Faculdade de Engenharia da Universidade do Porto

Automatic analysis of skin lesions in dermoscopy images: feature extraction and classification

Ricardo Pereira Cunha de Sousa Lé

MASTER THESIS

Integrated Masters in Bioengineering – branch of Biomedical Engineering

Supervisor: Prof. Doutor João Manuel R.S. Tavares (FEUP/DEMec)
Co-supervisor: Doutor Zhen Ma (FEUP/INEGI)

22/06/2015

Agradecimentos

Gostava de em primeiro lugar agradecer ao Professor João Tavares pela oportunidade concedida de realizar este trabalho sob a sua orientação, pela sua disponibilidade e o aconselhamento que me forneceu ao longo de todo o periodo da sua realização, sem o qual este trabalho não estaria decerto concluído.

Gostava também de deixar um especial agradecimento ao Zhen Ma, por a ajuda que me prestou ao longo de todo o trabalho, tanto na parte da implementação dos algoritmos testados, como em todas as discussões sobre a análise dos resultados obtidos, que permitiram que o trabalho crescesse em valor e nas conclusões obtidas.

Quero também agradecer todo o apoio recebido pela família, pelos pais e pelos irmãos, que me ajudaram a lidar com todo o stress enfrentado, e conseguiram algumas vezes dar-me uma perspetiva diferente sobre os assuntos tratados que me ajudaram a progredir e a melhorar no trabalho realizado.

E gostava de por último agradecer profundamente à Francisca, sem a qual esta etapa teria sido mais difícil de ultrapassar. Por ter sempre tolerado os momentos de maior nervosismo, e dado sempre o seu incentivo optimista, conseguiu transmitir-me força para que terminasse este projeto da melhor forma.

"We can know only that we know nothing. And that is the highest degree of human wisdom."

Leo Tolstoy, War and Peace

Resumo

Melanoma é o tipo de cancro de pele mais mortal que existe. Foi o sexto cancro mais comum nos Estados Unidos (EU) em 2014, e aproximadamente dez mil pessoas morreram com esta doença. Além disso, a sua incidência tem vindo a aumentar em todo o mundo, o que faz com que este seja um preocupação muito séria para a saúde pública. Nas suas fases mais precoces, é difícil de distingui-lo de outras lesões benignas da pele, extremamente frequentes, mesmo para dermatologistas experientes, e a principal causa da mortalidade do melanoma da pele é o seu diagnóstico tardio, porque este cancro é capaz de desenvolver metástases em outras partes do corpo. Por causa disto, esforços têm sido feitos para melhorar a consciência das pessoas acerca desta doença, e também das principais características que alertam para a sua presença e que devem ser examinadas por um especialista.

Para aliviar este problema, a comunidade científica tem vindo a desenvolver sistemas assistidos por computador que consigam rapidamente identificar e caracterizar uma lesão a partir de imagens de dermatoscopia, tentando automaticamente reconhecer o risco de ser um melanoma nas suas fases mais precoces. Muitos trabalhos têm sido dirigidos a este problema e já foram obtidos resultados encorajadores, no entanto não houve muito foco em determinar a importância das características consideradas para o reconhecimento e o quanto elas são indicadores relevantes da malignidade de uma lesão.

O presente trabalho foi dedicado a esta lacuna, focando-se em três passos de um sistema automático de classificação, nomeadamente a extração de características, a seleção de características e a classificação. Na aproximação inicial, um grupo de características foi extraído das lesões disponíveis em duas bases de dados, tendo estas sido depois submetidas a uma avaliação por métodos de *ranking*, com os quais a sua relevância foi estudada. Foi concluído que as características de cor são as mais consistentes no reconhecimento de melanomas, com enfâse nas medidas estatísticas simples obtidas dos espaços de cor $L^*a^*b^*$ e $L^*u^*v^*$, no componente vermelho do espaço de cor RGB, e na saturação média. Depois deste processo, foi aplicada uma pesquisa empírica no espaço de características considerado, com o objetivo de encontrar as combinações de características que permitem a melhor performance de classificação usando uma máquina de vetor de suporte com uma função de base radial como kernel para avaliação nas duas bases de imagens. O uso de características de forma, cor e textura foi avaliado individualmente e comparado com o uso das três combinadas, e foi concluido que os melhores resultados são atingidos quando as três categorias são consideradas. A melhor sensibilidade obtida com o primeiro conjunto de lesões foi de 89.66% (4 melanomas falhados em 29), e para o segundo de 90% (4 melanomas falhados em 40). As combinações de características que atingiram melhores performances em ambos foram comparadas, e as características que apareceram em ambas foram realçadas.

Abstract

Melanoma is the deadliest type of skin cancer. It was the sixth most common cancer in the United States (US) in 2014, and approximately ten thousand people have died from this disease. Additionally, its incidence has been increasing around the world, which makes it a very serious health concern. In its early stages, it is difficult to distinguish them from other extremely frequent benign skin lesions, even for experienced dermatologists and the main cause of mortality by melanoma skin cancer is the late diagnosis, because it is capable of metastasizing to other parts of the body. Due to the latter, efforts have been dedicated to improving the awareness of the people about this disease, and the main alarming features of its presentation for which expert examination is recommended.

In order to alleviate this problem, the scientific community has been developing computer-aided systems that can quickly identify and characterize a skin lesion from dermoscopic images, attempting to automatically recognize the risk of it being a melanoma during its early stages. Several works have addressed this issue and obtained encouraging results, however there has not been much focus in determining the relevance of the features considered and how they may be important indicators of a lesion's malignancy.

The present work was dedicated to this issue, by focusing on three steps of an automatic classification system, namely the feature extraction, the feature selection and classification. In the initial approach a significant group of features was extracted from the lesions available in two datasets, after which their individual worth was evaluated using feature ranking methods. It was concluded that color features are the most consistent for melanoma recognition, with emphasis on simple statistics derived from the $L^*a^*b^*$ and $L^*u^*v^*$ color spaces, the red component of the RGB color space, and the average color saturation. After this an empirical search was performed in the feature space considered, in order to find the combinations of the features that lead to the best classification performances using a support vector machine (SVM) classifier with a radial basis function (RBF) kernel for evaluation in the two databases considered. The use of shape, color and texture descriptors alone was evaluated, and compared with the use of a combination of the three, and it was found that the best results could be achieved when using the three categories combined. The best sensitivity achieved for the first dataset was 89.66% (4 melanomas missed out of the 29 available) and for dataset 2 was 90.00% (4 melanomas missed out of the 40 available). The combinations of features that achieved the best classification performances were compared, and the features present in both were highlighted.

Index

List of Figures

List of tables

Symbols and Abbreviations

Abbreviation list

ANN	Artificial Neural Network
APN	Atypical Pigment Network
AUC	Area under the ROC Curve
CAD	Computer-Aided Detection and Diagnosis
CFS	Correlation based Feature Selection
CSLM	Confocal Scanning Laser Microscopy
CT	Computerized tomography
ELM	Epiluminescence Microscopy
FDG	Fluorodeoxyglucose
FN	False Negative
FP	False Positive
GLCM	Gray-Level Co-occurrence Matrix
HSL	Hue, Saturation and Lightness
HSV	Hue, Saturation and Value
KLT	Karhunen-Loève Transform
k-NN	K Nearest Neighbors
LDA	Linear Discriminant Analysis
LOO	Leave One Out
MIFS	Mutual Information based Feature Selection
MLP	Multilayer Perceptron
MRI	Magnetic Resonance Imaging
MSC	Melanoma Skin Cancer
NMSC	Non Melanoma Skin Cancer
NB	Naïve Bayes
PCA	Principal Component Analysis
PET	Positron Emission Tomography
QDA	Quadratic Discriminant Analysis
RBF	Radial Basis Function
RGB	Red, Green and Blue
ROC	Receiver Operating Characteristic
ROI	Region of Interest
SBFS	Sequential Backward Floating Selection
SBS	Sequential Backward Selection
SD	Simmetry Distance
SFS	Sequential Forward Selection
SFFS	Sequential Forward Floating Selection
SMOTE	Synthetic Minority Oversampling Technique
SN	Sensitivity
SP	Specificity
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
UV	Ultraviolet

Chapter 1

Introduction

1.1. Motivation

Skin cancer is one of the most common cancers diagnosed around the world. The disease can be roughly divided into non-melanoma skin cancer (NMSC) and melanoma skin cancer (MSC). In America, melanoma cancers account only for around 2% of all cases of skin cancer, but are responsible for the vast majority of skin cancer deaths (American Cancer Society, 2015a), thus being a major concern in public health.

The most promising strategy to reduce mortality of melanoma is to diagnose it early; however, differentiating it in its early stages from other benign pigmented skin lesions remains a challenging task even for experienced dermatologists. Several non-invasive *in vivo* imaging techniques have been developed that increase the amount of information that can be observed within a lesion and help assessing a more accurate diagnostic. The most commonly used in routine practice is the dermoscopy, a technique involving optical magnification using an immersion oil to render the epidermis translucent and thus enabling the visualization of morphological features not discernible with the naked eye (S. W. Menzies, 1999). To address the subjectivity that is still inherent to the diagnostic of melanoma, several screening and scoring methods have been implemented, but efficient methods to extend the diagnostic capability to general practitioners are still lacking.

These facts motivated the biomedical community to study the possibility of computer-supported skin lesion inspection and characterization, and the vast majority of research published in this field is dedicated to developing automatic means of melanoma diagnosis from dermoscopic images. These systems are intended to reproduce the biopsy decision making of the dermatologist when observing images of pigmented skin lesions, aiming to increase their specificity and sensitivity and reduce the morbidity related to lesion excisions (Korotkov & Garcia, 2012). The implementation of such systems in routine dermatological practice would be truly beneficial, due to the increased accuracy and reproducibility of the results that can be achieved by a computerized analysis.

Encouraging results have been reported that are capable of matching or even outperforming the average diagnostic performance reported for a dermatologist, but they were obtained on limited sets of data which cannot reproduce the wide variety of lesions that are encountered in clinical practice. Therefore, there is no single best solution that has been

obtained so far that can be implemented in routine clinical practice, and exists a compelling need to direct efforts to this field of research and discover knowledge that can create new or improve the existing solutions.

One common lack of information present in the reviewed research was the details about the feature selection procedure and the assessment of their relevance for the field of melanoma recognition. This motivated the development of the present work, which focuses on the implementation of feature extraction routines and an extensive study of their contribution to classification using challenging dermoscopic images datasets, consisting of early staged melanomas and benign melanocytic lesions.

1.2. Objectives

The first goal of this project was to conduct a literature review of the main research conducted in the field of the automatic classification of pigmented skin lesions from dermoscopy images. This step intended to identify the main steps required for the implementation of a computer assisted decision and diagnosis (CAD) system in dermatology, and survey the main algorithms that have been used for each. The main goal of this review was to emphasize the steps of *feature extraction*, describing the most commonly considered attributes for the task of automatic melanoma recognition and the methods used to extract them; *feature selection* and the existing methods to assess the relevance of the considered attributes; and the main *machine learning* algorithms used in this application and a comparison of their performance.

Following the literature review, the main goal of this study was to select relevant features to be extracted from pigmented skin lesions; to develop and implement computational methods for its automatic extraction from dermoscopy images; apply feature selection methods to assess the relevance of the features selected and evaluate their performance using state-of-the art machine learning algorithms. The purpose of each of these steps is briefly introduced below:

- *Feature Extraction* – Reduce the information present in a skin lesion to a group of relevant numerical attributes, capable of quantifying characteristics of the lesion that can be used to identify its malignancy;
- *Feature Selection* – Evaluate the attributes selected in order to discard irrelevant information and keep the most relevant for the classification step. This step was also used to provide information about the features more suited to the task of melanoma recognition;
- *Classification* – Evaluate the performance of a machine learning algorithm on the selected attributes to measure its ability to perform predictions on the skin lesions considered.

1.3. Document Structure

This document was divided in 4 chapters, with the exception of the introduction. A brief summary of the contents addressed in each is provided below:

- **Chapter 2 –** *Literature Review*: This chapter presents the main literature considered for the development of this work. Initially it presents a brief description of the theoretical biology concepts related to the formation of skin lesions; the main epidemiology data and risk factors associated with melanoma; and the techniques involved in assessing a diagnostic for it in clinical practice. The remainder of the chapter is dedicated to describing the main constituting blocks of a CAD system in dermatology, focusing on the recent works developed for the automatic classification of pigmented skin lesions;
- **Chapter 3** – *Methodology:* In this chapter the image datasets used, the software tools in which the work was developed and the general framework of the approach proposed are presented. Additionally, the main computational methods implemented for the development of this project, including the methods used for the extraction of the selected features from the images, the algorithms considered for the selection and evaluation of the features, and the machine learning algorithms considered for the evaluation of the prediction performance are described;
- **Chapter 4** – *Results and Discussion:* This chapter presents all the results obtained with the adopted approach together with their discussion. The main problems faced during the implementation of proposed methods and the results achieved will be made clear here;
- **Chapter 5** – *Conclusions:* This chapter consists of the concluding comments about the work developed, highlighting the main limitations on the strategies used and proposing improvements for future work on the subject.

