features impose risk of overfitting, a situation when a model excels in classifying training data but fails to generalize knowledge and hence misclassifies new samples.

Other approaches besides feature selection were also applied by some authors, to decrease the computational load incorporated into the classifier. In particular, principal component analysis (PCA), an unsupervised technique for dimensionality reduction that finds out which features are important for best describing the variance

in a data set. It detects the variance structure in the data and identifies the directions along which the data subspace exhibits high variance. Basically it tries to reduce dimensionality by exploring how one feature of the data is expressed in terms of the other features (linear dependency).

3.4 Classification

Lesion classification is the fourth and final stage of our algorithm's work-flow. During classification given objects are assigned into a predefined group or class based on a number of observed attributes related to that object. These attributes, known as explanatory variables, are the features that are fed into the classifiers, and that were previously described in Section 3.3. Depending on the system, the output of lesion classification can be binary (i.e.,malignant/benign), ternary (i.e.,melanoma/dysplastic nevus/common nevus) or even more, if we want to identify several skin pathologies.

A classifier is a mathematical function, which is modelled, by giving it a number of examples, each belonging to a certain class. They map input data to a category, and are able to process large amounts of samples collected from positively diagnosed and negatively diagnosed patients and use this medical data to greatly enhance diagnosis. Hence, they produce valuable information, that may be of vital importance, since these models can accurately predict new, previously unseen examples. Doctors and practitioners can benefit from this technology since these models can identify patterns or specific features that distinguish them and therefore provide reliable future decision-making.

Most of the authors relied on a machine learning technique to classify the skin images, which can either be supervised (all data is labelled and the algorithms learn to predict the output from the input data) or unsupervised (all data is unlabelled and the algorithms try to learn how to establish the existence of clusters or classes in the data). But how do we know what is the best machine learning algorithm to choose for our classification problem?

In Table 3.8, we present the classifiers that each article, aside some exceptions (further detailed), elected to test. According to this table, we can easily realize that supervised machine learning algorithms are largely preferred to unsupervised approaches, matter of fact the last approach wasn't even used, with the authors prefering to use the following classifiers: support vector machines (SVMs), artificial

neural networks (ANNs), adaptive boosting (AdaB), k-nearest neighbours (k-NNs), random forestss (RFs), logistic model trees (LMTs), linear discriminant analysis (LDA), decision trees (DecTs), naive bayes (NBayes). Above all, this is related to the nature of the classification problem, and to the high diversity of dermoscopic features that can point to the malignant or benign nature of a lesion. Thus, there are many sample lesions whose corresponding established diagnosis partially or completely contradicts the observed dermoscopic features, which makes it important to teach the classifier how to recognize such unusual manifestations.

SVM	ANN	DecT	RF	LMT	AdaB	NBay	k-NN	LDA	*References*
	✓								[10]
✓									[11]
			✓		✓	✓	✓		[12]
✓		✓					✓		[13]
✓	✓				✓				[14]
	✓								[15]
✓			✓		✓				[17]
✓									[18]
✓									[19]
✓									[20]
					✓				[22]
✓			✓	✓		✓			[23]
✓							✓	✓	[24]
✓									[25]
✓									[26]
	✓								[28]
✓									[29]

Table 3.8: Classifiers explored by the authors

Nevertheless, we cannot forget that even thought we can choose the best possible classifier, performance will always critically depend on the selected feature descriptors and the learning procedure. Therefore, the comparison of classification approaches should be performed on the same dataset and using the same set of descriptors in order to give optimal results. After summarizing classification results reported by others authors we decided to briefly understand some of these classifiers advantages and disadvantages. We analysed them in categories, although its quite hard to correctly detect them, but most importantly we focused on the articles that compared several algorithms without taking into account specific implementation characteristics.

As we can understand just by looking to Table 3.8 SVM was the most used classifier,

seeming to be one of the most popular techniques. SVM is known for achieving high accuracy while presenting nice guarantees regarding overfitting. With an appropriate kernel they can work well even if our data isn't linearly separable in the base feature space. However, SVM has a big disadvantage because it has several key parameters that need to be set correctly to achieve the best classification results for any given problem. A problem that can be kind of annoying since the user has to experiment several parameters: SVM type, kernel type and kernel-specific parameters. In terms of comparisons, in [13] it showed best overall performance than k-NN and Logistic Regression, while in [14] and [23] it was outperformed by an ANN and a RF respectively.

Additionally, another popular technique is the use of ANNs, a non parametric model that is easy to use and understand compared to statistical methods. ANNs have the remarkable ability of capturing non-linear and complex underlying characteristics from complicated or imprecise data, which is great for abstract problems like image recognition. ANNs take a different approach to problem solving because they cannot be programmed to perform a specific task. Instead of following a set of instructions they process information in a similar way the human brain does, learning by example. However, there are some cons associated to it, as they normally only work well with large data sets and also can be very unpredictable since they follow a black box learning approach that makes it hard to deal with uncertainties. Regarding the appliance of these networks several training methods like FeedForward and Back Propagation were used to implement them but the fact that brings the most interest was the use of a Deep Learning Neural Network by [14] which performed better than a SVM and AdaB classifiers. This type of network is an ANN but with multiple hidden layers that allows to achieve a high level of learning with low supervision.

Furthermore, we can also refer other types of classifiers based on the prediction of discrete classes and numeric quantities which is the case of tree induction methods. They are specifically designed to discover complex interactions among features, and can also give us the idea of how important a particular feature was for making a tree. When we think about tree induction techniques we immediately think about Decision Trees. This method is very useful because of its easiness to use and interpret. In comparison to most methods in classical statistics, decision trees are not based on any probability density function, also being non-parametric, so we do not have to worry about outliers or whether the data is linearly separable. However, they are very prone to overfitting and suffer from high variance (meaning that slightly different data might lead to a very different decision tree), but that is where random forests comes in, an upgraded version of the decision trees method.

Random forests are a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance. Essentially it takes the majority vote of the classification of all trees to predict the class of an observation, the problem here is that extreme events are very rare which makes them under-represented in the data and the majority vote might be too strict. This comes at the expense of a small increase in the bias and some loss of interpretability, but generally greatly boosts the performance in the final model. Additionally, we still have the logistic model tree classification model, a method that combines logistic regression and decision tree learning. These two schemes have complementary properties as the linear regression functions at the leaves provide high bias and low variance while tree induction fits more complex models which results in lower bias but higher variance.

In terms of comparison, on one hand both decision trees and logistic model trees demonstrated marginally worse performance against other classifiers. The first one was outperformed by SVM in [13] , while the second had inferior performance that a Random Forest in [23]. On the other hand Random forests despite performing slightly below AdaB in [12] they seem to be stealing the crown and gaining popularity, since in terms of comparison they outperformed SVM and AdaB in [17] while in [23] they were superior than LMT, NBayes and SVM.

Moreover, we also have Adaptive Boosting as a powerful classifier that works well on both basic and more complex recognition problems. AdaB uses a committee of weak base and inaccurate classifiers to vote on the correct class, allowing to create a highly accurate classifier. This technique is based on an iterative approach known to provide good generalization, which makes it less susceptible to overfitting than most learning algorithms. Yet sometimes, this method might be sensitive to noisy data and outliers. AdaB was evaluated against NBayes, k-NN and Decision Trees in [12] and showed superior performance. Nonetheless that didn't happen in [14] and [17] were AdaB was marginally worse than an ANN and a RF respectively.

As a further matter there are still some other techniques less used that are worth mentioning, such as Naive Bayes, k-NNs and LDA. The first one, is quite useful if its conditional independence assumption actually holds, as it allows it to converge quicker than discriminative models like logistic regression, and enables the use of less training data while also reducing overfitting possibility. This method is a good bet if we want something fast and easy that performs pretty well, yet in comparison to other classifiers it always felt short like it happened in [12] and [23] where it was outperformed by AdaB and RF respectively. That may be explained by the fact

that it cannot learn interactions between features.

The second one, is a powerful tool that stores all available cases and classifies new cases based on distance functions. It relies on the use of big training sets, otherwise they will potentially overfit since they are a low bias/high variance classifier. k-NN is an non parametric algorithm, pretty useful as most of the practical data does not obey the typical theoretical assumptions made. Its lazy learning means that it does not use the training data points to do any generalization. In other words, there is no explicit training phase or it's very minimal. This technique has one big disadvantage associated to it since its based on distance learning, and sometimes its not clear which type of distance and attributes should we use in order to produce the best results. When compared to other articles its performance was not the best since it was outperformed by a SVM and an ANN in [12] and [13] respectively.

The third and last one is a technique that seems to be quite unpopular for skin lesion classification since it was only applied once by [24] and to be evaluated in a decision template combination rule. It is a simple, mathematically robust technique that often produces models whose accuracy is as good as more complex methods. This method can be used in pattern recognition and machine learning to find a linear combination of features that best separates classes. Even so, like other classifiers it is sensitive to overfit and its validation might be problematic.

On a final note, in Table 3.9, besides some exceptions, we present the classifiers that achieved the best results in each article. The exceptions are related to [21], who used a binary classifier, but did not specified the one, and both [16] and [27], who opted to use a scoring classifier, mostly based on the rules used by dermatologists described in Subsection 1.1.4, where several parameters are taken into account to give a final score/classification. All these classification methods were tested on real data and compared to human diagnoses using statistical measures such as sensitivity[2], specificity[3] and accuracy[4]. Nonetheless, and despite all these comparisons, it's still difficult to establish an absolute hierarchy in the performance of these classifiers. The reason for this, besides the marginal differences in the numerical evaluation results, lies in the structure of the comparisons themselves whether we have different feature or datasets, different learning procedures or classifier parameters. For that reason, and concerning our best interest of achieving great sensitivity we should be prepare to test out a couple different classifiers, making sure to try different parameters within each algorithm as well, in order to select the most appropriate

[2]Rate of sick people who are correctly identified as having a condition.
[3]Rate of healthy people who are correctly identified as not having any condition.
[4]Rate of correctly classified images.

classifier.

Classifier Employed	Sensitivity	Specificity	Accuracy	*References*
ANN	83.0%	95%	91%	[10]
SVM	86%	96%	96%	[11]
AdaB	93%	87%		[12]
SVM			92%	[13]
ANN	94 %	90%	92%	[14]
ANN			94%	[15]
Scoring System			90%	[16]
RF	98%	70%		[17]
SVM	84%	94%		[18]
SVM			95%	[19]
SVM	90%	82%	87%	[20]
Binary[5]	94%	77%		[21]
ADA			76%	[22]
RF	84%		91%	[23]
k-NN/SVM/LDA[6]			80%	[24]
SVM	88%	91%		[25]
SVM	91%	94%		[26]
Scoring System			77%	[27]
ANN	67%	80%	75%	[28]
SVM	93%	92%		[29]

Table 3.9: Best classifiers applied along with the respective results

[5]Binary classifier, they don't specify it, could be a RF, DecT, SVM, amongst many others.
[6]Decision template combination rule was applied, so there's not a specific better classifier.

4

Methods

Throughout the course of this work, we tried to develop a computer-assisted prediction model utilizing the most significant methods we found in the literature singled out by our systematic review. However, as expected, not all of them provided good outcomings, and while understanding which of them were the most adequate we also tried to create our own approaches. In the end, we came to the conclusion of which were the best procedures, the ones who allowed us to visually discriminate between benign and malignant skin lesions in the most efficient way. The proposed methodology is divided into five main stages, all of them described, in detail, on the following sections. The flow diagram of the implementation, which was previously explained in Section 2.2, is outlined below.

① Image acquisition
② Pre-processing
③ Segmentation
④ Feature extraction
⑤ Classification

The application of these procedures was implemented using MATLAB R2017a, a high-performance language which integrates computation, visualization, and programming in an easy-to-use environment. The images that will be shown during the following stages of methodology were all obtained using this tool.

4.1 Image acquisition

The first stage of our system was image acquisition, an essential phase for the rest of the algorithm, considering that the images need to be acquired satisfactorily for the remaining components of the system to be achievable, otherwise the results will not be reasonable. In order to work with high quality images, we resorted to Coimbra's

University Hospital Centre, where we acquired a dataset of 114 dermoscopic images, alongside their ground truth, provided by histological diagnosis.

The dataset was comprised of both melanoma and other benign skin lesion images which were taken from different patients using a dermatoscope. Due to the existence of excessive unwanted illumination artefacts or lesions which didn't fit entirely within the image frame, 7.44% of the benign images and 27,78% of the melanoma ones were not reliable, consequently they had to be excluded, leaving us with 87 benign images and 13 melanomas. This was the final set of images used for us as a train and test bed to perform experiments and validate our proposed approach. All of them were 8-bit RGB colour images with 576 $\times$ 767 pixels as dimension.

4.2 Pre-processing

Pre-processing was the first big stage of our work, one that seems obligatory in the computerized analysis of skin lesion images, mostly because the majority of dermoscopic images contain extraneous artefacts and parts that are unrelated to the lesion itself, which need to be removed. Hence, the main goal of this phase was to improve image quality by reducing the presence of several unwanted artefacts such as hairs, air bubbles, small pores, shines, and reflections. This way we can keep the vital information and avoid affecting the image segmentation later on.

4.2.1 Histogram modification

One of most difficult tasks in these computerized analysis missions is the ability to provide a consistent automated system, and that was a main issue we had to deal with, since we needed our algorithm to perform equally well to all the images. Therefore, in order to solve this problem we decided to apply histogram modification to all our images based on a reference one, which we chose according to good lighting and contrast characteristics, as well as not having unwanted artefacts. In Figure 4.1 we show the reference image alongside a model image.

The purpose of applying histogram equalization relied on the attempt of transforming each image's intensity in a way that the histogram of their output intensity image became similar to the histogram of our reference image. This way we could apply threshold procedures trusting on the fact that our automatic techniques will be steady, since every image would have a similar histogram.

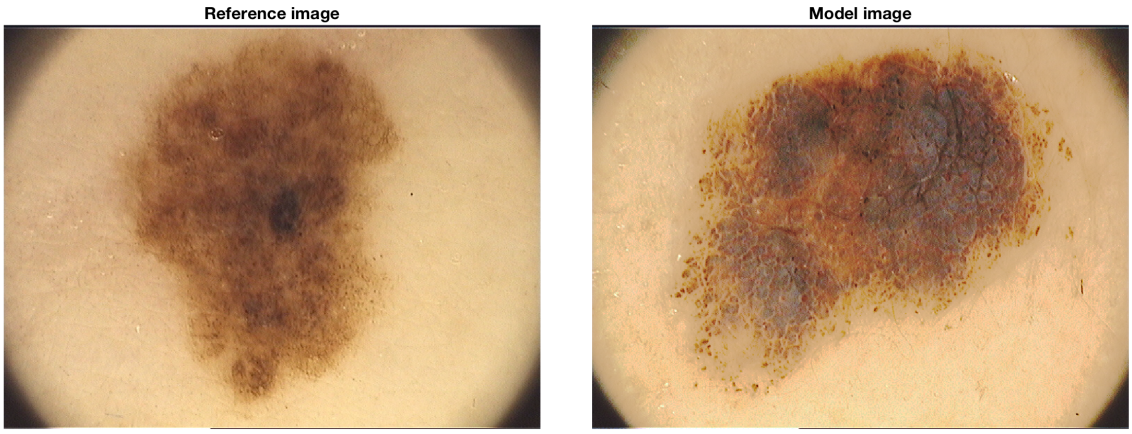

Figure 4.1: Reference image used for histogram modification side by side with a comparison image, as the legends highlights.

The results of the equalization are presented in Figure 4.2, where we can see the real effects of this procedure. On the left side of the image we have the reference's histogram, and on the right side we can see the model's original histogram (represented in blue) transformed into a new histogram (represented in orange), analogous to the reference's histogram.

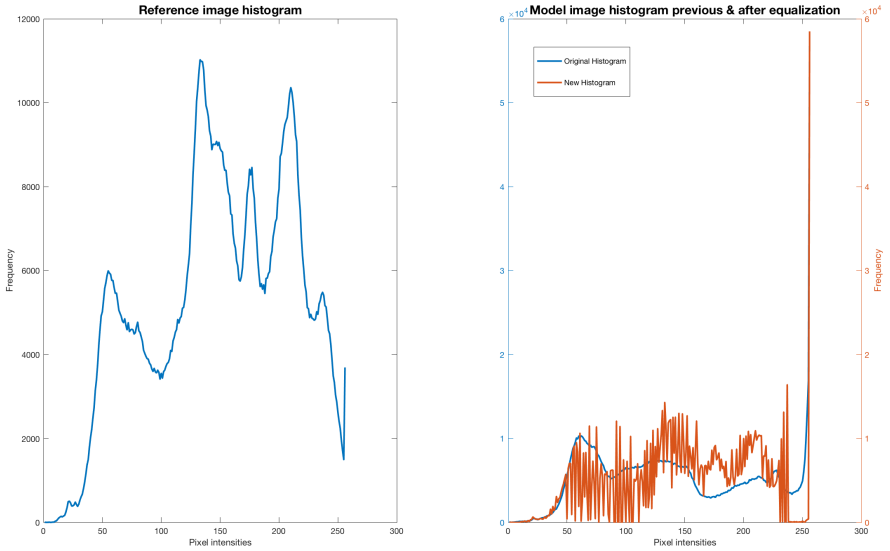

Figure 4.2: Reference's histogram side by side with the original and final histogram of the model's image.

4.2.2 Colour space transformation

The colour space transformation phase had the responsibility to provide the best possible plane for the detection and removal of unwanted artefacts. And in that

sense, we decided to use the L*a*b colour space instead of the RGB model, since the first correlates euclidean distance measures with perceived colour differences and the human visual system.

Owning to the fact that colour properties depend on a colour difference formula, which must be uniform, the use of L*a*b should provide more accurate results than other traditional colour spaces because of its perceptual uniform adaptation [26]. Subsequently, due to this ability of approximating human vision, all our RGB images were transformed into this colour space, hoping that we could accurately pre-process each skin lesion. It's worth mentioning that for the remaining steps of this stage, the luminance component (L*) of this colour space was used, as it closely matches human perception of lightness.

4.2.3 Various artefacts removal

After the colour space transformation, the next stage was image enhancement, or in another words, to reduce the influence of air bubbles, small pores, shines, reflections and other artefacts caused by the applied gel before image capturing. The idea was that if there was a transaction on edge detection of a source noised image, we would avoid locating other additional edges due to the presence of noise. Consequently, noise removal was a crucial process in our algorithm, therefore we invested careful thought in what technique we would use to suppress it. We came to the conclusion that the most appropriate method to smooth the image and reduce the intensity gradients[1] inside the lesion and in the surrounding healthy skin was a median filter technique [14]. This procedure replaces each pixel by the median value of the neighbouring pixels, and additionally since it's a non-linear filter, it allows edges to be preserved while removing the outlier pixels. After we selected median filtering as our choice, we decided that instead of using it directly for noise removal, it would be more useful to use it for noise detection. For this purpose, we used a simple thresholding formula, proposed by [15], which is described below:

$$\{(I(x,y) > \delta_{T1}) \wedge ((I(x,y) - I_{md}(x,y))\} > \delta_{T2}) \tag{4.2.1}$$

In equation (4.2.1), I represents the input image, while I_{md} represents the same image after a 11×11 median filter has been applied. If the left and right statements are bigger than threshold δ_{T1} and δ_{T2}, respectively, than it means that the pixel in

[1] Directional changes in the intensity of an image

matter will be classified as a reflection artefact. In Figure 4.3 we can see the outcome of this detection on a model image.

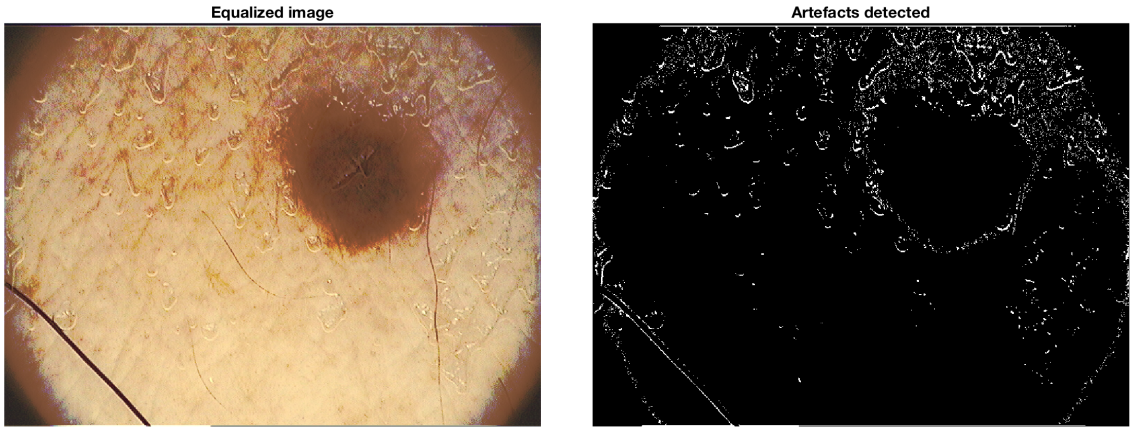

Figure 4.3: Detection of unwanted artefacts in a model image.

Afterwards, the detected artefacts were removed and an in-painting technique was utilized to repair the noise-occluded information. This operation was based on the removed pixel's neighbourhood, where the nearest edge pixels in 8 directions were scanned and their average intensity was used to replace the deleted pixels, a procedure illustrated in Figure 4.4. For the artefact detection, we used the luminance component and for repairing them, all three components were utilized.

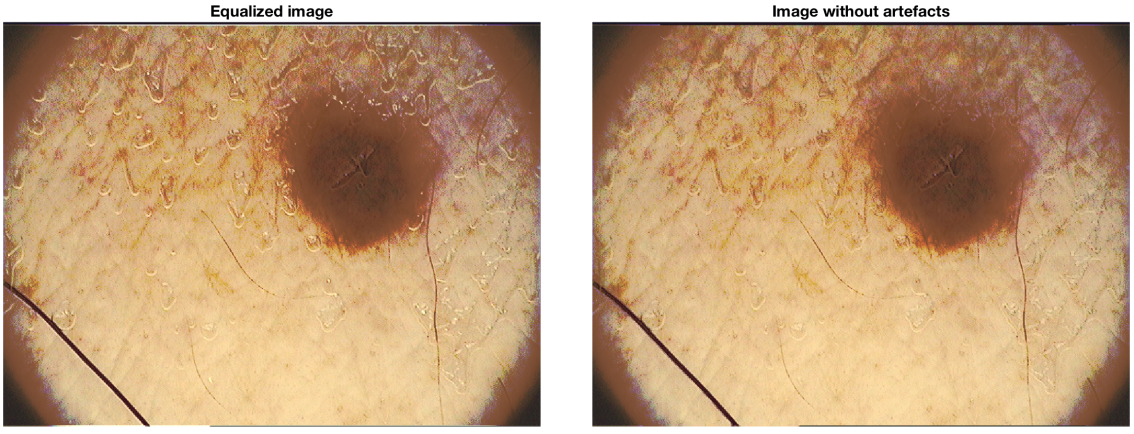

Figure 4.4: Artefacts removal and in-painting applied to Figure's 4.3 model image.

4.2.4 Contrast improvement

On the same subject of image enhancement, we also applied two extra procedures besides the removal of unwanted artefacts, with contrast improvement being the first

of them. For this end, we opted to use CLAHE, a pre-processing technique proposed by [14], which consists of an alternative method of adaptive histogram equalization. More specifically, it operates on small regions in the image, called tiles, and enhances each tile's contrast, thus, the output is more precise than improving the contrast of an entire image. In the end, a premeditated flat distribution histogram was obtained, with the results being shown in Figure 4.5.

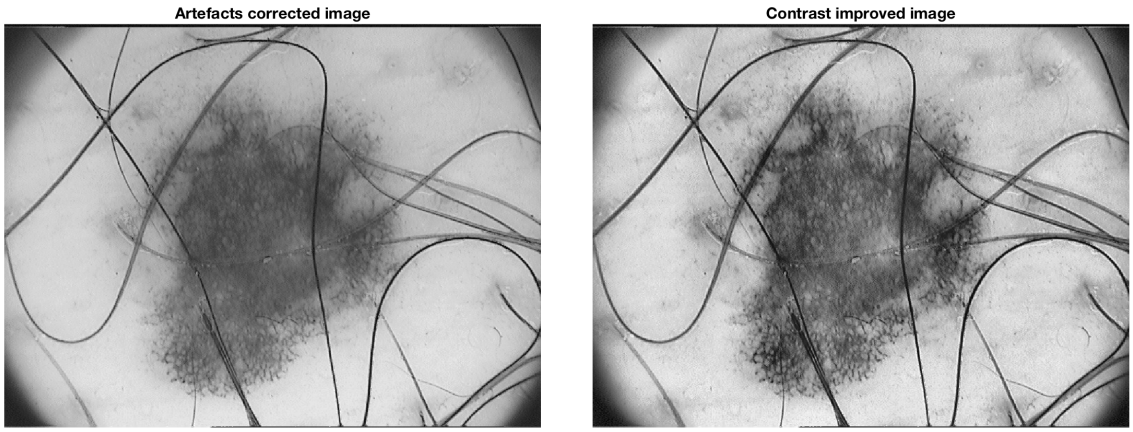

Figure 4.5: Contrast improvement of a model image after noise removal.

This time, histogram equalization had a different intent rather than the one described in Subsection 4.2.1, the goal was to stretch out the input histogram to produce an output whose histogram is approximately uniform, where the various pixel intensities are equally distributed over the entire dynamic range.

4.2.5 Correction of uneven illumination

Knowing that dermoscopic images often exhibit uneven illumination due to their acquisition process, we still had to correct this problem to facilitate further operations. For this reason, we decided to perform HTF, another generalized technique for image enhancement, suggested by [25], and described in detail afterwards.

The idea behind the HTF method was to divide an input image into illumination and spatial-distribution reflectance components, or, to put it another way, its goal was to linearly separate these components in the frequency domain. However, for that, their relationship had to be adapted to become additive rather than multiplicative, a procedure that was done by applying a logarithmic transform to the image. Afterwards, we still applied a 2-D fast Fourier transform on the logarithmic transformed image, in order to analyse these illumination and reflectance components in

the frequency domain.

Once we crossed these phases, we needed to effectively correct the illumination problem, by selecting the right frequencies to attenuate or amplify. In practice, we understood that the illumination component had low spatial variation, which meant that it had low frequency unlike the reflectance component who proved to have higher frequency thanks to its high spatial variation. Consequently, the next step in line was to allow high frequencies to get through while attenuating low frequencies, and that was done using a high-pass butterworth filter elaborated as follows:

$$\left(1 + \left(\frac{c}{(u^2 + v^2)^{0.5}}\right)^{2\times n}\right)^{-1} \tag{4.2.2}$$

In equation 4.2.2, u and v are the resulting coordinates from the previous Fourier transform, while c and n represent the cut-off frequency and attenuation coefficient of the filter, respectively. With that said, we applied the mentioned filter to the previous Fourier transform, and followed it with the computation of its inverse Fourier transform and exponentiation in order to get the final homomorphic filtered image. In Figure 4.6, we present the final outcoming image.

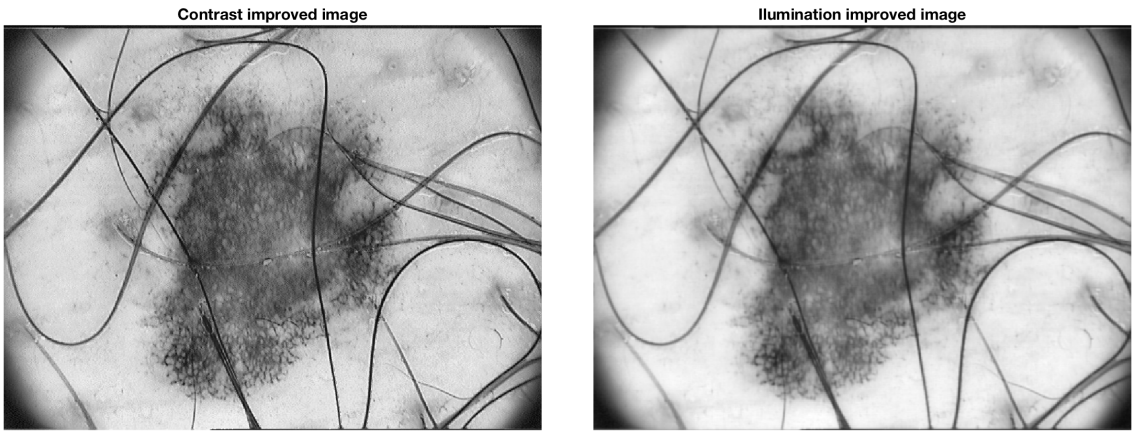

Figure 4.6: Ilumination correction of a model image after contrast improvement.

4.2.6 Hair removal

Finally, we arrive to the last phase of pre-processing, known as hair removal. Considering that hairs may cover several parts of an image and make the segmentation and feature analysis impossible, this last procedure was among the most necessary artefact rejection steps, if not even the most important. Our hair-repair system was divided into three steps: ① Hair detection ② Refinement ③ In-painting

Hair detection was the biggest challenge ahead of us, specially because most of the times hair had similar linear shape to the lesion's pigmented network, which could cause incorrect detections. A number of methods have been developed for hair detection in dermoscopic images, however we decided to create a different approach than the ones applied by the literature mentioned in Section 3.1.

Based on a concept idea proposed by [30], we decided to start our detection using a morphological image processing technique called black top-hat filtering. This procedure, as descried below, calculates the difference between an image and its morphological closing.

$$I - I_{mc}(s,\theta) \tag{4.2.3}$$

In 4.2.3 I represents the input image and I_{mc} exhibits the morphological closing of image I, whether s represents the structuring element to be used and θ its respective angle. Assuming that hairs, are thin linear structures, the top-hat operator was used with a line structuring element to detect them, which means that the morphological closing will merge together the line features in the image that are close together. The black top-hat operation was repeated several times with 16 different structuring element orientations, since the hair direction was not know. In the end, each outcome was an image, which contained elements that were darker than their surroundings and smaller than the structuring element, allowing us to distinguish hairs from other local structures.

Once we finished applying the black top-hat filtering techniques, the following step, as the coming condition suggests, was to use all the resulting images to find the maximum value at each pixel's (x,y) location, and compare it with a specific threshold, calculated using otsu's threshold method, a clustering-based image technique that gives a threshold which minimizes the intra-class variance of the background and foreground pixels [31]. Ultimately, if a pixel's value was smaller than this threshold, it would be removed, otherwise, the pixel was kept unchanged, ending up with a binary image as the final outcome.

$$Max_{j\in\{1,2,\ldots,N\}} \times I_{mc}(x,y) > \delta_{otsu} \tag{4.2.4}$$

Where I_{mc} represents the N black top-hat transformation images we previously obtained, with N being the number of different angle orientations we applied, and δ_{otsu} the implemented threshold which was calculated empirically.