1.4. Principal Contributions

The main contributions provided by the developed work may be divided in two important sections, namely the literature review and the discrimination of relevant features for the problem of automated melanoma recognition.

The literature review presents the reader with important theoretical background knowledge on the subject studied. It includes a summarized introduction of the biologic mechanisms underlying the formation of pigmented skin lesions, as well as the main types of lesions existing and the most important diagnostic criteria followed in the clinical practice. These are important concepts that inspire and integrate many of the computational methods implemented in this subject. This review also provides the reader with a fundamental introduction in the main computational methods developed in the context of the automatic classification of pigmented skin lesions, including the methods for performing the pre-processing of the lesion images; automatic segmentation of the lesions from the surroundings; the main features investigated; feature selection and classification algorithms. An overview of the classification performances achieved in the reviewed works is also presented in this section.

The discrimination of relevant features involved three steps, namely the *feature extraction*, *feature selection* and *classification*. In the *feature extraction* stage, significant features were selected from the reviewed literature, and the computational methods used for their extraction were based on previous algorithms developed, except for the calculation of

the lesion's asymmetry. To the best of the author's knowledge, the use of the asymmetry index as calculated, the 7 Hu's invariants, the solidity and the distance between the average (a^*, b^*) and (u^*, v^*) of the lesion and the surrounding skin had not been intensively studied before, for which the group of features used was unique and provided an important contribution to the knowledge in the field. For the *feature selection* stage, an extensive evaluation of the features' worth was performed, making use of state-of-the art *ranking* methods, *subset evaluator* methods and implementing an empirical search through the feature space, which had not yet been used in this context of application. The main contribution of this work consisted of this stage, through which the most relevant features from the group selected were determined and studied, and their prediction performance was evaluated using a support vector machine in two very heterogeneous datasets, validating the assumptions made and helping to determine the most relevant to the problem of automatic melanoma recognition. It is thought that the choice of the computational methods for each stage of the adopted approach was appropriate, and that the choice of features was representative of the problem at hand. An additional contribution was made by evaluating the performance of classification using individual groups of descriptors (shape, color and texture) and combining them.

Chapter 2

Literature Review

This chapter presents the reader with the main literature considered for the development of this work. It is presented with the goal of introducing the essential theoretical concepts that are used throughout this thesis and review the main research conducted on its subject.

Section 2.1 is dedicated to presenting introductory knowledge on the field of pigmented skin lesions, emphasizing the malignant skin cancer addressed in this work, the melanoma. It covers the basic underlying physiologic mechanisms responsible for the formation of pigmented skin lesions as well as its most common types. It also presents the main statistics related to the incidence and mortality of skin cancer and the risk factors associated to its development. This topic is concluded by presenting the steps involved in assessing a diagnosis for a skin lesion, focusing on the differentiating factors of the malignant melanoma. It includes the main imaging technologies used in clinical practice, as well as the existing scoring algorithms followed by the dermatologists.

Section 2.2 reviews the state of the art in the field of automatic classification of pigmented skin lesions acquired from dermoscopy images. It introduces the general pipeline adopted in the systems designed for this task, focusing on the computational methods used to extract relevant information from these images and the classification algorithms considered for its evaluation. This section also highlights the main results obtained from the reviewed literature, providing a benchmark to aid in evaluating the outcome of the approach proposed in this work.

2.1 Pigmented Skin Lesions

Pigmented skin lesions often appear in the skin's surface. The skin is the largest organ in the human body. It covers the internal organs and helps protecting them from injury, serves as a barrier to germs, prevents the loss of body fluids, helps to control body temperature, protects the body from ultraviolet (UV) rays and helps it producing vitamin D. The skin consists of three layers with distinct function and optical properties: the *epidermis*, the *dermis* and the *subcutis*, or subcutaneous layer (*Fig. 2.1*).

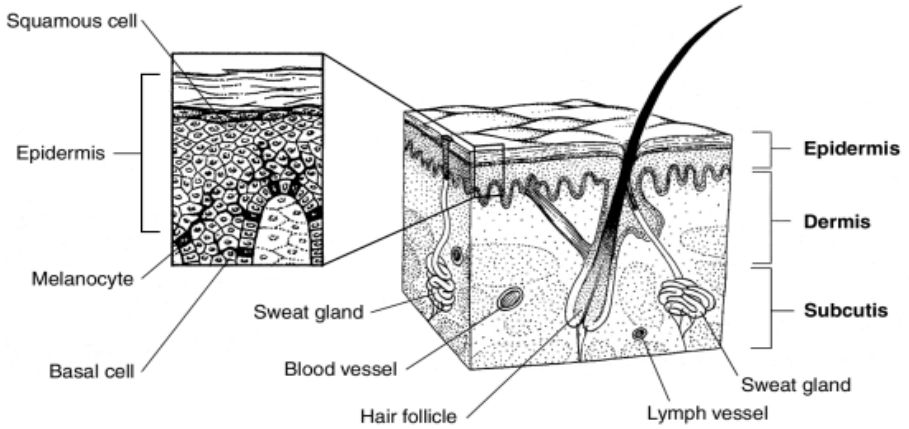


Figure 2.1 - Structural representation of the skin (American Cancer Society, 2015b)

The epidermis is the outer layer of the skin (see *figure 2.1)*. It is a very thin and tough layer, largely composed of connective tissues. Most of the cells of the *epidermis* are keratinocytes, responsible for producing keratin, a protein that helps the skin to protect the rest of the body against foreign agents, such as bacteria, viruses, heat or UV radiation (MacNeal, 2015). The keratinocytes are originated from the division of the cells in the deepest layer of the *epidermis*, the basal layer, and slowly migrate towards the outermost layer, the *stratum corneum*, where they become flattened and are gradually shed from the surface and replaced by newer keratinocytes pushed from below (MacNeal, 2015). When in the basal layer, keratinocytes are also named basal cells, which constantly divide to replace the cells in the *stratum corneum*. This layer is mostly composed of dead cells, the corneocytes, which are keratinocytes in their last stage of differentiation, flat cells with no nuclei or organelles, made up of mostly keratin, strongly contributing to the skin's barrier function. Due to their flat shape, these cells are also called squamous cells. Across the basal layer of the *epidermis* it is also possible to find melanocytes. Melanocytes produce a brown pigment named melanin, in response to stimuli such as UV radiation. This is the reason why, for most people, the exposure to the sun makes the skin to tan or darken (American Cancer Society, 2015b). The formation of melanin occurs in organelles inside the melanocytes' cell bodies, the melanossomes, which are carried to the intracellular region of neighboring keratinocytes and accumulate in the supranuclear region of these cells. In darker skinned individuals, the melanossomes contain greater amounts of melanin. This pigment strongly absorbs light in the blue part of the visible and the UV spectrum, and therefore acts as an important filter that protects the deeper layers of the skin from the harmful effects of the UV radiation (I. Maglogiannis & C. N. Doukas, 2009).

The middle layer of the skin is called the *dermis*. It is a thick layer of fibrous and elastic tissue (made mostly out of collagen, elastin, and fibrillin) that gives the skin its flexibility and strength. Within this tissue it is possible to find (see *figure 2.1*): hair follicles, responsible for producing the various types of hair found throughout the body, which helps in regulating body temperature; sebaceous glands, which secrete sebum into hair follicles, mainly to keep the skin moist and soft and also act as a barrier against foreign substances; sweat glands, that produce sweat in response to heat or stress; blood vessels, which carry nutrients to the skin and help regulating body temperature; and nerve endings, which provide the skin with its sense of touch, pain, pressure and temperature (MacNeal, 2015).

Below the *dermis* it is possible to find the *subcutis*, consisting of a network of collagen and fat cells. The *subcutis* serves as an energy store area, helps the body to conserve heat and has a shock-absorbing effect that helps to protect the body's organs from injury (American Cancer Society, 2015b).

The pigmented skin lesions can be divided into melanocytic and non-melanocytic, as they originate from melanocytes or not, respectively. Pigmented skin lesions refer to isolated growths appearing in the skin's surface that are distinctly colored from the surrounding skin. Although this is not the case with every skin lesion, they are the matter of focus in this work because most melanomas are pigmented and share important traits with the other pigmented skin lesions. These lesions may be divided into benign tumors, which are localized abnormal masses of cells that replicate at a normal rate and rarely pose a life threatening risk; or cancer, like melanoma, which are associated with an uncontrolled and abnormal growth of cells that form masses capable of invading other tissues and organs. A cancer in general is originated by a genetic mutation that occurs in a cell's DNA and may be caused by external (related to the environment) or internal (like genetic predisposition) factors. The mutated cell loses the ability to control its reproduction cycle, and starts multiplying and growing uncontrollably, leading to the formation of malignant masses. Cancerous cells may acquire the ability to detach from the malignant masses and migrate to nearby tissues, blood or lymph vessels. When they reach the vessels they can quickly disseminate to other organs and form more malignant masses, by when the tumor is said to have metastasized. Depending on the location and the cells affected, they present varying degrees of aggressiveness and risk of metastasizing. A skin cancer is a cancer that occurs in the cells of the skin.

The most common pigmented non-melanocytic benign skin tumors are the seborrheic keratoses (*figure 2.2 a)*), which are growths originated in keratinocytes that appear as elevated spots, with colors ranging from light tan to black, that have a waxy texture, and hemangiomas (*figure 2.2 b)*), which are growths originated from the blood vessels of the skin that appear as pinkish red regions (American Cancer Society, 2015b). With respect to the non-melanocytic skin cancers, the most frequent are the basal cell carcinomas (*figure 2.2 c)*), originated from the basal cells in the *epidermis* which often appear as growths that are either flat or small and raised, are colored pink or red, with translucent or shiny areas that may bleed following minor injury; and the squamous cell carcinomas (*figure 2.2 d)*), originated from the squamous cells in the *stratum corneum,* which appear as growing lumps, often with a rough surface, or as flat reddish patches that grow slowly (American Cancer Society, 2015a).

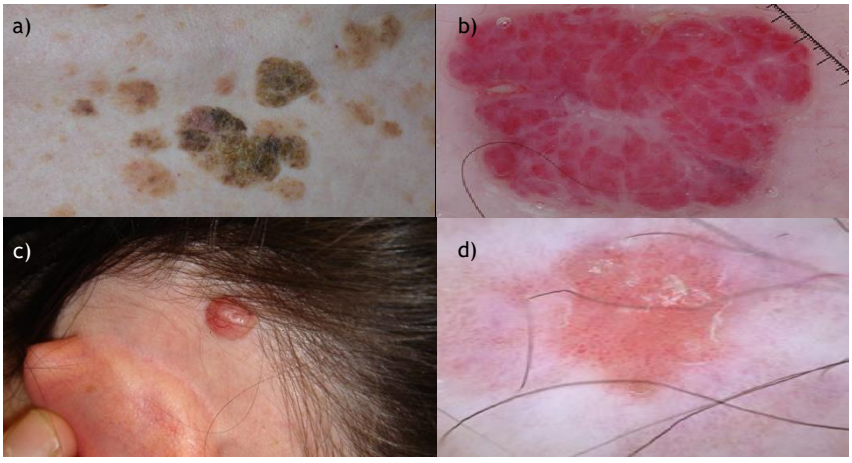


Figure 2.2 - Most common pigmented non-melanocytic skin tumors (Marghoob & Jaimes, 2015). Benign growths: a) seborrheic keratosis and b) hemangioma; Cancer: c) Basal cell carcinoma (BCC) and d) Squamous cell carcinoma (SCC).

The most common melanocytic benign skin lesions are the *melanocytic nevi*, or moles. These represent benign proliferations of a type of melanocyte known as a "nevus cell" (Schaffer & Bolognia, 2014b). The major difference between the ordinary melanocytes that reside in the basal layer of the *epidermis* and the nevus cells is that the latter cluster as nests within the lower *epidermis* and/or *dermis*, whereas epidermal melanocytes usually are evenly dispersed as single units. The proliferation of these melanocytes causes an excessive concentration of melanin, which creates coloured regions on the skin surface. The color of the mole will essentially depend on the localization of melanin in the skin. The color black is due to melanin located in the *stratum corneum* and the upper *epidermis*, light to dark brown in the *epidermis*, gray to gray-blue in the upper layer of the *dermis* and steel-blue in the deeper layers of the *dermis* (R. P. Braun, Rabinovitz, Oliviero, Kopf, & Saurat, 2005). The *melanocytic nevi* are usually classified as congenital or acquired. Congenital melanocytic nevi are present at birth or within the first few months of life, while the acquired nevi appear further in life. One additional pathologic difference between the two, is that congenital nevi tend to extend deeper into the *dermis* and subcutaneous tissue (Schaffer & Bolognia, 2014b). A typical congenital naevus is presented in *figure 2.3 a)*. The most frequent type of acquired nevi are the common nevi (see *figure 2.3 b)*), which have a wide variety of clinical appearances but usually present a small diameter, overall shape symmetry and a homogeneous surface. The denomination that is given to them is related to the location of the nests of melanocytes that originate them, which can be at the dermal-epidermal junction (*junctional nevi*), at the dermal-epidermal junction and in the *dermis* (*compound nevi)* or entirely in the *dermis* (*intradermal nevi) (Schaffer & Bolognia, 2014a)*. More unique cases of acquired nevi that share significant similarities to melanoma and often consist of difficult diagnosis are the blue nevi, Spitz nevi and the atypical nevi. A blue nevus is an intradermal nevus, for which the nests of melanocytes are located deep in the dermis, and the optical effects of light reflecting off the melanin there give it its blue or blue-black appearance (see *figure 2.4 c)*). A Spitz nevus typically appears in childhood or adolescence as a sharply circumscribed, dome-shaped, pink-red papule most commonly located in the face or the lower extremities. It can also appear with brown to black pigmentation, named the pigmented Spitz nevus, or even with heterogeneous shape organization, named the atypical Spitz nevus. Its clinical relevance lies in its close histologic resemblance to melanoma (Oakley, 2008). An example of an atypical Spitz nevus is presented in *figure 2.3 d)*. The atypical nevi (example in *figure 2.3 d)*), also known as dysplastic nevi, share some of the clinical features of malignant melanoma, such as asymmetry, irregular borders, multiple color and large diameter. Besides presenting features that make its differentiation from melanoma very difficult, they are also strong phenotypic markers of an increased risk of developing melanoma *(Schaffer & Bolognia, 2014a)*.

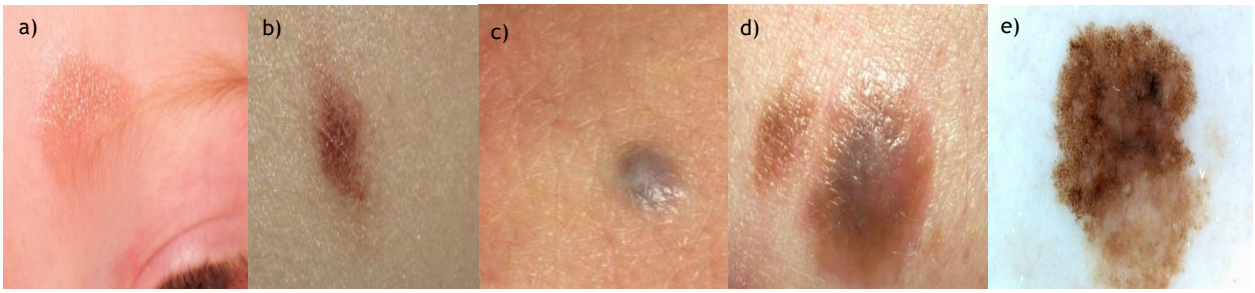


Figure 2.3 - Examples of benign melanocytic nevi: Congenital (Schaffer & Bolognia, 2014b) – a) congenital nevus; and Acquired (Schaffer & Bolognia, 2014a) – b) common nevus; c) blue nevus; d) atypical Spitz nevus and d) atypical nevus

Melanoma is the most serious form of skin cancer, and occurs when melanocytes become cancerous and start replicating and growing uncontrollably. Most melanomas often arise as superficial skin tumors that are confined to the *epidermis*, where they are termed *in situ*, and may remain for several years (I. Maglogiannis & C. N. Doukas, 2009). This stage is defined as the horizontal or radial growth phase, in which the tumors grow in size and develop irregular, asymmetric shapes. Because most melanocytes still produce melanin these tumors are usually presented with uneven distributions of brown or black color (American Cancer Society, 2015a). When the malignant melanocytes infiltrate the *dermis*, they leave melanin deposits there, thus changing the nature of the skin coloration and introducing characteristic hues in the visible tumors. When this happens, the melanoma is considered to be in a vertical growth phase and have metastatic potential. As the vertical phase develops, the melanoma becomes thickened and raised. Invasive melanomas can arise *de novo*, within the otherwise normal skin; from a horizontal growth phase melanoma; or less commonly from within other moles. The most common precursors to melanoma are the benign, atypical and congenital melanocytic nevi (Swetter & Geller, 2014).

There are four major types of invasive cutaneous melanoma. The superficial spreading melanoma (*figure 2.4 a)*) is the most common, accounting for approximately 70 percent of all melanomas (Swetter & Geller, 2014), and is the most often seen in young people. This melanoma grows along the top layer of the skin for a fairly long time before penetrating more. It starts as a flat or slightly raised brown to dark brown patch with irregular borders and asymmetrical form, and over time may present multiple shades of red, blue, black, gray and white, and a diameter ranging from just a few millimetres to several centimeters. The second most common is the nodular melanoma (see *figure 2.4 b)*), accounting for 15 to 30 percent of all melanomas (Swetter & Geller, 2014). This type has no identifiable horizontal growth stage, thus being the hardest to detect in its early stages. It usually appears as a darkly pigmented, pedunculated nodule. The lentigo melanoma (see *figure 2.4 c)*) is less frequent than the previous two, accounting only for 10 to 15 percent of all melanomas (Swetter & Geller, 2014), and most often appears in sun-damaged areas of the skin in older individuals. Its development is similar to the superficial spreading melanoma, beginning as a tan to brown lesion which gradually evolves in terms of size, shape and color. The least common subtype of melanoma is the acral lentiginous melanoma, which accounts for less than 5 percent of all melanomas (Swetter & Geller, 2014), and arise most commonly on palmar, plantar and subungual surfaces. It is the most common type of malignant melanoma among dark-skinned individuals, and can often advance more quickly than superficial spreading and lentigo melanoma.

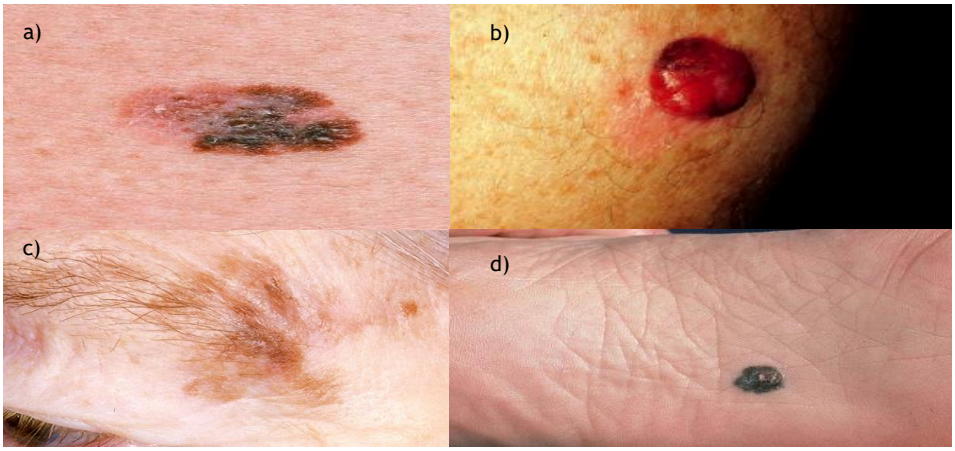


Figure 2.4 - Examples of the four major subtypes of malignant melanomas (Swetter & Geller, 2014): a) superficial spreading melanoma; b) nodular melanoma; c) lentigo melanoma and d) acral lentiginous melanoma.

2.1.1 Epidemiology of skin cancer

Skin cancer is divided into non-melanoma skin cancer (NMSC) and melanoma skin cancer (MSC). The most common forms of NMSC are the basal cell carcinoma (80%) and squamous cell carcinoma (16%) (American Cancer Society, 2015b). The melanoma skin cancer is often referred to as malignant melanoma or simply melanoma, and is the most aggressive form of skin cancer, as the following data confirms.

One study estimated that in America, 2006, 3.5 million cases of NMSC were diagnosed. However, the number of NMSC cases is difficult to estimate because these cases are not required to be reported to cancer registries, and most cases are highly curable (Rogers et al., 2010).

Regarding melanoma in the United States (US), in 2015, 73,870 new cases are estimated to appear, from which 9,940 thousand are estimated to be fatal (American Cancer Society, 2015a). It is the sixth leading cancer, representing 4.5% of all new cancer cases in the US. Melanoma accounts for only about 2% of all skin cancer cases, but is responsible for the vast majority of skin cancer related deaths. Its incidence has been increasing drastically for at least 30 years (see *figure 2.5*), while mortality rates have remained fairly constant. In 1975, the overall age-adjusted incidence across all population was of 7.9 new cases per 100,000 inhabitants, which almost tripled until 2012, to approximately 22.9 new cases per 100,000 inhabitants (National Cancer Institute, 2015). Although this indicates a significant increase in melanoma incidence, this can be in part due to a rise in screening for melanoma, leading to the detection of more, less developed tumors.

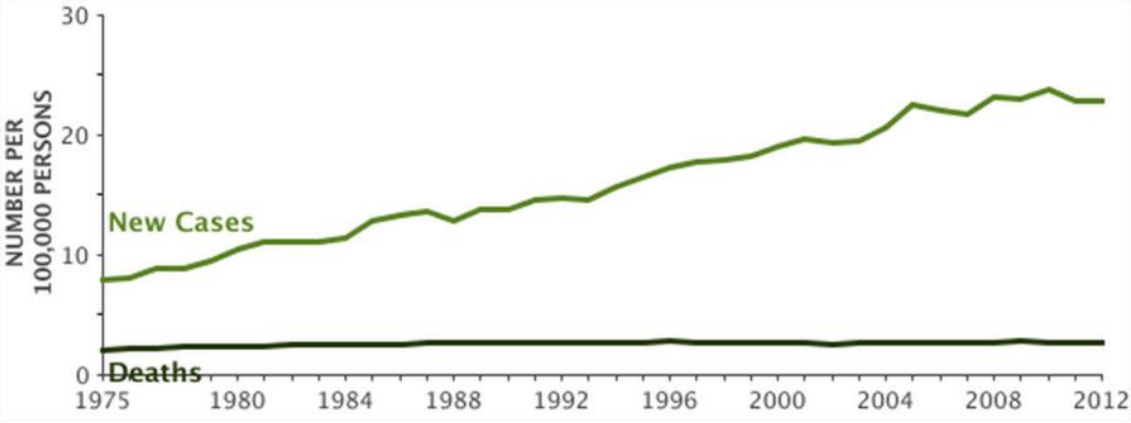


Figure 2.5 - Overall age-adjusted incidence and mortality of melanoma in the US from 1975 to 2012. Adapted from Surveillance, Epidemiology and End Results (SEER) data of National Cancer Institute (National Cancer Institute, 2015).

The most important indicator of the probability of survival from a melanoma is the cancer stage at diagnosis, which refers to the extent of the cancer in the body at the time it is diagnosed. The following results were obtained from the SEER data of the National Cancer Institute (National Cancer Institute, 2015) in the US, obtained for the years of 2005-2011, including cases from all races and both sexes. If the melanoma was found in its horizontal growth stage (84% of the cases diagnosed), the prognostic for the patient was very good and the 5-year survival rate (the percentage of patients diagnosed with the disease that has survived 5 or more years) was 98.3%. When it had entered its vertical growth phase (9%) but was still confined to the regional lymph nodes, the average 5-year survival rate decreased to 63%. Once the cancer has metastasized (4%) the 5-year survival rate was only 16.6%. This data highlights the importance of the early detection of melanomas.

There are some important risk factors that are associated with a higher probability of developing melanoma. However having a risk factor, or even several, does not necessarily mean that a person will get the disease. The following list summarizes the most important (Curiel-Lewandrowski, 2015):

- *Ultraviolet (UV) light* is considered a major factor for most melanomas. Although no direct causal relationship has been proved experimentally between exposure to UV radiation and melanoma, the evidence from indirect studies leaves little doubt that it is a major risk factor. This includes natural and artificial UV radiation, originated from the sun and from tanning lamps and beds, respectively;
- *Strong family history of melanoma:* having a close relative who has had the disease is also one of the most important indicators for increased risk of developing melanoma;
- *Previous melanoma or melanoma in situ:* a person that has had a melanoma in the past is at increased risk of having the disease again;
- *Multiple atypical nevi:* a person that has multiple atypical nevi is more likely to develop a melanoma;
- *High nevus count:* there is also a strong association between high common nevus counts and melanoma;
- *Increasing age:* most melanomas are diagnosed in people in the late fifties. According to the SEER data from 2008-2012, the median age of diagnosis was 63 years old (National Cancer Institute, 2015);
- *Male gender:* men are at slightly increased risk of developing melanoma. Based on the SEER data from 2008-2012, for the white-skinned population, the number of new cases per 100,000 inhabitants for men and women were, respectively, 33 and 20.2 (National Cancer Institute, 2015);
- *Light-colored skin:* the white-skinned population develops melanoma at rates 10-25 times higher than the dark-skinned, including African-Americans, Asians, American Indian and Hispanics. For example, the number of new cases per 100,000 persons for white American men was 33 in average from 2008-2012, while for the African-American men it was around 1.2 (National Cancer Institute, 2015).