Finally, we arrive to the last step of hair detection, which was based on thresh-

olding operations, like we display in equation 4.2.5. For this purpose, several sets of properties measurements were used, for each 8-connected component, present in the previously obtained binary image, including major axis length, eccentricity[2] and extent[3] properties.

$$\{maioraxislength > \delta_{th1} \wedge eccentricity > \delta_{th2}\} \vee \{extent < \delta_{th3}\} \tag{4.2.5}$$

In equation 4.2.5 we used three thresholds (δ_{th1}, δ_{th2}, δ_{th3}), which were calculated empirically, and the three property measurements, as previously mentioned (*maioraxislength*, *eccentricity* and *extent*). If the component's measurements allowed this condition to be true, then all the pixels belonging to this connected component were labelled as hair. In 4.7 we illustrate the final result of hair detection in a model image.

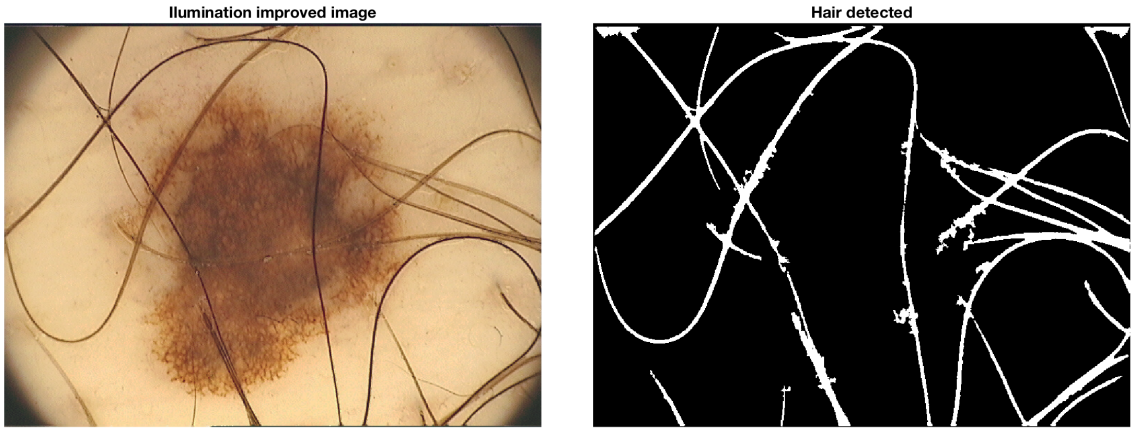

Figure 4.7: Hair detection performed on a model image.

The next step in line was to use morphological operations in order to refine these detected lines. Firstly, we had to correct some segmented hair lines, which contained contour or curvature like objects. For this reason, we applied thinning and pruning morphological conditions in order to get the hair-detected skeleton and remove some of these unwanted branches that could be noticed after hair detection. Furthermore, before correcting the occluded-hair, we also needed to smooth and fill some broken hair lines, therefore we used another series of morphological operations to perform these procedures.

We started by using a morphological closing function in order to link some lines

[2]Ratio of distance between the focus of the ellipse (that has the same normalized second central moments as the region) and its major axis length.

[3]Scalar that specifies the ratio of pixels in the region to pixels in the total bounding box.

which had gaps between them, following it with two more morphological operators in order to fill the holes in the image which defined the outline of each line, namely filling and dilation conditions. Last but not least, we wanted to filter some unwanted objects which had remained in the binary image, so we ended up the refinement stage applying an area opening function. It's worth reminding that all these transformations were performed using specific structuring elements, which were calculated empirically to better suit our purposes.

Afterwards, and in a similar way to Subsection 4.2.3, the detected hairs were removed and an in-painting technique was utilized to repair the hair-occluded information. Once again, this operation was based on the removed pixel's neighbourhood, where the nearest edge pixels in 8 directions were scanned and their average intensity was used to replace the deleted pixels, a procedure illustrated in Figure 4.8. For detecting hair-like regions, we used the luminance component of the L*a*b* enhanced image and for repairing, all three components were utilized.

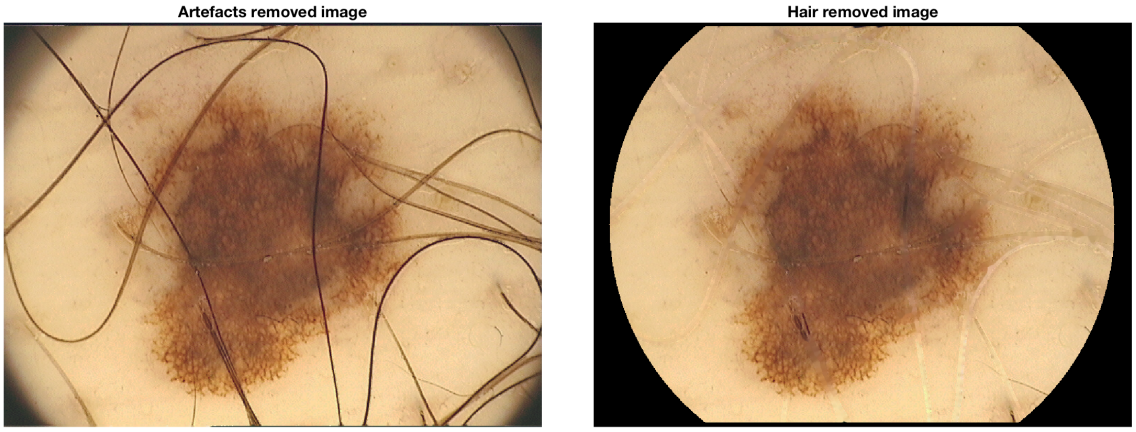

Figure 4.8: Hair in-painting results on the model image.

4.3 Segmentation

The next stage ahead of us was lesion segmentation, a challenging and crucial step in the computerized analysis of skin lesion images. The aim of image segmentation was to extract the lesion area from the healthy skin, a step achieved via a complete automatic method, which had the difficult task of extracting useful information to locate and delineate the lesion region present in the image. The border detection accuracy was of vital importance as it greatly affects subsequent feature extraction and classification. In order to provide a tool for segmentation, many procedures have been developed, with different methodologies being followed and also proposed

by researchers. We implemented and tested several of these methods, and decided to create a segmentation procedure based on the one purposed by [19], as it suited better our purposes. Our method's details will be precisely described below.

4.3.1 Otsu's threshold

Our segmentation algorithm started with the transformation of the previously obtained RGB hair removed image to the L*a*b colour space, in order to use the luminance plane component for the segmentation procedure, in a similar way to Subsection 4.2.2 of the pre-processing. Additionally, we applied a Gaussian filter to the respective channel, as a means to facilitate the border detection phase by increasesing the gradient on the lesion boundary and decreasing it inside a lesion or in the background. The Gaussian filter is depicted below:

$$\left(\frac{1}{2 \times \sqrt{2 \times \pi}}\right) \times \exp\left(\frac{(x - xc)^2 + (y - yc)^2}{2 \times \sigma^2}\right) \tag{4.3.1}$$

Where xc and yc represented the center of the gaussain filter while x and y were the coordinates of the pixels belonging to the $M \times M$ sized filter and sigma was the standard deviation parameter. Afterwards, we used Otsu's threshold to perform an initial segmentation and approximate the lesion localization. Pixels were classified as part of a lesion if their value was higher than this threshold value and were classified as background otherwise. Additionally, we had to elaborate a mask to deal with the white corners that resulted from otsu's threshold procedure, replacing them with black pixels. The resulting binary image is ilustrated in Figure 4.9.

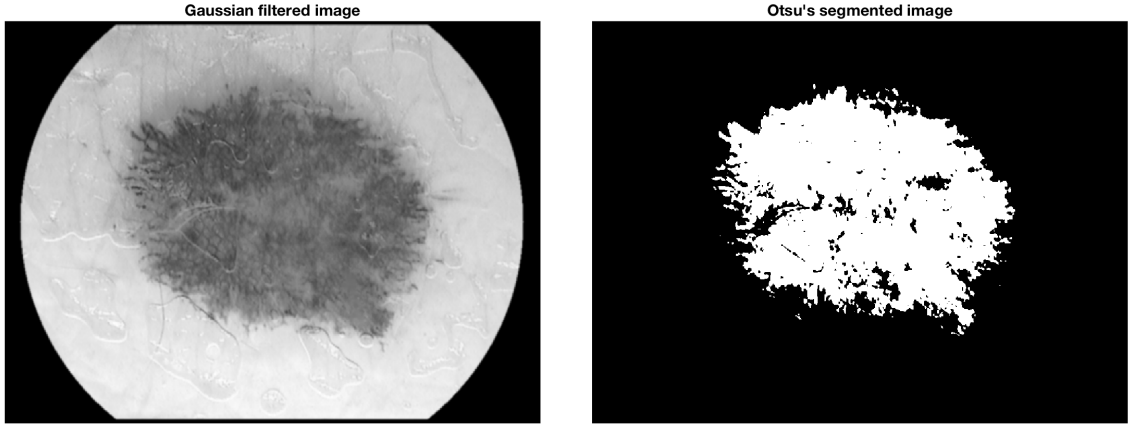

Figure 4.9: The Gaussian filtered image side-by-side with the outcoming binary image after Otsu's threshold.

4.3.2 Morphological refinement

After analysing the previous image it was clear that some of the edges of the lesion had irregular shape, therefore we needed to smooth them. In a similar refinement process to the one applied in Subsection 4.2.6, we performed several morphological operators, using a disk-shaped structuring element in all of them, as a means to preserve the circular nature of the lesion.

We started by applying a closing operation to fill the small gaps that had remained in the binary lesion, and followed it with an erosion operator, which enabled some unwanted regions to be separated from the edge of the lesion region. Additionally, we wanted to dilate the lesion area, since it had been thinned by the previous operator, hence, we used a dilation condition to expand the lesion area. Finally, after the dilation, we carried out a filing operation to make sure the hole regions were filled in, and immediately upon this step we found the largest region among the remaining ones, and kept it while eliminating other isolated regions. A final binary image was obtained in which the lesion was distinct from the surrounding healthy skin, as we can see in Figure 4.10.

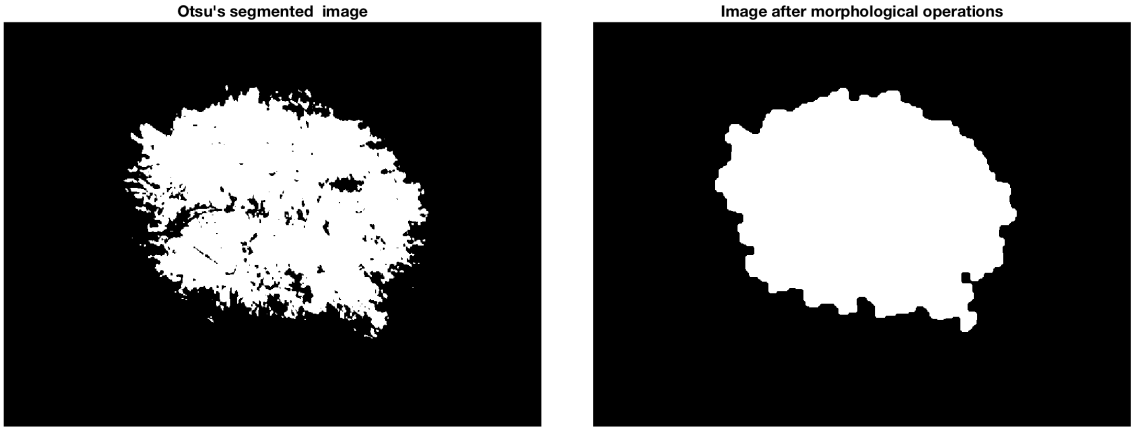

Figure 4.10: The outcoming binary image after Otsu's threshold alongside with the mask that resulted from the morphological refinement step.

4.3.3 Sparse-Field level-set method

Further on, we decided to perform an additional segmentation, using the previously refined mask as an anchor to this step. The goal was to apply an active contour function, a procedure that allows a contour to deform iteratively to partition an image into foreground and background regions. Active contours are often implemented with level set methods, which are widely used tools in computer vision because of

their power and versatility. However, sometimes, their implementation can be slow to compute, therefore we decided to apply a very efficient fast level-set algorithm, proposed by [32], called Sparse-Field method. This approach implements an active contour evolution, which combines both the efficiency of the parametric boundary tracing and the flexibility and robustness of the level set method.

The basic idea behind this concept was to start with an initial boundary shape position represented by the mask obtained in Subsection 4.3.2, and iteratively modify it by applying shrink/expansion operations according to the constraints of the image. Those contour evolution operations, were performed by the Chan–Vese model, which is based on the minimization of an energy function, and accomplished by the level set technique. On the left-side of Figure 4.11 we show the outcoming binary image.

Finally, we applied some subtle morphological operators (with a disk structuring element) to the output binary image, by using the dilation condition to expand the boundary to be larger, and the filling condition to correct the small gaps that had remained from the active contour segmentation procedure. The resulting image from this final enrichment represented the mask, that later on, would be used to segment the pre-processed image. On the right-side of Figure 4.11 we present the same obtained mask.

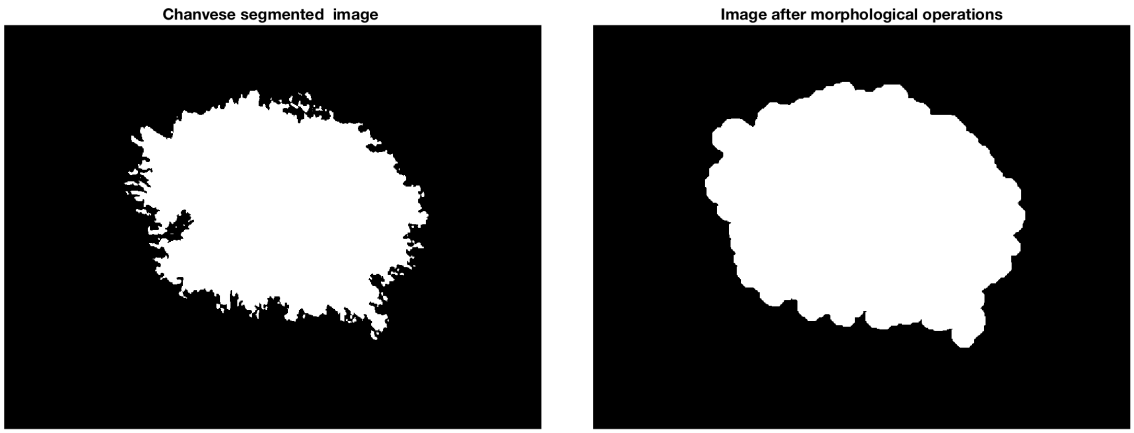

Figure 4.11: The Sparse-Field level-set segmented image before and after enrichment.

4.3.4 Freeman chain code

Finally, it's also worth mentioning an extra step we performed during the segmentation procedure in order to evaluate the effectiveness of our boundary pixels detection, named Freeman chain code.

The idea behind it, is that the single shape in the image could be described by recording a starting point on its outer boundary and then tracing the movements around the shape boundary. Knowing that in a continuous curve, each point is dependent on the previous one, the method devised by Freeman encoded the path from the centers of connected boundary pixels using a sequence of numbers between 0 and 7, where each digit represented a directional code describing a specific movement.

In brief, once the starting pixel for the chain code was determined, the algorithm continues tracing the boundary of the shape until we return to the starting point, generating a list of consecutive points, or to put it another way, a clockwise or counter-clockwise of adjacent pixels. A binary object contour, drawn using a chain-code procedure, is depicted on the left-side of Figure 4.12. As for the image on the right-side of Figure 4.12, it represents the final segmented image derived with the mask obtained in Subsection 4.3.3.

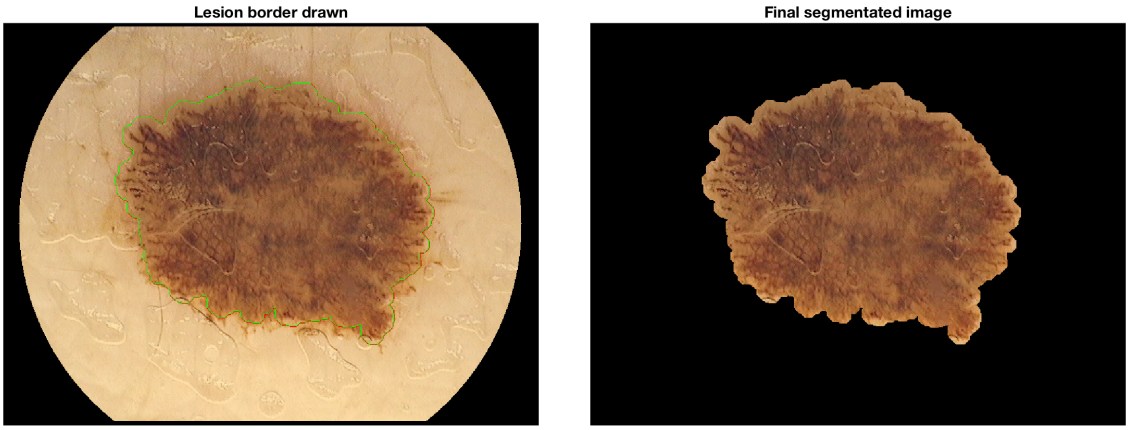

Figure 4.12: Freeman chain contour side-by-side with the final segmentation of a model image.

4.4 Feature Extraction

Feature extraction is one of the most crucial phases of skin lesion detection, since it has the hard job of ensuring that melanoma and benign lesions can be distinguished. Therefore, the aim of this procedure was to find several robust parameters, that combined together, could help us to correctly classify these images, since each attribute alone is not sufficient to diagnose a lesion precisely. The features were extracted over the entire lesion region using a global feature vector. In a similar fashion to the literature described in Section 3.3, we decided to extract four types of features, namely, colour, texture, border and shape features. The criteria of choice

was to find attributes from the referred literature which had proven their usefulness or that were commonly linked to the ABCD rule (Subsection 1.1.4). A particular problem we noticed in the related literature was that a significant number of studies did not report the details of their feature extraction procedure, thus, in order to enhance the reproducibility of this study, in the following subsections, we describe our set of extracted features, in detail. It's worth mentioning that we extracted 36 features from the overall feature extraction procedure.

4.4.1 Colour features

The analysis of a lesion's colour is a vital source of information, when determining a lesion's type, specially the malignant ones. In order to quantify the colours present in a lesion, the following features were extracted: colour occurrence, histogram analysis, and centroid distances. Knowing that the RGB colour representation does not allow reliable measuring of perceptual colour differences and their coefficients do not provide an intuitive description of colour, we used three different colour spaces to perform these extractions, according to specific criteria, that will be described further on.

We started by performing histogram analysis, a widely used colour feature descriptor which gave us two features. To that end, we employed two types of colour spaces, the HSV and L*u*v colour spaces. The first was used to handle photometric[4] and geometrical[5] variations, while the second was chosen because it provides perceptual uniformity [29].

To construct the HSV histogram we coarsely quantized H with 16 bins, S with 4 bins and V with 4 bins, while for L*u*v we used 4 bins for L*, 8 bins for u* and 8 bins for v, leading to a final descriptor size of 256 bins for each one of them. Note that for both of the referred colour spaces, instead of concatenating the histograms of each colour channel independently, we calculated the joint distribution of all the components, giving us a multivariate distribution (3-dimensional) of the colour features per colour space [18]. Finally, we had to analyse the shape of the two resulting multivariate distributions, and for that purpose we decided to use the kurtosis to measure if the pixel distributions were peaked or flat relative to a normal distribution. Ultimately, we extracted 2 features from this procedure.

Next in order, we had the extraction of several shades of colour, more specifically

[4]Shadow, specularities and changes of the light source
[5]Viewpoint, zoom and object orientation

their percentage of occurrence in each lesion. For that purpose, we determined that the L*a*b* colour space was the most obvious choice for this extraction, as it is designed to approximate the human visual system by providing a good correlation between perceptual difference of colours and measured colour distance.

Normally, to recognize early melanoma, dermatologists extract six shades of colour to evaluate a skin lesion, in particular, light-brown, dark-brown, white, red, blue, and black as [26] suggests. Hence, we decided to analyse this combination of dominant colours to effectively differentiate between lesions. To measure the occurrence of each and every one of these colours, we clustered the L*a*b colour space using the K-means clustering algorithm, where the euclidean distance between clusters was applied as the criteria for choosing the pair of clusters to merge at each step [13]. The outcome of this method on a model image can be seen in Figure 4.13, where every pixel is represented by the average colour of its class neighbours. In the end, 6 features were calculated corresponding to the percentage of occurrence of each colour.

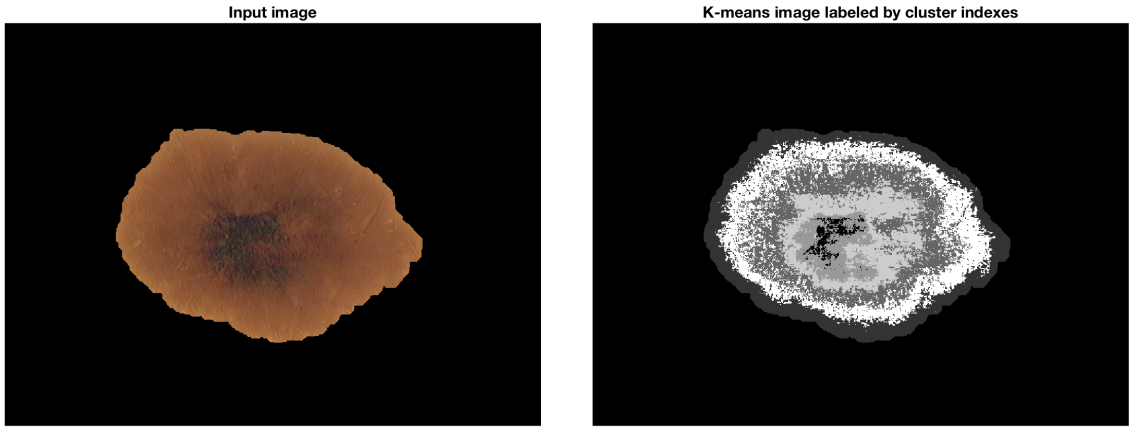

Figure 4.13: K-means result on a model image.

At last, we defined the centroid distances to extract the last colour features. For this reason, we required a colour space capable of dealing with images that were acquired in uncontrolled imaging conditions, therefore we decided to use the HSV colour space, extracting 1 feature for each of its component. Given a segmented lesion object, the centroid distance for a channel was calculated as the distance between the geometric centroid (of the binary object) and the brightness centroid of that channel. The brightness centroid represented the maximum intensity value of the corresponding channel. The idea behind this was that the centroid distance of a channel would be small, when the pigmentation in that particular channel was homogeneous, causing the brightness centroid to be close to the geometric centroid

[10].

4.4.2 Texture features

Texture analysis was responsible for giving us the information about the spatial arrangement of colours or intensities, while also evaluating the randomness in our images. In order to describe the quantitative properties of a lesion's texture, we decided to use a well-known descriptor for texture analysis, called gray-level co-occurrence matrix.

GLCM is a statistical method that scrutinizes texture characteristics that rely on the spatial relationship between pixels. In this approach, a co-occurrence distribution is calculated representing the occurrence probabilities of all pairwise combinations of the gray levels in a specified window, enabling texture features to be extracted based on statistical measurement of co-occurrence probabilities.

In this framework, we quantized the images to 32 grey levels and calculated co-occurrence probabilities given the distance of 2 pixels. In order to obtain rotation invariant features, the normalized GLCM was computed for each of the four orientations (0°, 45°, 90°, 135°) and the statistics calculated from these matrices were averaged [29],[24]. Although many statistics can be derived from the GLCM, only two gray level shift invariant statistics were used in our study, mostly because our overall feature extraction procedure is based on a global feature vector, and texture features are known to perform better for local feature vectors. The 2 extracted statistics were entropy and contrast, both described in the following equations:

$$\begin{aligned} Entropy &= \sum_{a,b}(a-b)^2 M_{ab} \\ Contrast &= -\sum_{i} P(x_i) log_2(P(x_i)) \end{aligned} \tag{4.4.1}$$

In the first of the above mentioned equations, we assume that GLCM is a square matrix M, where the (a,b)th entry of M represents information about the frequency of occurrence of such two adjacent pixels, where one of them has intensity a and another has intensity b. In the second equation P contains the histogram counts x_i. The entropy value of the image gives the randomness measure to characterize the texture of the image, whether the contrast value, as the name suggests, gives the intensity contrast between a pixel and its neighbourhood pixels over the entire image.

4.4.3 Border features

The next step in line was the extraction of border features, known to be a significant diagnostic information used by doctors. Border features were used to characterize the curvature function of the lesion border, a very important factor when evaluating a malignant lesion, since they have higher tendency to exhibit protrusions and indentations.

Firstly, we extracted our border features by building a time series for the lesion, an ordered sequence of values measured at successive equally spaced time intervals. Starting from an arbitrary pixel on the border, we calculated the distance between each border pixel and the centroid of the lesion (Equation 4.4.2), ending up with the corresponding boundary series [23].

$$distance_i = \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2)} \qquad i = 1,...,m \tag{4.4.2}$$

In the above equation m represents the number of pixels in the border while (x_i,x_i) and (x_c,x_c) are the coordinates of the ith boundary pixel and lesion centroid, respectively. The outcoming boundary series of a model image can be seen in Figure 4.14.

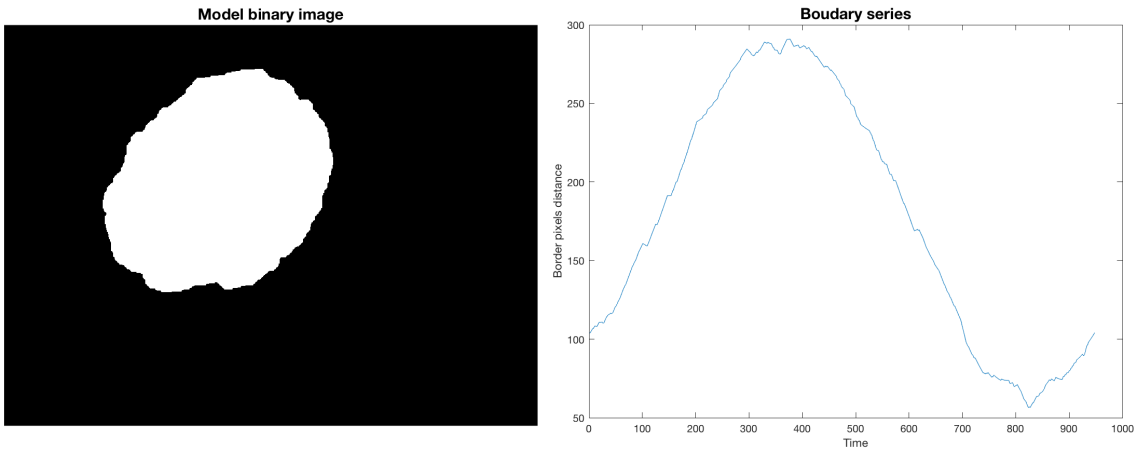

Figure 4.14: Boundary series of a model image.

Once we had the boundary series we decided to apply a histogram and a three-level wavelet transform to it, in an attempt to analyse both its spatial and frequency domains. The histogram was calculated using 10 bins, while the wavelet decomposition, into approximate and detail components, was performed up to the third level. The boundary series histogram and the three pairs of detail components that resulted

from the wavelet transform were the analysed signals. For the mentioned analysis we used four statistical measures, more specifically, kurtosis, skewness[6], mean and standard deviation. Therefore, we extracted 16 features using this procedure.

Afterwards, 2 more attributes, such as equivalent circular diameter and fractal dimension, were extracted regarding border quantification. The first, is a known method for measuring the circularity of a lesion, and was calculated as described in Equation 4.4.3 [26]. As for the second, its purpose was to analyse the scale of the edge structure, a procedure quantified by Hausdorff's dimension using a box counting method, as defined in Equation 4.4.4 [33].

$$ECD = \frac{\sqrt{(4 \times AreaOfLesion)^2}}{\pi} \tag{4.4.3}$$

$$Fractaldimension = \frac{log(N(R))}{logR} \tag{4.4.4}$$

The fractal dimension method analyses how many elements of size $M \times M$ do we need to describe the border of a lesion. In 4.4.4 R represents the number of boxes of size $M \times M$ that fit into the image, while $N(R)$ represents the number of boxes that contain a portion of the edge.

4.4.4 Shape features

Shape features were the last type of descriptors extracted in our algorithm, being used, mainly, to describe lesion's outline, as its irregularity can help distinguishing between malignant melanomas and benign lesions. Therefore, we used some standard geometrical features such as filled area, convex area and solidity, complementing them with other important features mentioned in the literature, like compactness and asymmetry.

The first 3 geometric properties extracted were filled area, convex area and solidity, which will be described below. These aforementioned features, according to [12] study, proved its importance, by detecting abnormal deviations when examining melanoma images.

- Convex Area - Scalar that specifies the number of pixels of the convex hull[7].
- Filled Area - Scalar specifying the number of lesion pixels in the binary image

[6]Measure of the asymmetry of a probability distribution based on its mean
[7]Smallest convex polygon that contains the binary image.

with all holes filled in.
▶ Solidity - Ratio specifying the proportion of the pixels in the convex hull that are also in the region.

The next shape attribute to be extracted was compactness, a measure that enabled us to compare the binary object to a circle, which is the most compact shape. In equation 4.4.5 we demonstrate how does this characteristic was computed.

$$Compactness = \frac{(4 \times \pi \times Perimeter^2)}{Area} \tag{4.4.5}$$

Finally, the last shape feature and the one that completed our global feature vector was asymmetry. According to the ABCD rule of dermoscopy (Subsection 1.1.4), asymmetry is the highest weighted criteria for differentiating malignant tumours from benign lesions.

A large number of studies have been carried out on quantifying it, with us deciding to evaluate it by comparing the two halves of the lesion according to the principal axis [23]. Thus, in order to evaluate the lesion asymmetry, we started by calculating the major axis orientation (θ) of the object, and used it to rotate the object θ degrees clockwise to align the major and minor axes with the image axes. Afterwards, the lesion was hypothetically folded along the x-axes and the difference between the two halves of the lesion was calculated by applying a XOR operation on the overlapping folds. The resulting difference was then transformed into a percentage value, and used as the final asymmetry feature. In Figure 4.15 we present a model image with its principal axes aligned.