2.1.2. Clinical evaluation of skin cancer

As the previous data suggests, the most effective way to deal with skin cancer is to detect it early. The best way to detect suspicious lesions is to evaluate all major areas of the skin regularly, and recognize new or changing skin growths, particularly those that look different from the surrounding moles. Every suspicious lesion should be presented to a health professional, either a general practitioner or preferably, if possible, to a dermatologist. If the analysis confirms the suspicion of a skin cancer, the lesion needs to be biopsied by a pathologist, which will perform an histologic analysis of the lesion sample and provide an explicit diagnosis (I. Maglogiannis & C. N. Doukas, 2009).

Although the most common cases of skin cancer are basal cell and squamous cell carcinomas, they are almost always cured without complications by a simple surgical excision, because they have little potential to metastasize (American Cancer Society, 2015a). Also, they present distinctive features from most of the existing skin lesions, which facilitate their diagnostic. In the case of melanoma, it can also be cured in almost all cases by surgical excision alone, if detected during its horizontal growth phase, when it is confined to the *epidermis* (Swetter & Geller, 2014). However during this stage, most melanomas appear as benign melanocytic nevi and their diagnostic is not trivial, even for experienced dermatologists. When the melanoma is in a vertical growth phase, it is considered to have metastatic potential. The probability of metastases with invasive, vertical growth-phase melanoma is most strongly predicted by measuring the thickness of the tumor (the Breslow depth), in millimeters, from the granular cell layer of the *epidermis* to the deepest malignant cell in the *dermis* (Swetter & Geller, 2014). In general, the deeper the measurement, the more chances there are for metastasis and the worse is the patient's prognosis, as suggested earlier. The following paragraphs focus on the distinctive clinical features of early staged melanoma and the main techniques used to assess its diagnostic.

The most common strategy for skin screening procedures by health professionals is a total body skin examination (TBSE). This method consists of analyzing every individual lesion in the body and look for specific clinical criteria that facilitate the recognition of early melanoma. Traditionally in a clinical setting, the lesions were analyzed by visual inspection often aided by a magnifying glass and, if available, compared to previous registries of the respective lesions, usually saved as photographs. However, the recognition of an early melanoma by visual inspection may be very challenging, even for the most experienced dermatologists (Swetter & Geller, 2014). Several sets of criteria have been developed to identify lesions that are suspicious for melanoma, and help general practitioners to decide lesions that should be referred for further evaluation by a specialist. The two methods more commonly applied are the ABCDE rule (Abbasi et al., 2004), and the revised Glasgow seven-point checklist (MacKie, 1990). The ABCDE rule is an extension of the ABCD rule (Friedman, Rigel, & Kopf, 1985) proposed in 1985, which is a semi-quantitative analysis of the parameters most crucial for the diagnosis of superficial forms of melanoma: the (A) asymmetry, by comparing the similarity between the two halves that result from a bisecting the lesion; the (B) border irregularity, which looks for ill-defined and irregular borders; the (C) color variegation defined by the variety of colors present in the lesion body (brown, red, black or blue/gray, and white); and the (D) diameter, which is associated to melanoma if over 6 mm. The ABCDE rule introduces the (E) evolving parameter to incorporate the essential concept of change of the lesion, including a modification over time of a preexisting nevus or the development of a new lesion. According to the seven-point checklist, the lesion should be evaluated according to three major criteria, namely changes in size/new lesion, shape and color; and four minor criteria, namely diameter larger than 7 mm, inflammation, crusting or bleeding, and sensory change. If any of the major criteria is observed, the lesion should be referred for further observation, while the presence of minor criteria reinforces the need for referral. However, limited information can be obtained through visual inspection. To address this issue, different non-invasive *in vivo* imaging techniques have been developed that improve the amount of information that can observed within a lesion and therefore can help in obtaining a more accurate diagnostic. These will be now briefly introduced, emphasizing the dermoscopy, since it is the most used method and how the images used in this work were obtained.

The most important and used imaging tool to aid in the diagnosis of pigmented skin lesions is the dermoscopy. Other synonyms include epiluminescence microscopy (ELM), incident light microscopy or skin-surface microscopy. It is a non-invasive imaging technique that links clinical dermatology and dermatopathology by enabling the visualization of subsurface skin structures in the *epidermis*, *dermoepidermal* junction, and upper *dermis*, not discernible with the naked eye. This technique involves the use of a hand-held incident light magnifying device (microscope) and immersion fluid with a refracting index that makes the *stratum corneum* more transparent to light and eliminates reflections (S. W. Menzies, 1999). In this mode of operation, contact of the microscope's glass and the skin is required, but cross-polarized lighting can also be used to obtain similar structure visualization without direct contact with the skin (Soyer, Argenziano, Chimenti, & Ruocco, 2001). Because melanoma's dermoscopic characteristics are well correlated to histopathological features, dermoscopy is an inexpensive and useful tool to aid dermatology practitioners, and it has been proved to increase dermatologists' correct assessment of malignant lesions (Lorentzen et al., 1999). An example of a benign melanocytic lesion and a melanoma acquired using a dermatoscope can be seen in *figure 2.6 a)* and *b),* respectively. These images give a clear idea of the additional information that can be obtained when using dermoscopy. In the atypical nevus, it is possible to distinctly perceive the presence of a pigment network, small areas with no pigment and a significant asymmetry of shape; while for the melanoma it is possible to identify the presence of a small blue-white area in the middle of the lesion (blue-white veil), an uneven distribution of color and also significant asymmetry of shape.

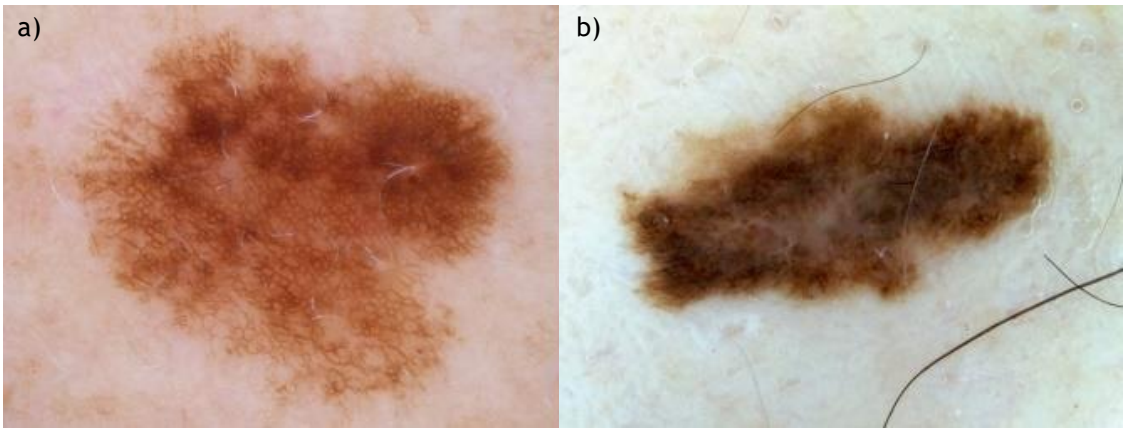
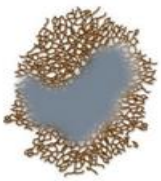


Figure 2.6 - Examples of pigmented skin lesions acquired using a dermatoscope (Marghoob & Jaimes, 2015): a) atypical nevus; b) superficial spreading melanoma.

To make use of the additional information provided by the dermoscopic images, new diagnostic methods were created and existing clinical criteria were adapted. The new diagnostic methods introduced were the pattern analysis (R. P. Braun et al., 2005), Menzies method (Argenziano et al., 2003) and the CASH algorithm(Henning et al., 2007). The adapted criteria were the ABCD rule of dermoscopy (Nachbar et al., 1994) and the seven-point checklist (Argenziano et al., 1998), which although presenting similar names to the ones previously described, incorporate new relevant information for assessing the diagnosis of a suspicious lesion. All of these methods incorporate in their evaluation visible dermoscopic structures that are presented differently across melanocytic lesions. In an attempt of summarizing some of these structures and their definitions, *table 2.1* was constructed, using illustrations found in (Marghoob & Jaimes, 2015). The use of illustrations was thought to be more general and easier to present than using regions of dermoscopic images.

Table 2.1 - Summary of the main dermoscopic structures present in melanocytic lesions (Marghoob & Jaimes, 2015).

Dermoscopic structure	Illustration
Pigment network: The typical pigment network is presented as grid-like network consisting of pigmented lines and holes without pigmentation.	
Negative Network A negative network consists of serpiginous interconnecting lines with low pigmentation, which surround irregularly shaped pigmented structures resembling elongated curvilinear globules.	
Dots and globules Dots are distinct small circular colored spots, and represent an accumulated localized pigment. Globules are usually defined as large dots (diameters greater than 0.1 mm).	
Streaks Streaks are radial projections at the periphery of the lesion, extending from the tumor toward the surrounding normal skin. This projections may be either in the form of *pseudopods*, which are fingerlike projections with small knobs on their tips, or *radial streaming*, which are the same structures without the formation on the tip.	
Blue-white veil Blue-white veils consist of irregular, structureless areas of confluent blue pigmentation with an overlying white “ground-glass” film.	
Blotches Blotches consist of usually homogeneous areas of dark brown to black pigment that obscure visualization of any other structures of the lesion.	
Regression areas Regression areas usually appear as white, scar-like depigmentations (lighter than the surrounding skin) and are often combined with adjacent blue-gray areas or peppering.	

In the ABCD rule of dermoscopy, the (A) asymmetry evaluates not only the asymmetry of shape but also the asymmetry of distribution of color and structures, and a score is given from 0 to 2; the (B) border is evaluated for the presence of abrupt cutoffs of pigment, the lesion is virtually divided into a pie with 8 equivalent segments, giving a score of 1 for each segment presenting an abrupt cutoff of pigment at the border; the (C) color is evaluated by giving a score of 1 for each of six colors present in the lesion (white, red, light brown, dark brown, blue-gray, and black); and in the (D) differential dermoscopic structures, the lesion is evaluated for the presence of any of five structures including pigment network, homogeneous/structureless areas, branched streaks, dots, and globules, and a score of 1 is given for each. The different parameters weight differently, and are summed up to obtain a final quantitative result that relates to the probability of a lesion being a malignant melanoma.

The seven-point checklist is based on the analysis of seven dermatologic characteristics typically found in melanoma, and they are divided into major criteria (atypical pigment network, blue-whitish veil and atypical vascular pattern) and minor (irregular streaks, irregular pigmentation, irregular dots/globules and regression structures) criteria, which value 2 and 1 points each, respectively. With this algorithm, a score of 3 or more points indicates the presence of a melanoma.

The Menzies method is based upon the evaluation of two negative features that are never present in melanoma, and nine positive features that are highly correlated with melanoma. The negative features include the symmetry of pattern around any axis through the centre of the lesion and presence of a single color (tan, dark brown, gray, black, blue and red). The nine positive features considered are the blue-white veil, multiple brown dots, pseudopods, radial streaming, scar-like depigmentation, peripheral black dots/globules, multiple (5 to 6) colors, multiple blue/gray dots and broadened network. The presence of both negative features virtually excludes the diagnosis of melanoma, while for all other lesions the presence of any of the positive features raises the suspicion for melanoma.

The CASH (Color, Architectural disorder, Symmetry and Homogeneity/Heterogeneity of dermoscopic structures) method is based upon evaluating a pigmented lesion for (C) the presence of few versus many colors (scoring 1 point for each color); for (A) architectural order versus disorder (score ranging from 0 to 2 points); for (S) symmetry of shape and pattern versus asymmetry (score ranging from 0 to 2 points) and for (H) homogeneity versus heterogeneity of dermoscopic structures (scoring 1 point for each structure). With this algorithm, a score of 8 or more is suspicious of melanoma. Only the ABCD rule of dermoscopy and the CASH algorithms take into account both the contour and distribution of colors and structures.

The pattern analysis is the most complex method of dermoscopic diagnosis and is based upon the association of an image with a recognition template developed from previous experience. This is the method preferred by experienced dermatologists but is not recommended for non-experts, since it requires significant knowledge and recognition of the global and local patterns of benign nevi and melanoma. In broad terms, with this method the clinicians try to identify a benign lesion by the presence of characteristic global features, which generally include having one or few colors, architectural order and symmetry of pattern or homogeneity; and a melanoma which usually presents a disordered distribution of structures, multiple colors, and an asymmetry of pattern.