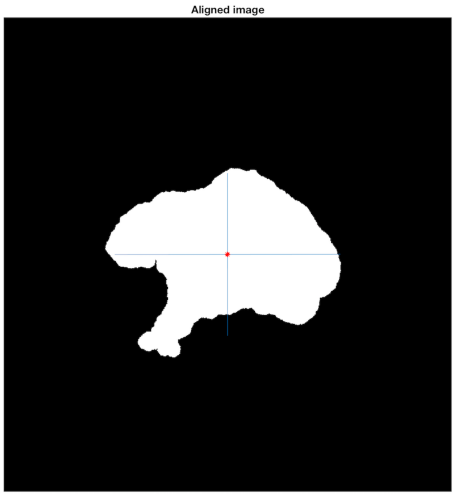

Figure 4.15: Model image with its principal axes aligned.

4.4.5 Feature Selection

Feature selection was the last step before classification, and a very important one in our machine-learning algorithm. Its goal was to find an optimized sub-set of features which could provide the highest discriminating power when employed by the classifier. In another words, its purpose was to reduce the dimensionality of our feature space by eliminating redundant, irrelevant or noisy features, while also reducing the computation cost.

Before performing feature selection we still had a issue to address, and that was feature normalization. Since our extracted features exhibited very different ranges of values, a normalization procedure was required to ensure their scale-, rotation-, and translation- invariance. This way we secured the proper-work of classifiers that were based on the analysis of distances between points in the feature space, consequently, the risk of a situation in which a feature with larger range of values dominated other features was eliminated. Feature normalization was applied as follows:

$$Z_{ij} = \frac{x_{ij} - \mu_j}{\gamma_j} \tag{4.4.6}$$

where x_{ij} was the value of the jth feature of the ith sample, and u_j and γ_j were the mean and standard deviation of the jth feature, respectively. After obtaining the normalized features, we needed to perform dimensionality reduction, and for that purpose we performed two types of feature selection, namely, PCA and RelieF. Their comparison is presented in Chapter 5.

The first technique used to carry out feature selection was PCA, a widely known technique for dimensionality reduction, which detects the variance structure in the data and identifies the directions along which the data subspace exhibits high variance. The idea was to get the dimensions having most of the variation, therefore, we only selected the dominant eigenvectors, retaining 95% variance of the data. In the end, we used the resulting eigenvectors coefficients to compute the features leverage scores using the norms of each vector in the new space.

The second technique used to perform feature selection was ReliefF, a very fast filter model that relies on general characteristics of the data to select a subset of features without involving any learning algorithm. The idea was to estimate the quality of attributes according to how well their values distinguish between samples that are near to each other. By other words, for each selected sample, the values of its

features were compared to those of the nearest neighbours and the relevance scores for each feature were updated accordingly [29]. In order to find the optimal number of neighbours for the ReliefF algorithm, we decided to test the behaviour of several attributes weights according to the variance of the k nearest neighbours. We came up with an optimal K value of 25, which means that around this value the weight of the attributes starts to stabilize, which means that adding more neighbours does not give a better modelling of the data.

4.5 Classification

Classification defined the last phase of our algorithm, having the complex task of discriminating between benign or malignant lesions, using the previously mentioned features described in Section 4.4. Having said that, our main desire was to provide classifiers that allowed a good detection of melanoma lesions, since general practitioners do not often observe them, specially the less representative lesions. In our study, we decided to provide a comparison between three types of classifiers, namely, Support Vector Machines, Random Forests and Adaptive Boosting. A brief description of each classification method is given below.

4.5.1 Support Vector Machines

SVM was the first classifier we trained, a widely known method that has become popular recently, due to its solid theoretical foundation and excellent practical performance. SVM training involves the optimization of a convex cost function, being based on structural risk minimization, where the aim is to find a classifier that minimizes the upper boundary of the expected error. Additionally, they are less prone to over-fitting, when compared to other learning algorithms, which implement the empirical risk minimization principle, and might lead the model to become too strongly tailored to the particularities of the training set as it minimizes its average loss function [34].

In another words, SVM performs classification by finding the optimal hyperplane which maximizes the margin of separation between two distinct classes. That can be done by minimizing the norm of the normal vector of the hyperplane with the constraint that no points should lie in the margin. The decision function of SVM can be described as follows:

$$y(x) = \sum_{i}^{N} \alpha_i K(T_i,x) + b \tag{4.5.1}$$

Where T is a set of N trained samples and α the coefficients learned during the training procedure. As for the K property it represents the kernel function to be used while x is a vector containing the new samples to be evaluated [18].

Regarding the aforementioned kernel function parameter, represented in equation 4.5.1, we opted to use a radial basis function (RBF) as our kernel, in order to determine the decision boundary of the SVM. The RBF kernel is governed by two parameters, known as C (penalization parameter) and γ (kernel width), therefore model selection is required to identify the optimal values for them, in such a way that they give the maximum prediction accuracy on new unseen data. For this purpose, a grid-search was applied to better adjust these constants, using exponentially growing sequences of values for each parameter. To evaluate the goodness of a particular combination of parameter values a 10-fold stratified cross validation assessment method has been used. Afterwards, the SVM classifier was trained with the elected optimal parameters [13]. Finally, it is also worth mentioning that we used the publicly available LibSVM implementation to perform our experiments.

4.5.2 Random Forests

Random forest is a collection or ensemble of decision trees, and was the second classifier we created. In the classification stage, multiple decision trees are trained, with a random subset of training data being generated to train each new tree, but with the same distribution as the previous ones. To put it another way, each random bootstrap sample is used for training one decision tree and at each node of the decision tree, the best split among the randomly selected subset of descriptors is chosen. Each tree is grown to its maximum length without any pruning, and as the number of trees in the forest starts to become large, the generalization error will start to converge, depending on the strength of the individual trees and the correlation between them. This approach of randomly creating vectors of features and building smaller trees, helps preventing overfitting, an issue which often affects other methods, like the decision trees, which is built using the whole set of training features [35].

In the end, the output of the RF is based on the majority vote approach, as each tree votes for a particular class and the class which gets maximum number of votes

is the predicted class. In another words, each individual tree classifies their given feature vector, and the final class assignment is labelled by the number of votes from all the trees. Regarding the number of trees we wanted to build before taking the maximum voting or average prediction, we knew that a higher number of trees would make our predictions stronger and more stable but would likely make our code slower, so we did some experiments to observe our processor's behaviour, and came up with a final number of 500 trees.

4.5.3 Adaptive Boosting

Finally, the last classifier we created in our study was Adaptive Boosting, an ensemble learning algorithm where multiple learners are employed to build a stronger learning algorithm. The idea behind it, is that, instead of attempting to determine a single complex prediction rule, training data is used to generate a large collection of very simple crude rules-of-thumb[8], ending up with a cascade of weak classifiers that combined together produce a powerful classifier, more resistant to overfitting than many other machine learning algorithms.

Initially, the AdaB classifier chooses a base learning algorithm and assigns equal weights to all the training examples. Afterwards, the classifier is called iteratively, and at each step of iteration, the base algorithm is applied to the training set and the weights of the incorrectly classified examples are increased, while those that are correctly classified get their weights decreased. The process continues until a previously-set number of iterations has been performed, in our case we decided to use 500 training iterations. Once completed, the final pool of weak learners is combined, and the output for the ensemble model is taken as the sum of the weighted predictions, like described as follows [14].

$$A = \sum_{i=1}^{j} \alpha_i A_i \tag{4.5.2}$$

Where A_i is the output of the jth weak classifier, and α_i the weight vector of the aforementioned weak classifier. The weight vector is updated based on the difference between the training set pattern accuracy of the ith classifier and $(i-1)$th classifier. On a final note, it's also worth mentioning, that we used decision stumps as our base learning algorithm. These weak classifiers represent short decision trees, that only contain one decision for classification.

[8]Principle that is not intended to be strictly accurate or reliable for every situation.

5

Results

5.1 Dataset Splitting

In this study, like we mentioned previously in Chapter 4, we had a database of 100 images, consisting of benign and melanoma lesions. Images of benign lesions represented 87% of the overall database, while the remaining 13% represented malignant lesions. Therefore, the main idea behind the use of machine learning classifiers was to allow separation of all these images into two independent sets. To make this classification decision, researchers normally divided their classes into training, validation and testing sub-sets. However, we decided to only train and test our independent sets, leaving the validation procedure out. The reason behind this, was related to the fact that our melanoma class was very small, therefore we wanted to use the biggest number of samples possible for the training and testing steps, as they are the most important ones. Additionally, it is worth mentioning that the validation set would only be useful for the SVM classifier, since there is no need to use it for the RF and AdaB, as they automatically generalize the data and do not need any additional validation set. For each class set, we end up using 55% of data for training, and 45% for testing, which meant that for the benign class, we used 48 lesions for training, and 39 for testing, while for the melanoma class, 7 lesions were used for training and 6 for testing.

5.2 Statistical Analysis

In order to evaluate the effectiveness of the proposed classifiers, 3 different metrics have been selected to assess the performance of these classifiers, more precisely, sensitivity, specificity and accuracy measures. The objective was to analyse the confusion matrix of each classifier, in order to understand the discrepancies between

the classification results and our ground truth. The definitions of the used metrics are given below:

$$Sensitivity = \frac{TP}{TP + FN} \tag{5.2.1}$$

$$Specificity = \frac{TN}{FP + TN} \tag{5.2.2}$$

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{5.2.3}$$

Where TP (true-positive) represents the correctly diagnosed positive instances, while TN (true-negative) represents the correctly diagnosed negative instances. As for FP (false-positive), it represents the misclassified positive instances, while FN (false-negative) represents the misclassified negative instances.

Also, it did not escape our notice that we were dealing with a class imbalance problem, therefore we knew that both the sensitivity and specificity metrics should focus in evaluating the results of the most important class in terms of medical diagnosis, the malignant one. Accordingly, we used the sensitivity metric for calculating the rate of malignant lesions correctly identified as having the condition, and the specificity metric for calculating the rate of non-malignant lesions which were correctly identified as not having the condition. Otherwise, from a medical point of view, we could be guaranteeing misleading results, knowing that the benign lesions represented the majority class. Having said that, we knew that the accuracy metric, which measured the rate of correctly classified lesions, was not the most appropriate way of quantifying performance, as it could be strongly biased to favour the majority class, however we still decided to use it as an overall performance indicator.

5.3 Subset of Features

Once we performed dimensionality reduction (Subsection 4.4.5), either with PCA or RelieF, we had all the extracted features ranked individually according to their weights. Afterwards, we needed to decide how many of these highest ranked features would be fed into our classifiers. Initially, we knew that the smallest class we had was the melanoma one, with 7 images for training, therefore we needed to use less than 8 types of features, otherwise, the number of feature weight coefficients to find would be bigger than the number of equations, leading up to an "endless" problem.

We still had the tough decision of selecting the most appropriate number of attributes to use, a matter where there are several viewpoints, some saying that we could use one feature for each 1 or 2 observations in the smallest class, others saying we should use one per each 3 or more observations, a divided opinion nonetheless. For that reason, we performed some experiments, exploring a range from 3 to 7 attributes, and realized that the classifiers did not show significant differences, meaning that the changes were approximately between 0.1 to 1%. Taking that into consideration, we inferred that 5 features could be sufficient to yield optimal results in the classical least squares regression, thus we decided to use this number of features for the remaining classification procedures. There was some apprehension when we chose this number, knowing that some of the selected features could have high correlations between them and a smaller subset should be better, however the reasoning behind this number selection was based on the fact that sometimes one of the selected features could present noise, and if some of the other remaining features exhibited an higher correlation to the aforementioned feature, it might have a positive effect, by making the noise cumulative and consequently diminishing it.

5.4 PCA vs RelieF

The first evaluation we did, was the test performance comparison between the dimensionality reduction techniques, with the intention of understanding which selected the better attributes. In order to do this, we established that we would compare the sensitivity metric between the three proposed classifiers, as presented in Figure 5.1. As mentioned previously, from a diagnostic point of view, the melanoma detection is the crucial assessment we want to make, hence, specificity and accuracy metrics were left apart in this analysis. For the aforementioned comparison, we ran each classifier 500 times, and afterwards we analysed the resulting sensitivity vector using four quantification measures, namely, mean, median, standard deviation and 95% confidence interval. The mean represented the average of all the sensitivity values, while the median was the middle number in the organized sequence of sensitivity values. The standard deviation measured how spread out the sensitivity values were, while the confidence interval meant that if we ran the same classifier another 500 times, we would expect the mean parameter to fall within this interval 95% of the time. Since the last two measures are less known, we depict their definitions below:

$$StandardDeviation = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}} \tag{5.4.1}$$

$$ConfidenceInterval = \overline{x} \pm z\frac{s}{\sqrt{N}} \tag{5.4.2}$$

Where $\{x_1,x_2,...,x_N\}$ are the observed values of the sample items, $\overline{x}$ is the mean value of these observations, and N is the number of observations in the sample. The s value represents the standard deviation value while z is the Z-value for the 95% Confidence Interval, 1.96 more precisely.

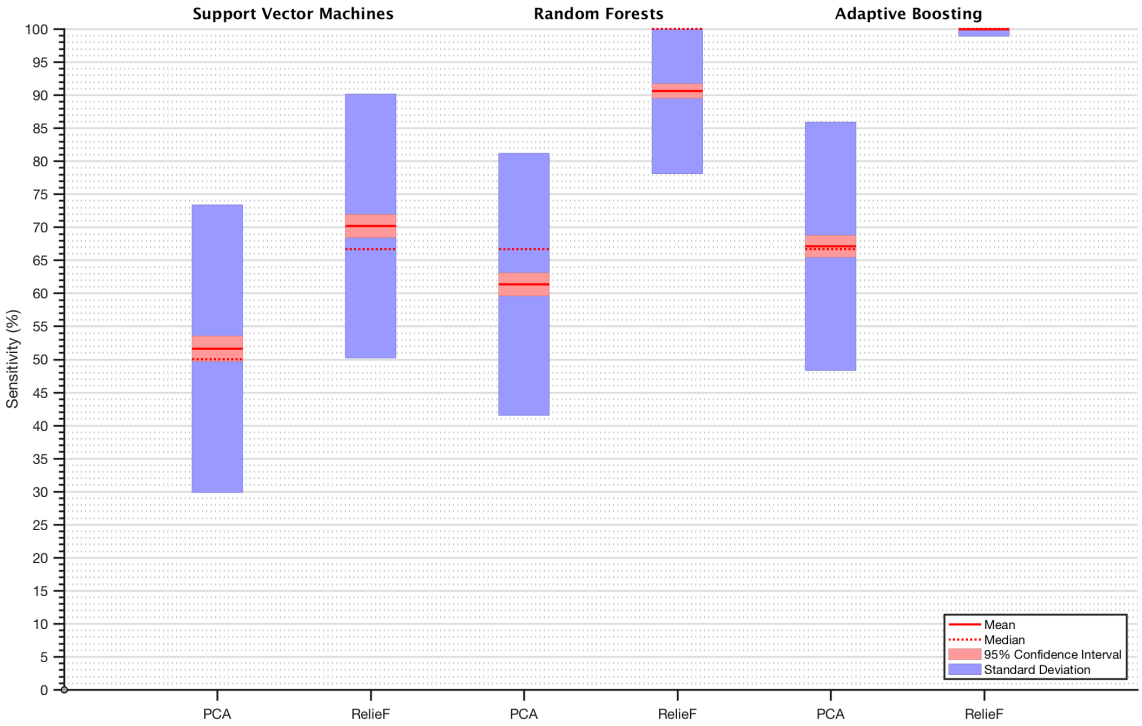

Figure 5.1: Boundary series of a model image.

Looking at each classifier, we noticed that there were notable changes regarding the mean values, which ranged from 18.5% to 32.8%. A fact that can be sustained by the size of the confidence intervals, since for the majority of the classifiers it lies in the 4% range, meaning that the mean value increases were not a fluke. Additionally, despite the standard deviation values being quite spread overall, it was clear that the median values could also corroborate the previous statements, has they also suffered massive increases.

In the end, it is safe to say that there were significant differences between the use of these dimensionality reduction techniques, with RelieF clearly prevailing over PCA. Thus, further on, only the RelieF features were used for classifying comparisons, more precisely, the compactness of the lesion, the border's spatial mean, the hue's centroid distance, and the two histogram distances from the L*a*b and HSV colour

spaces.

5.5 Box-plot Analysis

Once we understood which was the best dimensionality reduction procedure, the following step was to analyse the overall test performance of each classifier, which included the three metrics mentioned in 5.2. In that sense, knowing that each classifier was ran 500 times, we decided to graphically summarize the spreading of the data, by breaking it into quartiles, namely, the lower and upper quartiles, which represent the data points at the 25th and 75th percentiles, respectively. This means that 25% of the data is smaller-than/equal-to the first quartile (Q1) value, while other 25% is bigger-than/equal-to the third quartile (Q3). The difference between these two is called interquartile range (IQR), and was used to describe the range of the middle half of the scores in the distribution. Additionally, we also used two extra percentiles to better understand the biggest outliers in the data, more specifically the 10th (LP) and 90th (HP) percentiles. In the following Figures we present the mentioned graphical analysis, for the SVM, RF, and AdaB classifiers.

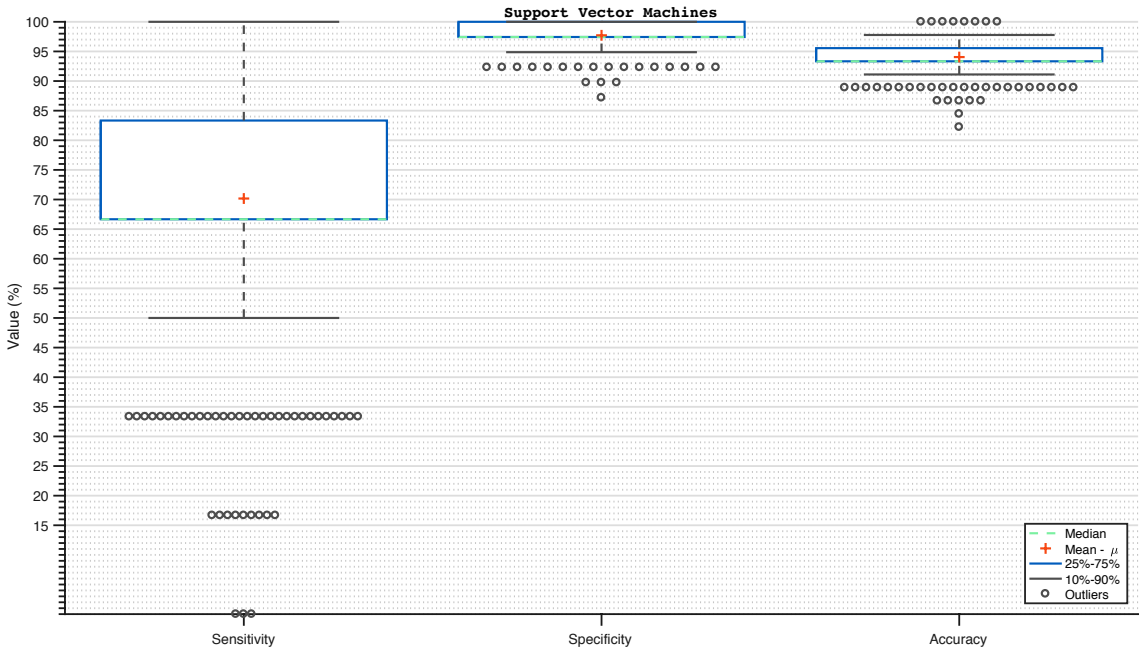

Figure 5.2: Box-plot graphics for the SVM metric analysis.

Sensitivity: LP=50.0%; Q1=66.7%; IQR=16.6%; Q3=83.3%; HP=100.0%;
Specificity: LP=94.9%; Q1=97.4%; IQR=2.6%; Q3=100.0%; HP=100.0%;
Accuracy: LP=91.1%; Q1=93.3%; IQR=2.3%; Q3=95.6%; HP=97.8%;

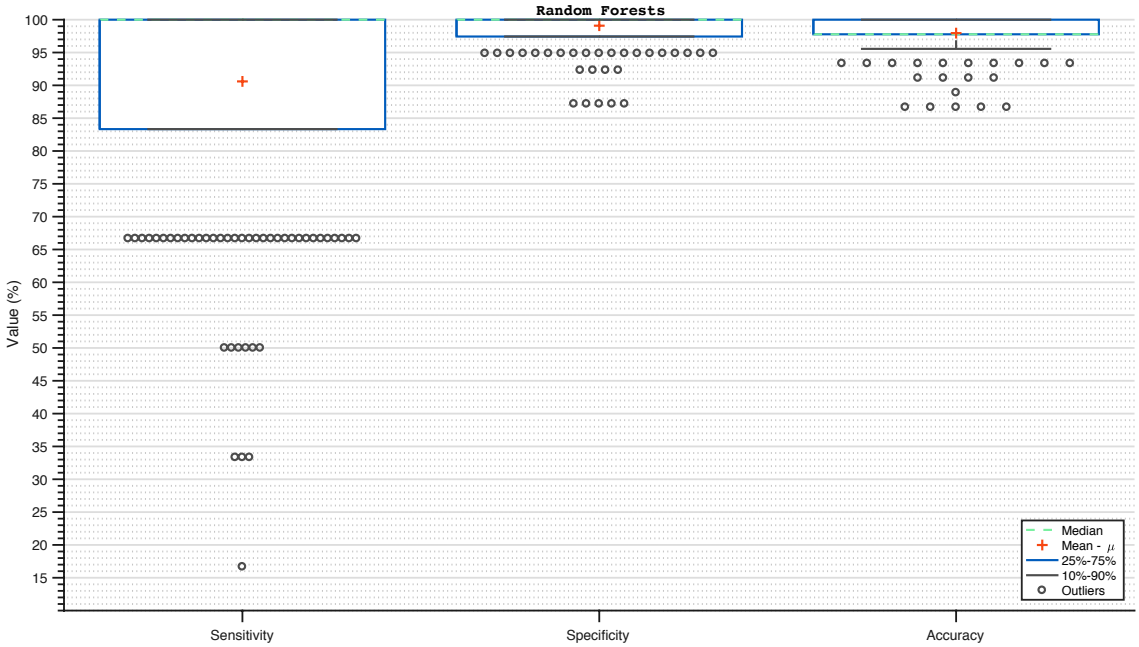

Figure 5.3: Box-plot graphics for the RF metric analysis.

Sensitivity: LP=83.3%; Q1=83.3%; IQR=16.6%; Q3=100.0%; HP=100.0%;
Specificity: LP=97.4%; Q1=97.4%; IQR=2.6%; Q3=100.0%; HP=100.0%;
Accuracy: LP=95.6%; Q1=97.8%; IQR=2.2%; Q3=100.0%; HP=100.0%;

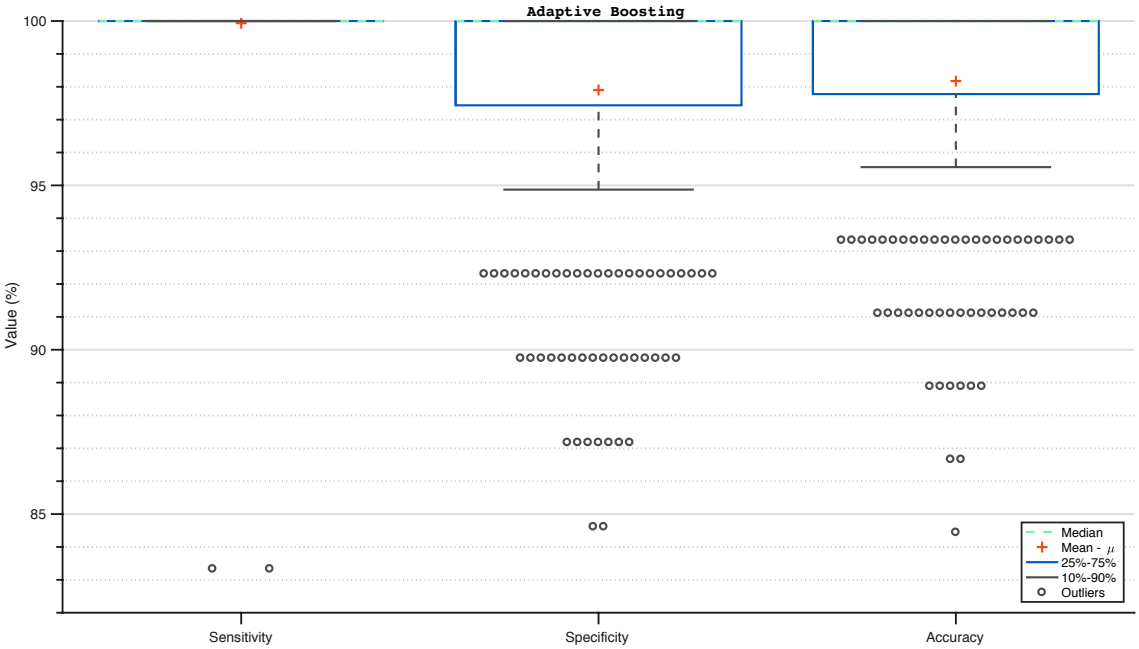

Figure 5.4: Box-plot graphics for the AdaB metric analysis.

Sensitivity: LP=100.0%; Q1=100.0%; IQR=0.0%; Q3=100.0%; HP=100.0%;
Specificity: LP=94.9%; Q1=97.4%; IQR=2.6%; Q3=100.0%; HP=100.0%;

Accuracy: LP=95.6%; Q1=97.8%; IQR=2.2%; Q3=100.0%; HP=100.0%;

Finally, after visualizing all these summary statistics, we got an initial preview of the classifier behaviours. Regarding the sensitivity metric, there were significant differences in the spreading of the data, as the SVM and RF classifiers, showed 66.7% and 83,3% Q1 values, while the AdaB classifier presented an amazing Q1 value of 100%. Additionally, unlike the closely packed values presented by the last classifier, both the first two classifiers, presented medium-sized IQR values, as their Q3 values were 83.3% and 100.0%, respectively. Even further, both the first two classifiers showed big outlier margins, a negative factor when compared to the nonexistent margin presented by the AdaB classifier.

Concerning the other two metrics, specificity and accuracy, we noticed very good results, with all the analysed summary statistics showing similar results, quantity and quality worth-wise. The first metric presented a noteworthy 97,4% Q1 value for all the classifiers, complemented by a very small IQR 2.6% value, which meant very low variability in the observations. As for the second metric, once again the classifiers presented similar low variability in the IQR values [2.2;2.3], while the Q1 values ranged from 93.3% to 97.8%, with SVM exhibiting the worst performance. Both these metrics showed some unwanted outliers, however only by a small margin.

In the end, the use of these percentiles was a useful measure of spread and central tendency, indicating that the means, despite some exceptional cases of skewed data, were not far way from the median values, therefore they could be considered a fairly robust parameter. Overall, the Adab classifier seemed to show the best values among the classifiers, specially in terms of variability, where it was way lower than the others. As for the SVM classifier, its values seemed the worst of all three, meaning that it is the task with the most potential for improvement.

5.6 Classifier Comparisions

After we got a glimpse on the general distribution of the data, it let us with the impression that the AdaB classifier outperformed the others, while the SVM classifier seemed to show the worst box-plot percentile values, specially in terms of the sensitivity metric. Nonetheless, we still wanted to corroborate this idea, so we choose to compare each statistical metric individually for the previously showed classifier performances. In that regard, and knowing that we had established that the mean could be considered a good analysis parameter, in the next Figures the comparisons

were performed using the aforementioned mean related measures we had already used in Section 5.2, with the additional representation of all the observation values.

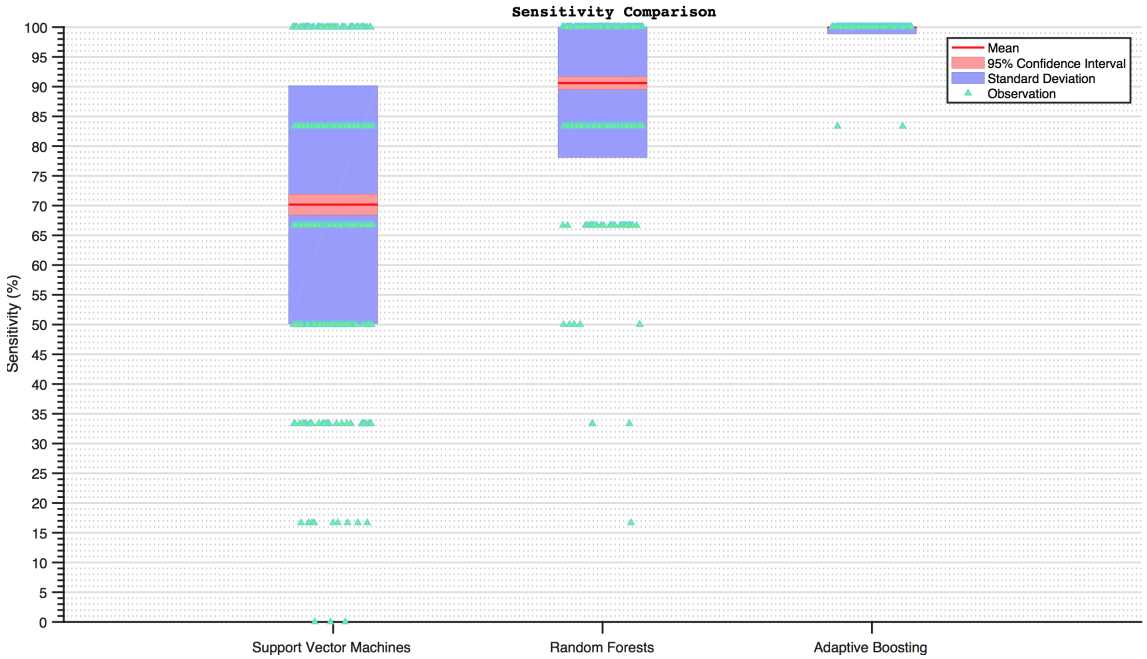

Figure 5.5: Sensitivity comparison.