The algorithms described are commonly used today in dermatology practice, and are all capable of achieving similar sensitivities (accuracy of detection of melanoma). However, since most algorithms define low thresholds for a lesion to be referenced as suspicious to avoid misclassifications, the average specificity (accuracy of detection of benign lesion) is very low. The four algorithms presented were reported to achieve an average sensitivity ranging from 78% to 94%, and an average specificity ranging from 46% to 83% (Marghoob & Jaimes, 2015). These values were obtained by averaging the ranges of values reported for each individual algorithm, and show that they can help achieving good detection rates for melanoma, especially when applied by an expert user, even with the likelihood of referring some unnecessary lesions for excision. The large range observed within the values is related to the varying experiences of the clinicians evaluated in the studies. Some algorithms are more suited to be used by general practitioners, such as the ABCD rule of dermoscopy, Menzies method and the seven-point checklist, because of their simplicity, overall accuracy and reproducibility.

However, even with the quantitative results these methods produce, the analysis is often still subjective and dependent on the user's experience. Additionally, the scope of observable structures is still limited when compared with other imaging techniques. Some additional imaging tools studied in the field of dermatology are briefly described next:

- *Commercially available photographic cameras* (Feit, Dusza, & Marghoob, 2004; Loane et al., 1997) which have the advantage of being affordable, provide easy data management and are useful in the follow-up management and easy comparison of lesions; however, they provide limited morphologic information due to low resolution and the variable illumination conditions limit the potential of a correct assessment.
- *Multispectral imaging* (Elbaum et al., 2001), which allows a quantitative and more objective analysis but the results are of difficult interpretation because of the complexity of the optical processes of scattering and absorption;
- *Confocal Scanning Laser Microscopy* (CSLM) (Lorber et al., 2009) and *reflectance confocal microscopy* (Pellacani, Cesinaro, & Seidenari, 2005) allow imaging of nuclear, cellular and tissue architecture of the epidermis and underlying dermal structure without a biopsy and allows recognition of abnormal intraepidermal melanocytic proliferation, but cannot evaluate tumor invasion depth reliably and is technically sensitive and expensive to use in routine clinical application;
- *Multifrequency electrical impedance* (Aberg et al., 2004) is a technique based on the eletrical impedance of a biological material which allows inferring about the molecular structure of the sample;
- *Magnetic Resonance Imaging* (MRI)(Premkumar et al., 1996), from which it is possible to obtain information on the depth and extent of the underlying tissue involvement and can be used to measure melanoma thickness or volume, but requires an adequate number of images per sequence for discriminating skin lesions;
- *Positron Emission Tomography* (PET) employing fluorodeoxyglucose (FDG) (Pleiss, Risse, Biersack, & Bender, 2007), which has proven to be highly sensitive and suitable in assessing the staging of various neoplasms;

2.2. CAD systems in Dermatology

Computer-aided detection and diagnosis (CAD) systems are increasingly being used as an aid by clinicians for detection and characterization of diseases. The fact that computers have the capability of storing and processing large amounts of data, perform complex calculations with high reproducibility makes them useful for implementing decision support systems. Machine-learning theory is a solid tool for implementing predictive models built on data acquired from actual cases. These are used in a variety of medical domains for diagnostic and prognostic tasks. Nowadays, there are commercially available CAD systems worldwide; for example, for breast cancer detection on mammograms (Elter & Horsch, 2009) and lung nodule detection on chest radiographs or on thoracic computerized tomography (CT) (Chan, Hadjiiski, Zhou, & Sahiner, 2008).

In the dermatology field, numerous works have been published regarding computerized diagnostic systems using digital images acquired by ELM. The general scheme of a CAD system for pigmented skin lesions can be seen in *Fig. 2.7*. The first step is acquiring the images obtained from ELM, which are preprocessed to reduce the negative effects of the artifacts they may contain, and enhance contrast between the pigmented lesion and the surrounding skin to facilitate the following steps. It is followed by the detection of the lesion region by image segmentation techniques. After the lesion's region is determined, different chromatic and morphological features can be quantified. A subset of these features should then be selected to avoid redundant or irrelevant features that can affect the classification accuracy and computation time, after which the subset can be used to classify the lesion being evaluated.

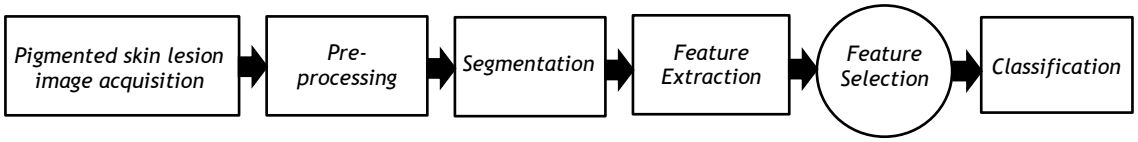


Figure 2.7 - Pipeline of a standard CAD systems for pigmented skin lesions in images.

In the following sections, the pre-processing and segmentation techniques found in the literature are briefly described, as they will not be the focus of the future work, and a comprehensive analysis of the techniques used for feature extraction, feature selection and classification is provided. A summary of the obtained results in the referred works is also presented.

2.2.1. Pre-processing

Pre-processing is a crucial step for an efficient CAD system of pigmented skin lesions. Its purpose is to prepare the image for the following segmentation procedure, by eliminating possible undesired artifacts and enhancing the image contrast between the lesion and the surrounding skin in order to facilitate the border detection step, while retaining its most important features. Dermoscopy images often contain artifacts such as uneven illumination, dermoscopic gel, black frames, ink markings, rulers, air bubbles and intrinsic cutaneous features that complicate the border detection, like blood vessels, hairs, and skin lines and texture (I. Maglogiannis & C. N. Doukas, 2009). The input digital images are often in RGB (red, green and blue) coordinates, each channel with 256 possible intensity levels, so each

pixel in the image can have one of almost 17 million possible colors. Because of this, other common pre-processing steps are color space transformation and color quantization.

One way to address the removal of these artifacts is to smooth the image using general purpose filters such as the median filters (Messadi, Bessaid, & Taleb-Ahmed, 2009), Gaussian filters (Maglogiannis, Zafiropoulos, & Kyranoudis, 2006) or anisotropic diffusion filters (Oliveira, Tavares, Marranghello, & Pereira, 2013). By choosing the adequate parameters, these can reduce the effects of most artifacts without causing loss of important image information. Alternatively, it is possible to use specialized methods for each artifact type; for example, the DullRazor technique (Messadi et al., 2009) for the removal of thick hairs. It is important to note that some filters used can cause blurring of the edges if not used carefully.

The segmentation of pigmented skin lesions is often poor due to insufficient contrast and smooth transition between the lesion and the surrounding skin. To enhance the contrast, the most common techniques used in image processing are histogram stretching, histogram equalization, homomorphic filtering, and high pass filtering. A contrast enhancement method based on independent histogram pursuit was proposed in (Gómez, Butakoff, Ersbøll, & Stoecker, 2008), by adopting an algorithm that linearly transforms the RGB image to a decorrelated color space in which the lesion and the background skin are maximally separated.

Due to computational simplicity and convenience of scalar (single channel) processing, the input RGB color image is often converted to a scalar image using different methods like retaining only the blue channel as lesions are often more prominent in this channel, applying the luminance transformation or the *Karhunen-Loève* transform (KLT) and retaining only the channel with the highest variance (Elgamal, 2013).

To minimize the computation requirements for processing the large amount of color information, it is common to perform color quantization in the pre-processing phase (Celenk, 1990). This process consists of two steps: the palette design, i.e. the selection of a small set of colors that represents the original image colors; and pixel mapping, the assignment of one of the palette colors to each input pixel. It has been shown (Celebi, Aslandogan, & Bergstresser, 2005) that to achieve precise quantization this method should reduce the number of colors in the image to 20.

2.2.2. Segmentation

Segmentation serves the purpose of delineating the region of interest (ROI), which in a CAD system for skin lesions is the lesion's area. This is an essential step for a reliable functioning of the whole system, and accurate skin lesion segmentation is a challenging task. A correct segmentation allows the following morphologic and chromatic features to be extracted from the lesion region alone, which in turn leads to a more accurate classification. Celebi et al. (M Emre Celebi et al., 2008) obtained the lesion's borders in the input images manually under the supervision of an experienced dermatologist. They argued that manual border detection was better than computer-detected borders because it separated the problem of feature extraction, focus of the work, from the problem of automatic border detection. However, for the development of an automated diagnostic system for skin lesions, it is important to develop automatic segmentation techniques. Numerous works have been published regarding the implementation of automatic segmentation algorithms for pigmented skin lesions. The main methods found in the literature for this step can be roughly divided into: histogram thresholding, region-based, active contours and clustering.

Image segmentation based on histogram thresholding (Emre Celebi, Wen, Hwang, Iyatomi, & Schaefer, 2013; Sezgin, 2004) uses the existing quantitative differences between the pixels in a skin lesion and the surrounding healthy skin to define a threshold upon which the pixels are divided into homogeneous regions.

Region-based approaches to segmentation involve splitting the image into smaller components and then merging the subimages that are adjacent and similar according to some feature, morphological or statistical criterion (Emre Celebi et al., 2008; Schmid, 1999).

Active contours algorithms include deformable models, snakes and their variants, and the principle behind this approach is the detection of the object's contours using curve evolution techniques, i.e. using colors or some other feature characteristic of the lesion as forces that deform an initial curve so that it evolve to match the object's boundaries (Erkol, Moss, Joe Stanley, Stoecker, & Hvatum, 2005; Z. Ma & J. M. R. Tavares, 2015).

Clustering algorithms involve the partitioning of a color or feature space into homogeneous regions using classification algorithms, like neural networks and fuzzy logic and have also been used in the segmentation of skin lesions (Vennila, Suresh, & Shunmuganathan, 2012).

2.2.3. Feature extraction

As referred previously, malignant melanomas are difficult to differentiate from other pigmented skin lesions, especially in its early stages. However, it is crucial that an automated computerized system for the diagnostic of skin lesions is capable of doing this differentiation with at least the same accuracy of a dermatologist for it to be accepted in a clinical setting. It is not intended to replace the professional opinion, but it would constitute an important tool to improve biopsy decision-making.

Similarly to the traditional visual diagnosis procedure, the computer-based systems search for features in the lesion regions and combine them to characterize the lesion. They have to be measurable and of high sensitivity, i.e. high correlation with skin cancer and high probability of true positive response and also high specificity, i.e. high probability of true negative response (Ilias Maglogiannis & Charalampos N Doukas, 2009). The lesions will therefore be characterized by a feature vector, an n-dimensional vector, containing the measures of the selected features that correspond to the objects of interest in the image. From the reviewed literature, it was understood that the extracted features can be divided into four major groups: shape features, color features, texture features and high-level features. The features studied for each group are described in the following sections.

a) Shape Features

Shape features are relevant for the diagnosis of melanoma because it often has asymmetric shape, irregular borders, and disordered architecture, while benign lesions normally do not. Many shape parameters are of simple computation, which makes them valuable if they prove to have good differentiation potential.

The most used shape features are the area, perimeter, compactness (M Emre Celebi et al., 2007; Messadi et al., 2009; Ruiz, Berenguer, Soriano, & SáNchez, 2011), thinness ratio/circularity index (M Emre Celebi et al., 2008; Harald Ganster et al., 2001), greatest diameter (M. E. Celebi et al., 2007; Messadi et al., 2009), eccentricity (M. E. Celebi et al., 2007), ellipticity (M. E. Celebi et al., 2008), symmetry distance (Vincent TY Ng, Benny YM Fung, & Tim K Lee, 2005), aspect ratio (M. E. Celebi et al., 2007; Ruiz et al., 2011),

asymmetry index (M. E. Celebi et al., 2007; Messadi et al., 2009) and border irregularity (P. Hall, Claridge, & Smith, 1995; Messadi et al., 2009). Other shape features have been explored, but these are the most commonly found in the literature.

Area and perimeter, A and p, respectively in the following equations, are two basic shape features that are used to calculate more relevant features of the lesion objects. Area is determined by the number of pixels inside the lesion's border. However, in (M. E. Celebi et al., 2007) the authors argued that this method is not very accurate for lesions with rough borders and so calculated the area using the method of bit quads (Bishop, 2006). The perimeter of a lesion is determined by the number of pixels that make up the border.

Compactness (C) is defined as the ratio of the lesion's area to the area of a circle with the same perimeter (M. E. Celebi et al., 2007; Messadi et al., 2009). In (Messadi et al., 2009), compactness is calculated as in *equation 2.1*. In (M. E. Celebi et al., 2007), an alternative way is used to avoid using the perimeter, which is also hard to estimate for objects with irregular borders. It is calculated as the ratio between the equivalent diameter (the diameter of a circle that has the same area as the object of interest) and the maximum diameter (maximum distance between two points in the object's border).

$$C = \frac{p^2}{4\pi A} \tag{2.1}$$

The aspect ratio (A_R), as calculated in (M. E. Celebi et al., 2007), is defined as the ratio of the length of the major axis (L_1) to the length of the minor axis (L_2):

$$A_R = \frac{L_1}{L_2} \tag{2.2}$$

where $L_{1,2}$ can be calculated as follows:

$$L_{1,2} = (8\left(\mu_{02} + \mu_{20} \pm ((\mu_{02} - \mu_{20})^2 + 4\mu_{11})^{1/2}\right))^{1/2} \tag{2.3}$$

Here m_{pq} and μ_{pq} are the *(p+q)*th order geometric and central moments of an object, respectively (see *equations 2.4 and 2.5*), where $(r_0,\ c_0)$ represent the coordinates of the object's centroid (see *equation 2.6*). The indexes i and j are used to denote the row and column of the lesion's pixel, respectively. Fikrle et al. (Ruiz et al., 2011) defined the aspect ratio as the ratio between the perimeter and area of the lesion.

$$\mu_{pq} = \sum_{i=0}^{rows}\sum_{j=0}^{cols}(i - r_0)^p \times (j - c_0)^q \tag{2.4}$$

$$m_{pq} = \sum_{i=0}^{rows}\sum_{j=0}^{cols} i^p j^q \tag{2.5}$$

$$(r_0, c_0) = (m_{10}/m_{00}\ ,\ m_{01}/m_{00}) \tag{2.6}$$

The ellipticity (ELL) of a lesion, a measure of its elliptical shape, can be determined as (M Emre Celebi et al., 2008):

$$ELL = \begin{cases} 16\pi^2 A_1 \ if\ A_1 \leq \frac{1}{16\pi^2} \\ \frac{1}{16\pi^2 A_1} \quad otherwise \end{cases} \tag{2.7}$$

$$\mu_{pq} = \sum_{i=0}^{rows}\sum_{j=0}^{cols} I(i,j)(i - r_0)^p (j - c_0)^q \tag{2.8}$$

$$A_1 = \frac{\mu_{20}\mu_{02} - \mu_{11}^2}{\mu_{11}^2} \tag{2.9}$$

where *I* is the binary lesion image, $(r_0,\ c_0)$ are the coordinates of the lesion's centroid, and μ_{pq} is the *(p+q)*th central moment of the object.

Eccentricity (ε) is a measure of an object's elongation, and can be calculated as (M. E. Celebi et al., 2007):

$$\varepsilon = \frac{(\mu_{02}-\mu_{20})^2+4\mu_{11}}{(\mu_{02}+\mu_{20})^2} \quad (2.10)$$

An object is asymmetric, if after dividing the object over its main axes and hypothetically folding one half of the object over the other, the two halves do not match. This is a very important shape feature for visual detection of a melanoma. This concept can be translated into a quantitative measure (M. E. Celebi et al., 2007; Messadi et al., 2009), also known as asymmetry index. Basically, the idea is to start by calculating the orientation (θ) of major axes of the object:

$$\theta = \frac{1}{2} tan^{-1}\left(\frac{2\mu_{11}}{\mu_{20}-\mu_{02}}\right) \quad (2.11)$$

Then rotate the object θ degrees clockwise in order to align them with the image (x and y). Then, a hypothetical folding is performed over the x axis, and the area difference (A_x) between the overlapping folds is taken. After repeating the same over the y axis, another area difference (A_y) is obtained. These can be used to calculate two asymmetry measures A_1 and A_2:

$$A_1 = \frac{\min(A_x,A_y)}{A} \times 100\% \quad (2.12)$$

$$A_2 = \frac{A_x+A_y}{A} \times 100\% \quad (2.13)$$

The symmetry distance (SD)(V. T. Ng, B. Y. Fung, & T. K. Lee, 2005) calculates the average displacement among a number of vertexes as the original shape is transformed into a symmetric shape. It is an alternative way of measuring the asymmetry of an object. It is determined by the amount of effort required to transform the original shape (P) into a symmetrical shape ($\hat{P}$), with n representing the number of vertexes considered:

$$SD = \frac{1}{n}\sum_{i=0}^{n-1} \left|\left|P_i - \hat{P}_i\right|\right| \quad (2.14)$$

The border irregularity requires accurate detection of the lesion's borders to be determined. The most common way of assessing the form irregularity is by calculating the compactness of an object. Authors (Messadi et al., 2009) have also used fractal analysis to calculate a border irregularity parameter. The first step was to determine the fractal dimension (d_{frac}) using the box counting algorithm. For this, they started by dividing the original image into an image of $m \times m$ pixels, which was then divided into cells of $s \times s$. The dilatation ratio, k, is the ratio m/s. At this point the number of cells (n) that contain a portion of the edge can be calculated:

$$d_{frac} = \log(n) \Big/ \log(k) \quad (2.15)$$

The plot of d_{frac} is a line passing through the origin, and its coefficients provide the fractal dimension in the $m \times m$ image.

b) Color Features

Color features can be very important in discriminating melanoma from other benign skin lesions, and are the most commonly used for automatic skin lesion characterization (I. Maglogiannis & C. N. Doukas, 2009). As referred previously the malignant melanoma tends to show a variety of colors across its surface, due to the melanin deposits out of the epidermis, the increased blood supply on the lesion periphery or even the lack of blood in the areas

destroyed by the cancer. Additionally, non-melanocytic skin cancers present particular color patterns that can be of interest for their detection.

The digital dermoscopy images are obtained in RGB components. To measure color features of skin lesions, the input RGB information has been directly used (Lucio Andreassi et al., 1999; M. E. Celebi et al., 2007; T Fikrle & K Pizinger, 2007; Messadi et al., 2009; Ruiz et al., 2011), but often combined with transformations of this color space that prove to be advantageous in dealing with uncontrolled imaging conditions, as is the case with the acquired dermoscopic images. The most common color spaces used are the normalized RGB (*rgb*) (M. E. Celebi et al., 2007); chromaticity coordinates (M Emre Celebi et al., 2008); HSL (hue, saturation, lightness) (Ruiz et al., 2011) and HSV (hue, saturation, value) (Barata, Ruela, Francisco, Mendonça, & Marques, 2014; M. E. Celebi et al., 2007), the most common cylindrical coordinate representations of points in an RGB color model and the International Commission on Illumination (CIE*)* $L^*u^*v^*$ (Barata et al., 2014; M. E. Celebi et al., 2007) and $L^*a^*b^*$ (Barata et al., 2014; M. E. Celebi et al., 2007; Umbaugh, Moss, & Stoecker, 1991) color spaces. Relative colors (M Emre Celebi et al., 2008; Y. I. Cheng et al., 2008; William V Stoecker et al., 2011), which equalize variations in the normal skin color among individuals, are able to compensate for variations caused by illumination and/or digitization process and are more natural from a perceptual point of view have also been used for feature extraction. In (Messadi et al., 2009) the image grid is divided in the RGB color space that best represents all colors present in all tumors using the k-means algorithm.

Color features extracted from these different channels are often statistic parameters, such as the mean, standard deviation and variance. Additional features include skewness, energy and entropy calculated from the red, green and blue channels histogram (Y. I. Cheng et al., 2008), color asymmetry (M. E. Celebi et al., 2007; T Fikrle & K Pizinger, 2007) and centroidal distance (M. E. Celebi et al., 2007)

The chromaticity coordinates (M Emre Celebi et al., 2008) are used to quantify the absolute color of a pixel. They are determined as the ratio between the R, G and B value of a pixel and the sum of all three. The advantage of using chromaticity over the raw R, G and B values is that it is invariant to illumination direction and intensity. In (W. V. Stoecker et al., 2011) the average absolute R, G and B chromaticity was used as well.

The relative color image in (Y. Cheng et al., 2008) was obtained as follows: the image was masked with the lesion's region, leaving only the skin untouched. The average value of the R, G and B values of the normal skin color were determined (R_S, G_S, B_S). The relative color image of the lesion was obtained by subtracting R_S, G_S and B_S to the R, G and B values of the lesion (R_L, G_L, B_L). This color space was then used for the extraction of color features. In (M Emre Celebi et al., 2008) , the relative color features extracted were the relative color difference, and relative color ratio. The first was obtained as described previously for obtaining the relative color image. The relative color ratios were obtained by dividing the R_L, G_L and B_L by R_S, G_S and B_S. In (W. V. Stoecker et al., 2011) nine relative color features were explored, including average of the relative R, G and B, average ratio between the relative G and B, average luminance ($luminance = 0.30 * R + 0.59 * G + 0.11 * B$) (Elgamal, 2013), calculated with relative colors, and introduced three new features, namely the average relative chromaticity, calculating the chromaticity in a similar way to the previously described, but using relative colors instead.

Celebi et al. (M. E. Celebi et al., 2007) extracted features from a combination of six color spaces, in order to achieve the most robustness in dealing with uncontrolled imaging

conditions. Additionally, for the calculation of color features, they divided the image into three significant regions: lesion, inner periphery and outer periphery. The region inside the border with an area of 10% of the lesion was not used for calculations, in order to reduce the effects of possible peripheral inflammation of the lesion and errors in the automatic border detection used. Inner and outer periphery was considered to be the region adjacent to the omitted region with 20% of the lesion's area. Color features were then extracted from the three regions in the various color spaces. The mean, a measure of the average color, and the standard deviation, a measure of the color variegation, were computed for the three regions in all three channels of the six color spaces used. Additionally, the ratio and differences of the two statistics over the three regions was also determined, as they believed they would provide significant diagnostic value; for example, information about the color transition from inside the lesion to the outside.

In their work, they used the color asymmetry to measure the asymmetry in pigment distribution. Its calculation was defined similarly to the shape asymmetry (A_1 and A_2) shown previously, but in this case the pixel value was incorporated in the calculations of the first order geometric moments and the second order central moments as weighting factors and the absolute brightness difference between the corresponding pixels in the two overlapping folds was accumulated. The calculation of color asymmetry was performed only in the RGB channels, in the three different regions. The centroidal distance was defined as the distance between the geometric centroid (r_0, c_0) of the lesion and the brightness centroid of that channel, calculated in a similar way to the geometric centroid, but including the pixel values as weight factors in the momentum calculations. If the pigmentation in a particular channel was homogeneous, the centroids would be close, and the centroidal distance would be small. Invariability to scaling was implemented by dividing the distance by the lesion diameter. The inner and outer regions were not considered in this calculation.

c) Texture Features

The extraction of texture features from skin lesion is also motivated by its diagnostic value in the differentiation of malignant and benign skin tumors. Texture information is an important and efficient measure to estimate the structure, orientation, roughness, smoothness or regularity of regions inside an image (Yuan, Yang, Zouridakis, & Mullani, 2006). Texture extraction methods include statistical, model and filtering-based methods. Some of the typically used filter banks are Laws masks, the dyadic Gabor filter bank, and wavelet transforms (Yuan et al., 2006).

In most works reviewed (L. Andreassi et al., 1999; M. E. Celebi et al., 2008; Shrestha et al., 2010; W. V. Stoecker et al., 2011), statistical texture descriptors were used. These are based on the gray level co-occurrence matrix (GLCM), the most commonly used statistical method of examining texture in images. The values of the GLCM are the result of calculating how often pairs of pixels with specific values and in a specified spatial relationship (in a defined pixel distance and angle offset) occurs in the image. Then, several statistical measures can be extracted. The most common measures extracted from pigmented skin lesion images are shift-invariant statistics, in order to obtain a texture characterization that is robust to linear shifts in the illumination intensity. These measures include maximum probability, energy, inertia, entropy, dissimilarity, contrast, inverse difference and inverse difference moment.

It is desirable to use the normalized GLCM, obtained by dividing each matrix element by the sum of all elements. In (Shrestha et al., 2010), the GLCM was constructed from the luminance plane of each RGB color image. Authors often construct the GLCM in each of the four directions {0º, 45º, 90º, 135º}, in order to achieve rotation invariance. The average and/or range of each statistic calculated over the four directions can then be used as the resulting features. In (M. E. Celebi et al., 2008), the average entropy, contrast and correlation were used. In (M. E. Celebi et al., 2007) the average maximum probability, energy, entropy, dissimilarity, contrast, inverse difference, inverse difference moment were used. In (Shrestha et al., 2010; W. V. Stoecker et al., 2011), the average and range of energy, inertia, correlation and inverse difference were calculated.