SVM: Mean=70.2%; SD=[50.2;90.1]%; CI=[68.4;71.9]%;
RF: Mean=90.6%; SD=[78.1;100.0]%; CI=[89.5;91.7]%;
AdaB: Mean=99.9%; SD=[98.9;100.0]%; CI=[99.8;100.0]%;

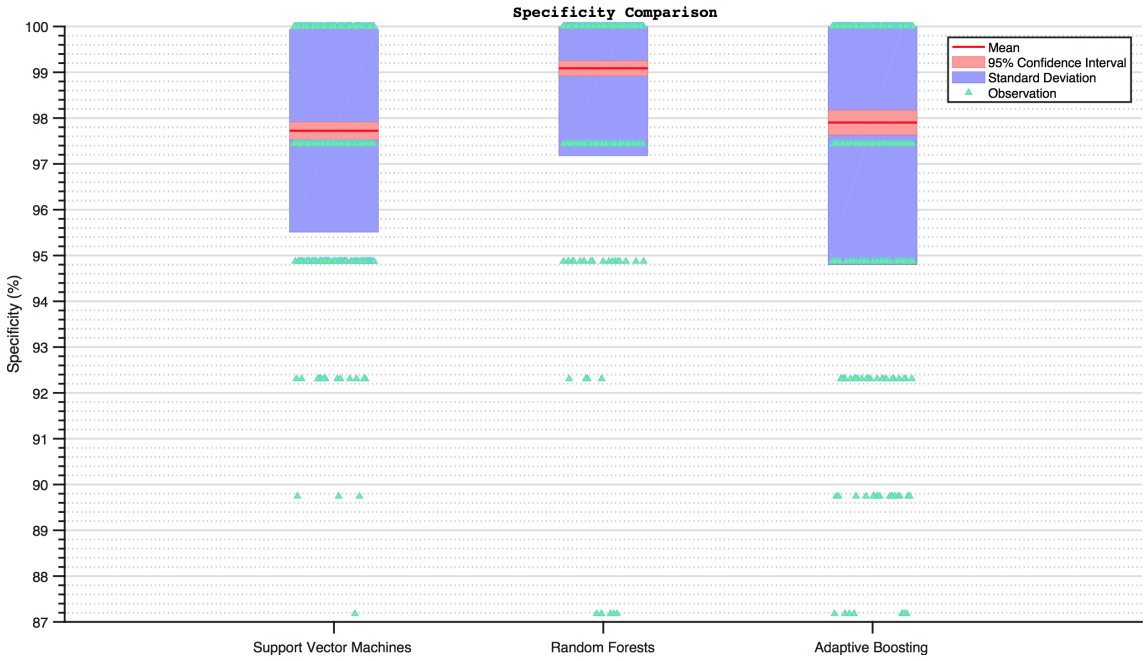

Figure 5.6: Specificity comparison.

SVM: Mean=97.7%; SD=[95.5;99.9]%; CI=[97.5;97.9]%;
RF: Mean=99.1%; SD=[97.2;100.0]%; CI=[98.9;99.3]%;
AdaB: Mean=97.9%; SD=[94.8;100.0]%; CI=[97.6;98.2]%;

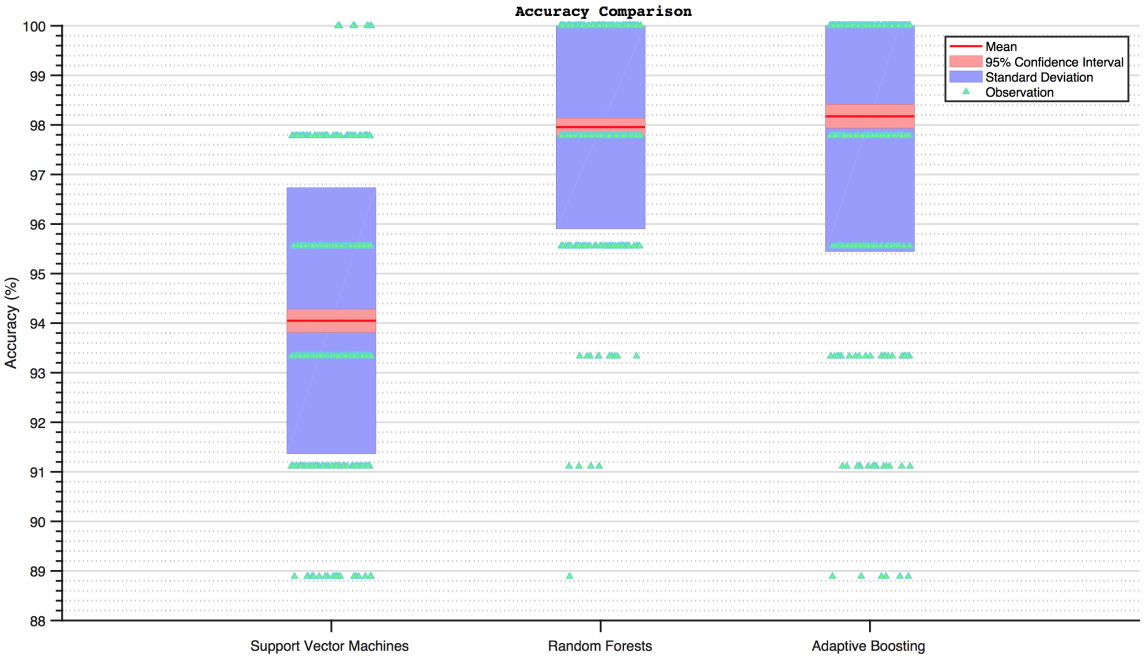

Figure 5.7: Accuracy comparison.

SVM: Mean=94.0%; SD=[91.4;96.7]%; CI=[93.8;94.3]%;
RF: Mean=98.0%; SD=[95.9;100.0]%; CI=[97.8;98.1]%;
AdaB: Mean=98.2%; SD=[95.4;100.0]%; CI=[97.9;98.4]%;

Among all the proposed framework of metrics the Confidence Intervals proved to be very small, meaning that the mean values in all these comparisons were not a coincidence. In the sensitivity metric, the AdaB classifier clearly achieved the highest mean among the classifiers (99.9%), followed by the RF classifier with a 9.3% difference, and lastly by the SVM classifier with a larger 29.7% difference. In the other hand, for the specificity metric, the RF's mean (99.1%) outperformed the one from the others, however the differences were not significant this time, as they ranged between 1.2% to 1.4%. Lastly, in the accuracy metric, the AdaB mean (98.2%) surpassed the RF and SVM means, but once again the differences were minimal, 0.2% and 4.2%, respectively. Regarding the standard deviation values, both the specificity and accuracy metrics presented low standard deviations, indicating that the data points were close to the mean, while the sensitivity metric showed rather higher deviations, mostly because this metric refers to a small sized class, meaning that the misclassification values are widely spread out.

In the end, the mean values provided clear results for the three metrics, and as we previously suspected, the AdaB classifier definitely achieved the best results, specially in the sensitivity metric, as it turned out to be the most effective classifier in recognizing malignant melanoma, the most dangerous of the lesions. The minor difference of 1.2% to the RF classifier in the specificity metric is not relevant in this case, as its value was also noteworthy (97.9%). AdaB can learn non-linear decision boundaries, hence, the reasoning behind this amazing performance could be related to the non-linearity of our features, otherwise it was all just a question of empirical results. That being said, both the RF and SVM classifiers could also have been affected by the class imbalance problem, since the training matrices might have been misled to favour the major benign class.

6

Conclusions

Due to the difficulty and subjectivity of human interpretation, the computerized image analysis techniques have become important tools in this research area. In this study, a computer-aided diagnosis system for the classification of dermoscopy images was presented. This methodological approach is fully automatic, and allows us to test and evaluate lesion discrimination between benign and malignant melanoma images.

The proposed framework covers the main diagnosis components, known as pre-processing, segmentation, feature extraction, feature selection, and classification. Both the pre-processing and segmentation stages were successfully applied, and except some minor discrepancies, they were able to provide good enhancement (either by contrast/illumination improvement or efficient artefact removal) and partition of the images, respectively. With regard to the feature extraction stage, a global feature vector, consisting of colour, texture, border, and shape features, was extracted. Furthermore, two dimensionality reduction techniques were applied for feature selection, with RelieF surpassing PCA as the most efficient one, since it provided significant improvements in the classification performances. Lastly, three classifiers were employed in the classification stage, namely, SVM, RF and AdaB. Among those, the AdaB classifier particularly stood out in relation to the others, revealing exceptional results in several statistical metrics.

More specifically, the AdaB classifier was ran 500 times, and in each of these iterations, a random test bed of 44 lesions (38 benign and 6 malignant) was used for testing. In the end, the outcoming values were averaged, and we achieved the outstanding sensitivity of 99.9%, a specificity of 97.9%, and an accuracy of 98.2%. Concerning the importance of melanoma detection, the sensitivity metric provided exceptional results, misclassifying melanoma only 0.1% of the times, while the specificity metric also yielded amazing results, misleading to lesion excision only 2.1% of the times. Additionally, the classifier showed high resilience to outliers, with all the

evaluated metrics showing most of the observations closely packed to the mean.

Although the results cited above were obtained through experiments conducted on a particular image set, the obtained performance of our system is in fact highly comparable with the literature's melanoma recognition systems reported in our state-of-the-art. However, we acknowledge that due to lack of a standard benchmark for dermoscopy imaging, it is not easily feasible to provide a comprehensive and quantitative comparative study among the existing classification methods.

Additionally, it is also worthwhile to highlight that this system is not designed to bring about complete autonomy in the diagnostic process or replace human judgement, but rather has potential as a complementary system that could be used to screen images and direct doctors attention to cases that have high risk. Moreover, it could be used to corroborate the diagnosis of even trained doctors, in order to provide better interpretation and definition of whether or not a lesion is likely to be melanocytic or malignant for instance. This way, unnecessary biopsies could be avoided and any suspicious lesion would be directed for excision or histopathologic confirmation.

It is worth mentioning that there are still several lines of research arising from this work which should be pursued. For instance, future work must extend our results by using a larger database of images, in order to have a fair and general representation of the data, and avoid class imbalance problems like we had. Furthermore, in this study, some images containing large (did not fit entirely within the image frame) or incomplete lesion objects had to be excluded, and despite knowing that this limitation is in line with the problems imposed by the majority of the literature, it might be an issue future work could fix. Moreover, further efforts could also be done to optimize our border segmentation procedure, by obtaining a comparable ground truth from an expert dermatologist.

Finally, the most interesting opportunity for extending the scope of this thesis still remains, more precisely, its application on the clinical level as an auxiliary evaluation tool for skin lesions. To put it another way, our work has demonstrated the potential for efficiently recognizing melanoma, hence, future work could focus on implementing our proposed system in clinical trials with several subjects, over a long period of time, to overcome the possible glitches and further optimize its performance. That being said, during the development of this work we elaborated a graphical-user-interface for research purposes (for more details, refer to Appendix A), which could serve as the initial framework to a future conception in terms of a real clinical useful application.

Bibliography

[1] K. Korotkov and R. Garcia, "Computerized analysis of pigmented skin lesions: A review," *Artificial Intelligence in Medicine*, vol. 56, no. 2, pp. 69–90, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.artmed.2012.08.002

[2] M. Filho, Z. Ma, and J. M. R. S. Tavares, "A Review of the Quantification and Classification of Pigmented Skin Lesions: From Dedicated to Hand-Held Devices," *Journal of Medical Systems*, vol. 39, no. 11, 2015.

[3] A. Pflugfelder, C. Kochs, A. Blum, M. Capellaro, C. Czeschik, T. Dettenborn, D. Dill, E. Dippel, T. Eigentler, P. Feyer, M. Follmann, B. Frerich, M. K. Ganten, J. G??rtner, R. Gutzmer, J. Hassel, A. Hauschild, P. Hohenberger, J. H??bner, M. Kaatz, U. R. Kleeberg, O. K??lbl, R. D. Kortmann, A. Krause-Bergmann, P. Kurschat, U. Leiter, H. Link, C. Loquai, C. L??ser, A. Mackensen, F. Meier, P. Mohr, M. M??hrle, D. Nashan, S. Reske, C. Rose, C. Sander, I. Satzger, M. Schiller, H. P. Schlemmer, G. Strittmatter, C. Sunderk??tter, L. Swoboda, U. Trefzer, R. Voltz, D. Vordermark, M. Weichenthal, A. Werner, S. Wesselmann, A. J. Weyergraf, W. Wick, C. Garbe, and D. Schadendorf, "Malignes melanom S3-leitlinie "Diagnostik, therapie und nachsorge des melanoms"," *JDDG - Journal of the German Society of Dermatology*, vol. 11, no. SUPPL. 6, pp. 1–126, 2013.

[4] G. Argenziano, I. Zalaudek, G. Ferrara, R. Johr, D. Langford, S. Puig, H. P. Soyer, and J. Malvehy, "Dermoscopy features of melanoma incognito: Indications for biopsy," *Journal of the American Academy of Dermatology*, vol. 56, no. 3, pp. 508–513, 2007.

[5] C. Garbe, A. Hauschild, M. Volkenandt, D. Schadendorf, W. Stolz, U. Reinhold, R.-D. Kortmann, C. Kettelhack, B. Frerich, U. Keilholz, R. Dummer, G. Sebastian, W. Tilgen, G. Schuler, A. Mackensen, and R. Kaufmann, "Evidence and interdisciplinary consense-based German guidelines: diagnosis

and surveillance of melanoma." *Melanoma research*, vol. 17, no. 6, pp. 393–9, 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17992123

[6] J. Neila and H. P. Soyer, "Key points in dermoscopy for diagnosis of melanomas, including difficult to diagnose melanomas, on the trunk and extremities," *Journal of Dermatology*, vol. 38, no. 1, pp. 3–9, 2011.

[7] G. Campos-do carmo and M. Ramos-e silva, "Dermoscopy : basic concepts," pp. 712–719, 2008.

[8] R. H. Johr, "Dermoscopy: Alternative melanocytic algorithms - The ABCD rule of dermatoscopy, menzies scoring method, and 7-point checklist," *Clinics in Dermatology*, vol. 20, no. 3, pp. 240–247, 2002.

[9] I. Maglogiannis and C. N. Doukas, "Overview of advanced computer vision systems for skin lesions characterization," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 5, pp. 721–733, 2009.

[10] F.-y. Xie, H. Fan, L. Yang, Z.-g. Jiang, R.-s. Meng, and A. Bovik, "Melanoma Classification on Dermoscopy Images using a Neural Network Ensemble Model," *IEEE Transactions on Medical Imaging*, vol. 0062, no. c, pp. 1–1, 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7762919/

[11] S. Kaya, M. Bayraktar, S. Kockara, M. Mete, T. Halic, H. E. Field, and H. K. Wong, "Abrupt skin lesion border cutoff measurement for malignancy detection in dermoscopy images," *BMC Bioinformatics*, vol. 17, no. S13, p. 367, 2016. [Online]. Available: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1221-4

[12] A. Pennisi, D. D. Bloisi, D. Nardi, A. R. Giampetruzzi, C. Mondino, and A. Facchiano, "Skin lesion image segmentation using Delaunay Triangulation for melanoma detection," *Computerized Medical Imaging and Graphics*, vol. 52, pp. 89–103, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.compmedimag.2016.05.002

[13] J. Jaworek-Korjakowska and P. Kłeczek, "Automatic Classification of Specific Melanocytic Lesions Using Artificial Intelligence," *BioMed Research International*, vol. 2016, 2016.