In (Shrestha et al., 2010), in order to reduce the computational cost of the computation of the GLCM and the derived statistics, the full dynamic range of the GLCM (256 gray levels) was reduced to 64 gray levels. This also reduced the effects of noise in the image. They also measured the referred features on the 3 different regions described in the previous section, together with the ratios and differences between the 3 regions for each feature. In (M. E. Celebi et al., 2008) the value of each feature in a pixel was considered to be the median of that feature in its 5x5 neighborhood, in order to avoid noisy results.

a) High-level Features

High-level features are related to the detection of dermoscopic structures (atypical pigment networks, globules/dots/blotches, streaks, granularity and blue-white veil) in images, as they have been used by dermatologists for differentiation between lesions. The introduction of contact dermoscopy in routine clinical diagnosis allows the differentiation of these structures in skin lesions, useful in detecting melanoma in its early stages. However, research on automatic detection of these structures is still scarce, due to the difficulty of relating lesion shape and color information to these structures, and the computational complexity involved. In the reviewed literature, works related to the detection of blue-white veil (M. E. Celebi et al., 2008), atypical pigment networks (Shrestha et al., 2010) and granularity (W. V. Stoecker et al., 2011) based on color and texture features were found. However, the detection was applied following manual determination of the structure's area by a dermatologist in the skin lesion area, from which the extracted features were used to determine the differentiation potential of these structures between malignant and benign lesions. An approach for the automatic detection of dots in picgmented skin lesions was studied by Skrovseth et al. (Skrovseth et al., 2010).

Blue-white veil (irregular, structureless areas of confluent blue pigmentation with an overlying white "ground-glass" film) is one of the most significant dermoscopic indicator of invasive malignant melanoma, with a sensitivity of 51% and specificity of 97% (Scott W Menzies, Crotty, Ingvar, & McCarthy, 2003). In the work addressing this structure (M. E. Celebi et al., 2008), a machine learning approach was used to detect its presence in dermoscopy images. This approach was based in classifying pixels from predetermined regions as being veil and non-veil pixels. From the initial image dataset, they selected a subset of images containing either sizeable pure veil regions or sizeable non-veil regions, on which a number of small circular regions containing either veil or non-veil pixels were manually determined. From each pixel of the resulting regions eighteen previously described features were extracted, fifteen color features and three texture features. Based on the extracted

features, a decision tree classifier was used to assign a label (veil or non-veil) to the analyzed pixels. They achieved sensitivity (percentage of correctly detected veil pixels) of 84.33% and specificity (percentage of correctly detected non-veil pixels) of 96.19%. With these results, the area of blue-white veil present in each lesion was calculated and used as a feature for lesion classification, from which the results are presented later.

An atypical pigment network (APN) is a black, brown, or gray network with irregular holes and irregularly distributed thick lines. On the contrary, a typical pigment network is light-to-dark brown, with regularly distributed thin lines and uniformly spaced holes (Shrestha et al., 2010). The presence of APN increases the likelihood of melanoma, but they may be present in a significant amount of benign melanocytic lesions as well, especially dysplastic nevi. Shrestha et al. (Shrestha et al., 2010) studied the possibility of discriminating malignant melanoma with APN and dysplastic nevi either having or not APN, using statistical texture descriptors alone. In order to do this, they selected twenty-eight malignant melanoma from the image database whose primary diagnostic features were atypical texture features (APN, branch streaks, radial streaming and pseudopods), not including those that displayed eccentric and irregular blotches, globules, blue and white areas and vascular features. Additionally, they selected seventy-eight dysplastic nevi images, with 12.8% of them containing an APN. A dermatologist then determined the area of each lesion that contained the most irregular texture, considered as the APN area, on all images. The texture features were extracted from the marked areas, allowing to study the discriminating potential of these dermoscopic structures.

Granularity is defined as an accumulation of tiny, blue-grey granules in dermoscopy images, and has been found to be significantly associated with the diagnosis of melanoma, with a sensitivity of 85%, specificity of 99% (R. Braun et al., 2007). Stoecker et al. (W. V. Stoecker et al., 2011) investigated the discriminating potential of color and texture features between melanoma and benign lesions, measured only from manually determined typical granular areas. Under the supervision of a dermatologist, the students marked the most typical granular portion of all melanomas from the image dataset, and marked the areas in the non-melanoma lesions that were as close as possible to the same color and texture as the granular spots. They extracted the ten texture features and nine absolute and relative color features referred in the previous sections from the delineated areas and proceeded to classification based on these features.

Skrøvseth et al. (Skrovseth et al., 2010) developed an automatic method for the detection of dots and globules. As the designation suggests, dots are distinct small circular colored spots, and represent an accumulated localized pigment. Globules are presented as large dots. The color they show is due to the melanin location, and is characteristic of the skin lesion. In the case of melanoma, they often appear as clusters of tiny dark dots, also referred to as cobblestone pattern, and are a key indicator of the malignant nature of the lesion. In (Skrovseth et al., 2010), they are found using a score for the regions of an image based on the binary contrast ignorant classification described in (Ojala, Pietikäinen, & Mäenpää, 2002). The first step used is finding the normalized cross correlation coefficient throughout the image with a reference image of a simple circular dot, finding candidate locations for the presence of a dot. Then, in the gray scale image, the gray value (g_c) of the central pixel is compared to the gray value (g_k) of the P surrounding pixels at a radius R, (see *equation 2.16*). The number of surrounding pixels, P, and the radius, R, for the analysis

should be determined empirically. The result is a score to the central pixel that will be high for dark spots, location of which is then determined by simple thresholding.

$$S_P(R) = \sum_{k=1}^{P}(g_c - g_k) \quad (2.16)$$

2.2.4. Feature selection

So far, a number of features that were extracted from pigmented skin lesions in the reviewed literature were described. In theory, they were all expected to provide significant value for the differentiation between benign from malignant skin tumors, but it was likely that some of them carried redundant or even irrelevant information for prediction on the datasets used. To address this issue, the group of features extracted from the images for classification was often submitted to the process of dimensionality reduction, to help selecting a small subset of features that allows the most effective classification. The benefits of this procedure are: reduced feature extraction time and storage requirement, reduced classifier complexity for better generalization behavior, increased prediction accuracy, reduced training and testing times, and enhanced data understanding and visualization [38].

Approaches to addressing the dimensionality reduction can be divided in *feature construction* methods, which project the original features into a new feature space with lower dimensions by performing combinations of the previous that are capable of improving prediction performance; and *feature selection* approaches that aim to select a small subset of features from the original that minimize the redundancy and maximize the relevance to the prediction of the target data.

Typical *feature construction* methods used are the principal component analysis (PCA), linear discriminant analysis (LDA) and canonical correlation analysis (CCA) (Tang, Alelyani, & Liu, 2014). PCA was applied in (Y. Cheng et al., 2008; Elgamal, 2013). It is a popular eigenvector-based multivariate technique which summarizes the variation in a correlated multi-attribute to a set of uncorrelated components, each of which is a particular linear combination of the original variables, the principal components (Y. Cheng et al., 2008). The PCA receives a matrix $X_{m,n}$ as input, with m being the number of features extracted and n the number of samples (lesion images), and outputs a matrix Y with the smallest number of features that account for the most variation of the original multivariate data.

Several algorithms have been proposed to implement *feature selection* for a classification problem, and are commonly categorised as *filters* or *wrapper* methods. *Filters* evaluate features independently of a machine learning scheme hence separating the bias of a learning algorithm from the bias inherent to a feature selection algorithm. They rely on general characteristics of the training data, such as distance, consistency, dependency, information and correlation (Tang et al., 2014). On the other hand, *wrapper* models use the prediction performance of a given machine learning algorithm to evaluate the attributes. One additional differentiation that can be made within *feature selection* methods is between methods that evaluate features individually, also called *feature ranking methods*, which are applied only for the *filter* approach, where the features are ranked according to a certain criteria and the higher ranked ones are considered to be the most valuable; and methods that evaluate subsets of features, the *feature subset evaluation* methods, which are applied both in *wrapper* and *filter* methodologies, and are differentiated by the search technique that is employed to investigate the feature space.

Feature ranking have been widely used as an approach to feature selection because of their simplicity, scalability, efficiency and good results (Guyon & Elisseeff, 2003). The main

disadvantage with this approach is that features are ordered according to some criteria, but they are not evaluated in the context of others, thus possibly missing the best combinations of features for a given task. It may be that the highest ranked features are redundant between them and that prediction deteriorates when used in combination, or that some of the lowest ranked features when used with one of the highest ranks provide the best class separability, for example. These methods provide significant speed and computational requirements reduction, since they require only the computation of a number of scores equal to the number of features used and their ranking, at the cost of some optimality in the result found. Celebi et al (M. E. Celebi et al., 2007) experimented with *feature ranking* methods in their work of classification of pigmented skin lesions, namely with the *ReliefF* and the *mutual information based feature selection* (MIFS) algorithms. The *ReliefF* algorithm starts by selecting from the data set a random number of samples (each sample being a vector with the values of the extracted features) and determines their nearest neighbors. The algorithm then compares the values of each sample's features with those of its neighbors, and scores the features by how well their values can distinguish samples that are near to each other. The MIFS method evaluates the mutual information between individual features and the class labels, and selects those that have maximum mutual information with class labels and are less redundant.

Feature subset evaluation methods usually cycle through the following steps: subset generation, subset evaluation, stopping criterion (Guyon & Elisseeff, 2003). The subset generation is a process of heuristic search, with each state in the search space specifying a candidate subset for evaluation. For the search method one must define: a search starting point, which can be either an empty subset, full or randomly selected subset of features; a search direction, dependent on the starting point, which can be forward (successively add features), backward (successively remove features) or bi-directional (add and remove simultaneously); and a search strategy, which can be sequential, successfully adding or removing features one at a time; random, generating subsets in a completely random manner; or complete, which completely searches the feature space. Each subset generated is evaluated according to the criteria defined and the process ends when the stopping criterion is reached. Feature subset evaluators provide the advantage of considering the value of features when used with others, possibly achieving combinations of features with higher prediction performance, but do it at the cost of increased computation time and requirements, and finding an optimal solution is still not guaranteed. The use of *feature subset evaluation* methods was found in (M. E. Celebi et al., 2007), namely with the use of the *correlation based feature selection* (CFS) algorithm. The goal of the CFS algorithm is to find a subset of the features that correlates well with the target class and has the minimum intercorrelation between features.

There is not one single optimal number of features that should be used in a classification problem, but the use of feature dimensionality reduction methods is very important to achieve the optimal number in a particular case. A very small number of features is not likely to be able to discriminate well between the classes, and a large number of features is not desired as it can lead to overffiting, i.e. the classification works perfectly in the training data but cannot generalize to unseen data, resulting in larger classification errors. Also the inclusion of irrelevant and redundant features often degrades the performance of classification algorithms, both in speed and prediction accuracy. However, in the reviewed literature it was found that details about the feature selection procedure are often absent

and that small relevance was given to the set of features that lead to the best reported classification. This results in an overall lack of information about which features are likely to perform better in the field of automatic classification of skin lesions.

2.2.5. Classification

The last step in a lesion recognition system is the classification, which is in charge of producing the diagnostic about the input images. That is, based on the previously selected measured features, the system needs to determine the class to which the lesion belongs to. This is the final goal of a CAD system for skin lesions. For the classification of dermoscopic images, there are two different approaches: one considers only a dichotomous distinction between the two classes (malignant of benign lesion); the other attempts to model $P(y|x)$, which not only assigns a class label to a lesion, but also a probability of each lesion belonging to a class, where y represents each of the possible class labels, and x the input skin lesion. The most used and effective techniques for a dichotomous classification are the support vector machines (SVMs) (M. E. Celebi et al., 2007; W. V. Stoecker et al., 2011). For the second approach, the most used are the k-nearest neighbors (k-NN) (Elgamal, 2013; Harald Ganster et al., 2001; Ruiz et al., 2011), artificial neural networks (ANNs) (Elgamal, 2013; Messadi et al., 2009; Ruiz et al., 2011; W. V. Stoecker et al., 2011), decision trees (M. E. Celebi et al., 2008; Yuan et al., 2006) and Naïve Bayes (NB) (Y. Cheng et al., 2008; Shrestha et al., 2010) classifiers. The task of classification involves two phases: the training phase, where previously classified data is used as input to a classifier to build a model that can best achieve the desired differentiation under each specific classifier modeling rules, and the test phase, where the input to the classifier should be unseen data and from which the classification results are used to assess the classification performance.

It is ideal that testing of the classifier is performed on a different data set from the one used to train the model in order to achieve an unbiased estimate of generalization error, and because the classification result may be overly optimistic if this is not done. If the original data set is too small to do this, n-fold cross validation can be used to make the best possible use of a limited amount of data. With this method, the data set is divided into n partitions, $n-1$ partitions are used for training, and the last piece is used for testing the classifier. Using n-fold cross validation, n classification models are built, to test on each n partition, and the classification results should be the average over all n test sets. The limit scenario, where only one sample is used for testing the classifier, is called the leave-one-out (LOO) method.