[14] J. Premaladha and K. S. Ravichandran, "Novel Approaches for Diagnosing Melanoma Skin Lesions Through Supervised and Deep Learning Algorithms," *Journal of Medical Systems*, vol. 40, no. 4, pp. 1–12, 2016.

[15] N. Alfed, F. Khelifi, A. Bouridane, and S. Huseyin, "Pigment network - based skin cancer detection," *37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 7214–7217, 2015.

[16] Jaworek-Korjakowska, J., and R. Tadeusiewicz, "Determination of Border Irregularity in Dermoscopic Color Images of Pigmented Skin Lesions," *Embbs, Ieee*, vol. 3, no. 3, pp. 6459–6462, 2014.

[17] M. Rastgoo, R. Garcia, O. Morel, and F. Marzani, "Automatic differentiation of melanoma from dysplastic nevi," *Computerized Medical Imaging and Graphics*, vol. 43, pp. 44–52, 2015. [Online]. Available: http://dx.doi.org/10.1016/j.compmedimag.2015.02.011

[18] F. Riaz, A. Hassan, M. Y. Javed, and M. T. Coimbra, "Detecting melanoma in dermoscopy images using scale adaptive local binary patterns," *Proc. Annu. Int. Conf. IEEE Engineering in Medicine and Biology Society (EMBS)*, pp. 6758–6761, 2014. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6945179

[19] O. Abuzaghleh, S. Member, and B. D. Barkana, "Noninvasive Real-Time Automated Skin Lesion Analysis System for Melanoma Early Detection and Prevention," vol. 3, no. October 2014, 2015.

[20] R. Amelard, J. Glaister, A. Wong, and D. A. Clausi, "High-Level Intuitive Features (HLIFs) for intuitive skin lesion description," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 3, pp. 820–831, 2015.

[21] C. Barata, J. S. Marques, and J. Rozeira, "A system for the detection of pigment network in dermoscopy images using directional filters," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 10, pp. 2744–2754, 2012.

[22] M. Sadeghi, T. K. Lee, D. McLean, H. Lui, and M. S. Atkins, "Detection and analysis of irregular streaks in dermoscopic images of skin lesions," *IEEE Transactions on Medical Imaging*, vol. 32, no. 5, pp. 849–861, 2013.

[23] R. Garnavi, M. Aldeen, and J. Bailey, "Computer-aided diagnosis of melanoma using border- and wavelet-based texture analysis," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 6, pp. 1239–1252, 2012.

[24] M. Faal, M. H. Miran Baygi, and E. Kabir, "Improving the diagnostic accuracy of dysplastic and melanoma lesions using the decision template combination method," *Skin Research and Technology*, vol. 19, no. 1, pp. 113–122, 2013.

[25] Q. Abbas, M. Emre Celebi, I. F. Garcia, and W. Ahmad, "Melanoma recognition framework based on expert definition of ABCD for dermoscopic images," *Skin Research and Technology*, vol. 19, no. 1, pp. 93–102, 2013.

[26] Q. Abbas, M. Emre Celebi, and I. Fondón, "Computer-aided pattern classification system for dermoscopy images," *Skin Research and Technology*, vol. 18, no. 3, pp. 278–289, 2012.

[27] J. H. Christensen, M. B. T. Soerensen, Z. Linghui, S. Chen, and M. O. Jensen, "Pre-diagnostic digital imaging prediction model to discriminate between malignant melanoma and benign pigmented skin lesion," *Skin Research and Technology*, vol. 16, no. 1, pp. 98–108, 2010.

[28] M. Messadi, A. Bessaid, and A. Taleb-Ahmed, "Extraction of specific parameters for skin tumour classification." *Journal of medical engineering & technology*, vol. 33, no. 4, pp. 288–95, 2009. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2683694{&}tool=pmcentrez{&}rendertype=abstract

[29] M. E. Celebi, "NIH Public Access," *Biophysical Chemistry*, vol. 31, no. 6, pp. 362–373, 2007.

[30] A. Afonso and M. Silveira, "Hair detection in dermoscopic images using Percolation," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pp. 4378–4381, 2012.

[31] W. K. Pratt, *PROCESSING DIGITAL IMAGE PROCESSING*, 2001, vol. 5.

[32] R. T. Whitaker, "A Level-Set Approach to 3D Reconstruction From Range Data 1 Introduction," *International Journal of Computer Vision*, vol. 29, no. 3, pp. 203–231, 1998. [Online]. Available: http://link.springer.com/10.1023/A:1008036829907{%}5Cnhttp://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.133.8076{&}rep=rep1{&}type=pdf

[33] Y. Kohavi and H. Davdovich, "Topological dimensions , Hausdor dimensions & fractals," *Bar-llan University*, no. May, pp. 1–16, 2006.

[34] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data mining and knowledge discovery*, vol. 2, no. 2, pp. 121–167, 1998. [Online]. Available: http://link.springer.com/10.1023/A:1009715923555{%}5Cnpapers3://publication/doi/10.1023/A:1009715923555

[35] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

Appendices

A

Proposed Framework

When we started this study we wanted to provide an efficient approach to the classification of dermoscopic images, in a way to which doctors would trust our algorithm to the extent of really taking it seriously, and that was our major focus. During this willingness for thriving, we were forced to try several different methods along the steps of our work-flow, some which led to good outcomings and others which did not. Throughout this course in search of the most proficient methods, we came to the conclusion that most of the authors, who tried to achieve a similar screening procedure to ours, also wasted a lot of time and effort investigating which methods suited better their purposes. In that sense, and after some careful thought, we decided that it would be helpful to provide a mutual framework to all of the researchers interested in this field of work, a tool that looks at the bigger picture and might lead the community to faster developments in the area. As a result, the idea was to create a graphical user interface (GUI), an operating system which is described on Section A.1. Nevertheless, our main goal remained the same, thoroughly scrutinizing what were the most adequate methods, the difference is that meanwhile we also started developing the previously mentioned framework.

A.1 Graphical User Interface

The GUI is the type of user interface that allows users to "interact" with high performance software. These systems provide point-and-click control of software applications, through the use of graphical icons and visual indicators, instead of text-based user interfaces, which require commands to be typed on a computer keyboard. For this reason, they are much easier to learn, since they eliminate the need to know any programming languages and commands do not need to be memorized in order to type them and run the application. From the developer's point of view this system seemed like a dominant way to provide end-users a useful framework, and that was

the main goal, to offer supplementary tools to new enthusiast's willing to work in this field and continue exploring new ways to improve the early diagnosis of skin cancer. Therefore, we used GUI front ends to design our automatic classification algorithm able to distinguish between benign and malicious lesions, an Interface which can be seen in Figure A.1. As we can see the Interface contains several controls such as push up buttons, panels, pop-up menus, check boxes and even axes, which we will explain in the following paragraphs.

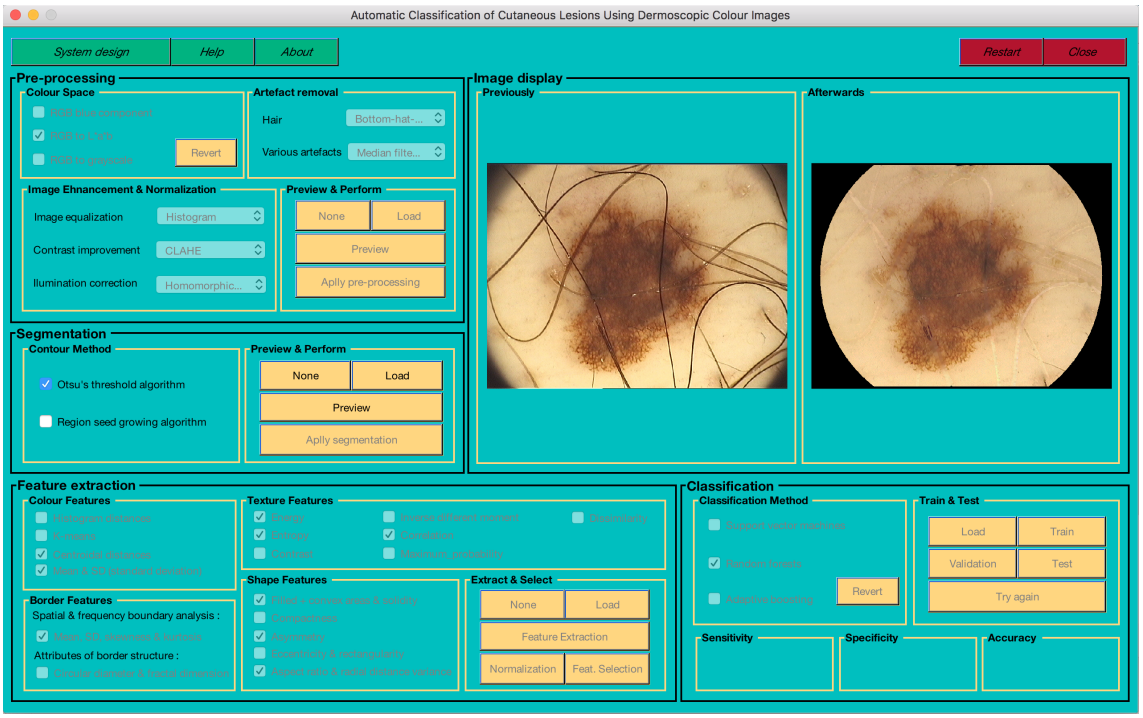

Figure A.1: GUI, showing an image before and after pre-processing.

The Interface was designed to focus on 4 major image computational techniques, known as "Pre-processing", "Segmentation", "Feature extraction" and "Classification", ordered respectively according to each method's placement along the workflow pipeline. Therefore, when we start the GUI only the "Pre-processing" panel is active, with the next panel in line ("Segmentation") only being activated when the user decides to end its research on the current active panel. The same process happens along the other panel transitions like "Segmentation"↪"Feature extraction" and "Feature extraction"↪"Classification", with the GUI disabling most of its functions when the researcher arrives to the "Classification" panel. Last but not least, there is still an extra panel which is always active and has not been referred to yet, named "Image display", which contains two axes crucial for the display of images previously and after they are changed according to the user choices. As

previously referred, each of these techniques has its own representative panel, where the user can choose methods (using the check boxes and the pop-up menus), and perform several tasks (using the push-up buttons), with the last being clarified later. In that sense, it is important to mention that in each one of these panels the default methods and check-marks which are selected represent the options that led to the best results in each technique. In Figure A.2 we present a screen shot of the Interface, in a case where the user has already gotten into to the classification panel, where we can see the test results from a "Random forests" classifier. Additionally, the "Image display" panel shows the last step of transformation performed on a model image (belonging to the analysed dataset), before arriving to the "Feature extraction" panel. In the left plane we can see the outcome of the pre-processing techniques and on the right plane we see the product of its segmentation.

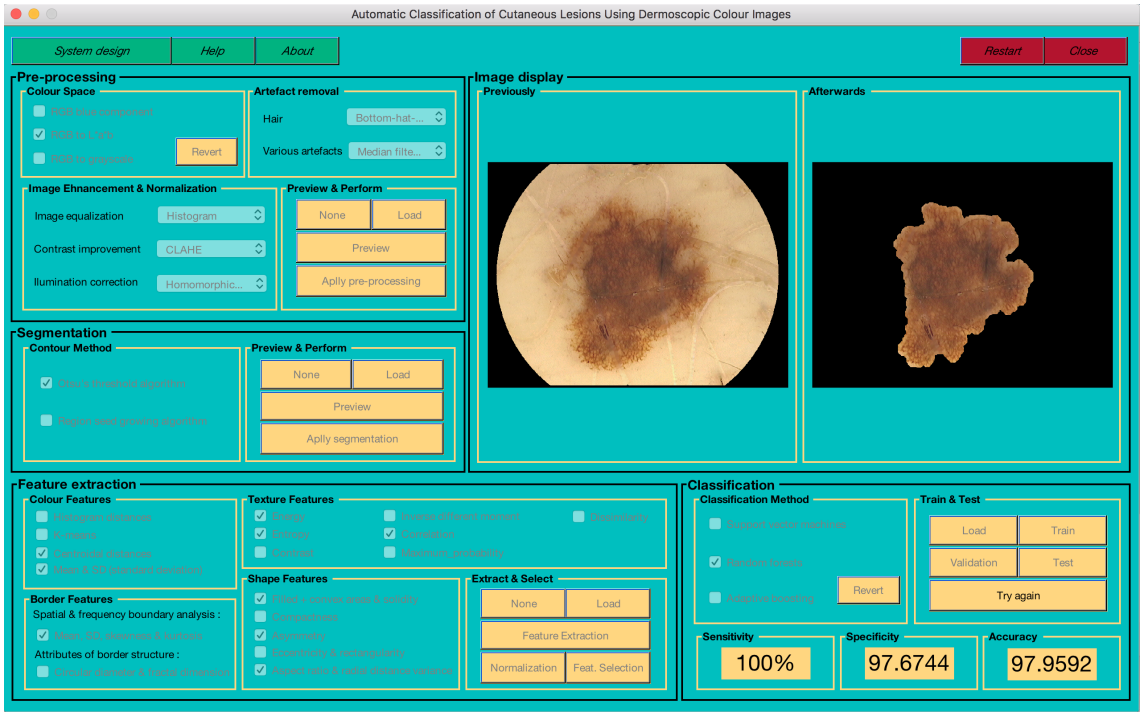

Figure A.2: GUI, showing an image before and after segmentation. Additionally the test results from a Random Forests classifier are also presented.

The GUI has five general push-up buttons, named "System design", "Help", "About", "Restart" and "Close", which are very intuitive as their name suggests. As for the panel buttons, the "Pre-processing" panel has four, named "None", "Load", "Preview" and "Apply pre-processing". The first is used when the user does not want to perform any pre-processing action, and wants to move on to the "Segmentation" panel, the second as the name intends, loads a dataset of previously pre-processed

images (originally the best ones) and makes the program shift to the next panel, the third is used to preview the effects of the pre-processing methods elected by the user, and only afterwards the fourth button becomes operational, bringing the user the option of applying the methods he has chosen to all the images, ending here the "Pre-processing" panel actions. In a similar way, the "Segmentation" panel has 4 similar buttons, named "None", "Load", "Preview" and "Apply segmentation", all with the same functions but this time referring to the segmentation of the images. Next in line, we have the "Feature extraction" panel, composed by five buttons, named "None", "Load", "Feature Extraction", "Normalization" and "Feat. Selection". The first two are similar to the previously referred panels, giving the ability to skip this panel without performing anything, or to load an array of previously extracted and selected features (originally the best ones), respectively. The third button enables the extraction of the user elected features, and makes the fourth and fifth buttons available, meaning that the user can now perform feature selection, with the option of previously normalizing the features he extracted. Finally, we arrive to the last panel of the work-flow, the "Classification" one, which has four buttons, named "Load", "Train", "Test" and "Try again". The first loads a previously trained matrix (originally the best one respective to the elected method), while the second button allows the user to train a new model regarding the type of classifier the user wants to explore. Whether the user chooses the first or second referred buttons, either way the "Test" button will become operational, and the user will be able to test new images with their trained matrix. When the user presses the "Test" button, the "Classification" panel disables all its buttons and method's selection , except for one, the "Try again" button which leaves one final option to the user in this final panel, which is to repeat all these classification steps he took before, which might be important if he wants to train new matrices and latter obtain better test results. In a final note, there is also one extra button in the "Pre-processing" and "Classification" panels, named "Revert", which allows the user to revert its colour space selection, or to revert its classification method selection, respectively.