One common problem in classification tasks, is that the desired classes are not equally distributed in the available inputs. That is, there is often a discrepancy within the available database between objects of different classes. In dermoscopy classification tasks, there is often much more images of benign skin lesions than malignant, which may lead to poor classification performance, because the classifier focus on learning the larger classes and fails in classifying the minority, especially when both classes overlap significantly within the selected features (Japkowicz, 2000). To deal with class imbalance, sampling can be applied, either under-sampling (removing majority class samples) or over-sampling (adding minority class samples). In (M. E. Celebi et al., 2007), two sampling methods were compared: random under-sampling, which eliminates randomly chosen majority class samples; and synthetic minority oversampling technique (SMOTE), which over-samples the minority class by taking

each minority class sample and introducing synthetic samples along the line segments joining n of the k minority class nearest neighbors.

Another important step to take before the classification task is to normalize the features that characterize the samples. They often have different ranges and this can introduce significant classification errors, because of the different weights that features with different ranges of values can introduce in the classification model. In (M. E. Celebi et al., 2007), the following normalization method is used:

$$z_{ij} = \frac{(\frac{x_{ij}-\mu_j}{3\sigma_j}+1)}{2} \tag{2.17}$$

where x_{ij} represents the value of the jth feature of the ith sample; μ_j and σ_j are the mean and standard deviation of the jth feature, respectively. This guarantees that if each feature is normally distributed, 99% of z_{ij} are in the range of [0,1], and the out-of-range values are truncated to either 0 or 1.

a) Support vector machines

Support vector machines are kernel-based learning algorithms derived from the statistical learning theory (Cristianini & Shawe-Taylor, 2000). They are capable of building optimal separating boundaries between data sets by solving a constrained quadratic optimization problem. The basic training algorithm for SVMs is only capable of constructing linear separators; however, it is possible to use different kernel functions (linear, polynomial, radial basis function (RBF), and sigmoid) to include varying degrees of nonlinearity and flexibility in the model. In (Yuan et al., 2006), a 4th degree polynomial kernel was used, and in (M. E. Celebi et al., 2007), the RBF was chosen. They argue the preference of RBF kernel over polynomial as the first is capable of handling nonlinearity in a more computationally stable fashion, and requires less parameter tuning.

SVMs present advantages over the more classic classifiers. Their training mainly involves the optimization of a convex cost function, so there is no risk of getting stuck at local minima. They are based on the structural risk minimization (SRM) principle that minimizes the upper bound on the generalization error. Another advantage is that they provide a unified framework in which different learning machine architectures (e.g. RBF networks and feed forward neural networks) can be generated through an appropriate choice of the kernel (Vennila et al., 2012). The main disadvantage is that the classification result is purely dichotomous, and no probability of class membership is given.

b) K-Nearest neighbor

The K-nearest neighbor algorithm (Elgamal, 2013; H. Ganster et al., 2001; Ruiz et al., 2011), is a nonparametric method for classification. It is based on the principle that the instances within a dataset will generally exist in close proximity to other instances that have similar properties. One important aspect of the K-NN classifier is that the algorithm uses the data directly for classification, without building a model first. The training phase consists of simply storing all known instances and their class labels (Elgamal, 2013). It is very important for a good performance of the K-nearest neighbor classifier that the training set has enough examples of each class of pigmented lesions to adequately represent the full range of measurements that can be expected from each class. The only adjustable parameter in this method is K, the number of neighbors to use for label assignment. By varying this, the model can be made more or less flexible.

The process for determining the class of an unclassified sample t is as follows: the distances between t and each instance in the stored dataset is computed; then the distances are sorted in increasing numerical order and the first K elements are picked; in the end the class represented by the majority of the K closest neighbors will be the class of t. The value of $P(y|x)$ is calculated as the ratio of members of class y among the K-nearest neighbors of x.

The major drawback of the K-NN algorithm lies in the calculation of the distance between samples. In most applications, it is not clear how to, other than by trial and error, define a metric in such a way that the relative importance of data features is reflected in the metric (Stephan Dreiseitl et al., 2001).

c) Decision trees

Decision tree classifiers are popular due to the simplicity of implementation, efficiency in decision making and generation of easy-to-understand rules. The algorithm repeatedly splits the data set according to a criterion that maximizes the separation of the data, resulting in a tree-like structure. To do this, the algorithm identifies a variable and a threshold in its domain that can be used to divide the data set into two groups (Clark, 1997). The most common criterion used is information gain; which means that at each split, the decrease in entropy due to the split is maximized. In the end, the estimate of $P(y|x)$ will be the ratio of the y class elements over all elements of the leaf node that contains the data element x.

The main advantage of the decision trees is that they are not black box models and can be easily expressed as rules (Clark, 1997). They are also often fast to train and apply. The major disadvantage is that given a large training set, these classifiers, in general, generate complex decision rules that perform well on the training data but do not generalize well to the unseen data (Oates & Jensen, 1998). In such cases, the classifier model is said to have overfit the training data.

In (M. E. Celebi et al., 2008), the C4.5 algorithm (Bishop, 2006) is used, which prevents overfitting by pruning the initial tree, that is, by identifying subtrees that contribute little to predictive accuracy and replacing each by a leaf. For this algorithm, two parameters need to be adjusted, namely the confidence factor (C), that controls the level of pruning and the number of samples per leaf (M).

d) Artificial neural networks

An artificial neural network (ANN) is a mathematical or computational model that is inspired by the structure and/or functional aspects of biological neural networks. A neural network consists of several small processing units (artificial neurons) that are highly interconnected. In most cases, an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. The neurons are arranged in layers. The neurons in the first layer, named the input layer, are related to external data, and receive the feature vector of an object. Information will flow from this layer until the output layer. If it is a multilayer network, there will be intermediate layers, called hidden layers.

The most commonly used, and the simplest type of ANN is the feed-forward neural network (Elgamal, 2013; Messadi et al., 2009; Ruiz et al., 2011; W. V. Stoecker et al., 2011), named like this because the information flows in one direction only, forward, from the neurons in the input layer, to the neurons in the output layer. They are also called supervised networks, because they require a desired response in order to be trained. In order to train a network, the most common algorithm used is the back-propagation algorithm (Elgamal,

2013). At the training stage, the feature vectors are applied to the input of the network, and the desired output classes are known. After the information reaches the output layer, where the input features result in a class label, the back-propagation algorithm runs the network in the opposite direction, updating the weights and biases, which are often initialized randomly, between the connected neurons until all examples are correctly classified or a stopping criterion is reached.

e) Naïve Bayes

A naïve Bayes classifier (Y. Cheng et al., 2008; Shrestha et al., 2010) is a simple probabilistic classifier based on applying Bayes' rule, but assuming that the features (f_1, f_2,…, f_n) are independent. The Bayes' rule for a classifier can be written as (Bishop, 2006):

$$P(C|f_1, f_2, \dots f_n) = \frac{P(C)\times P(f_1,f_2,\dots f_n|C)}{P(f_1,f_2,\dots f_n)} \quad (2.18)$$

Where the class C is dependent on several features. The denominator is a constant design to normalize the probabilities so that they add, but since it does not depend on C, they can be effectively ignored in the classification. Because all the features are assumed to be independent, to nominator can be re-written as a product of the component probabilities, and the *a posteriori* probability of a sample belonging to class C knowing the features (f_1, f_2,…, f_n), becomes (Bishop, 2006):

$$P(C|f_1, f_2, \dots f_n)\ \alpha\ P(C) \times \prod_{i=1}^{n} P(f_i|C) \quad (2.19)$$

With this, the Naïve Bayes classifier assigns to a test sample the class with the maximum *a posteriori* probability. Like all the probabilistic classifiers under the maximum *a posteriori* probability rule, it arrives at the correct classification as long as the correct class is more probable than any other class.

2.2.6. Performance evaluation

In order to present the results of the classification, various metrics are available. In this section, some of these will be presented as well as the results obtained in the reviewed literature.

It is ideal, in order to obtain good classification performance, that a large data set of manually classified images is used. As previously referred, it is also important that this data set is divided into one set for training the classifier, and another for testing. The learning and test sets must be exchanged for all possible combinations to avoid bias in the solution (I. Maglogiannis & C. N. Doukas, 2009). The final results obtained should then be the average of the results obtained in the classification of each test set.

Most classification approaches in the area of skin lesions perform a dichotomous classification, between malignant (melanoma) and benign (all others) skin tumors. Some (Elbaum et al., 2001; T. Fikrle & K. Pizinger, 2007; Harald Ganster et al., 2001) attempt to provide a more specific classification, namely dividing the input objects into melanoma, atypical nevi (dysplastic nevi) and common nevi. Patients with atypical nevi are routinely followed up as these present a risk factor to the development of melanoma, and therefore, the routine examination for these patients can help in its early detection (Rigel, 1992). For an automatic skin lesion diagnostic system to be accepted in the clinical environment, it must

have very high correct classification results, especially in the classification of melanomas, as they might be deadly if not treated early.

The result of classification for each input skin lesion image can be divided in: true positive (TP), a sample correctly classified as melanoma; true negative (TN), a sample correctly classified as benign; false positive (FP), a sample wrongly classified as melanoma; and false negative (FN), a sample wrongly classified as benign.

The most common performance measures are accuracy, sensitivity, specificity and the receiver operating characteristic (ROC) curve. Accuracy is defined as the percentage of correctly classified samples (see *equation 2.20*); sensitivity is the percentage of correctly classified positive samples, also called true-positive rate (see *equation 2.21*); specificity is the percentage of correctly classified negative samples, also called true-negative rate (see *equation 2.22*); and the ROC curve is the plot of the sensitivity versus specificity, obtained after varying a threshold on a classifier's continuous output between its extremes.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \tag{2.20}$$

$$Sensitivity = \frac{TP}{(TP+FN)} \tag{2.21}$$

$$Specificity = \frac{TN}{(TN+FP)} \tag{2.22}$$

Accuracy is not an appropriate measure of the classification performance when the data is unbalanced, because it is strongly biased to favor the majority class. To avoid this problem in unbalanced data sets, the ROC curve is often used, as it illustrates the behavior of a classifier without regard to class distributions or error costs. The area under the ROC curve (AUC) can better measure the predictive performance of a classifier, is independent of the decision threshold and invariant to *a priori* class distributions (M. E. Celebi et al., 2007).

In Table 2.2, an overview of the best classification results obtained in each of the reviewed works is present. It is important to refer that these results are not directly comparable as the images used originated from different sources and represented different cases, but it gives an idea of the achieved results so far in this area of application.

The presentation of these results is organized as follows: on the leftmost column, the reference for each work's results is presented in separate rows. On the remaining columns, the conditions of each experiment and the classification results obtained are presented. For the definition of the former, the number of selected features for the classification trial, the machine learning algorithm used and the total number of images considered together with the percentage of representation of each class (namely the melanomas, the common benign and dysplastic nevi, if provided) were used. The two last columns define the performance achieved in each classification trial by means of the values of sensitivity and specificity obtained. As referred, the sensitivity is defined as the percentage of the positive instances correctly identified, which are considered to be the malignant cases; and the specificity defines the percentage of benign lesions correctly identified. In some of the papers reviewed, information about these parameters was not provided as the assessment of classification performance was determined by the AUC or the average accuracy of the classification experiments, and the results for these are presented as such.

Table 2.2 - Summary of the classification results of the reviewed works.

Ref.	No. of selected features	Classifier	Total images	Melanoma (%)	Dysplastic nevi (%)	Benign (%)	Sens. (%)	Spec. (%)
(Elbaum et al., 2001)	13	LDA	246	25.7	45.1	29.2	100	85
(Elgamal, 2013)	8	K-NN ANN	40	49		51	100 95	95 95
(Messadi et al., 2009)		ANN	180	40		60	67.5	80.5
(M. E. Celebi et al., 2008)	2	Decision tree	545	34.1		65.9	69.4	89.9
(M. E. Celebi et al., 2007)	18	SVM RBF kernel	564	48		52	93.3	92.3
(Ruiz et al., 2011)	6	K-NN Bayesian ANN Combined	98	52		48	70.2 85.1 95.7 97.9	76.5 76.5 78.4 78.4
(H. Ganster et al., 2001)	21	K-NN	5363	1.8	18.8	79.4	87	92
(T. Fikrle & K. Pizinger, 2007)	2	Logistic regression	260	17.7	18.1	64.2	91.3	81-91
(L. Andreassi et al., 1999)	13	LDA	147	38.8		61.2	88	81
(Y. Cheng et al., 2008)	11 3	NB ANN	285	56.1	14.7	29.2	67.5 86	62.8 70
(W. V. Stoecker et al., 2011)	11	ANN	288	30.6		60.4	AUC: 0.964	
(Yuan et al., 2006)	200	SVM Polynomial kernel	44	50		50	Average accuracy (%): 70	
(Shrestha et al., 2010)	10	Six classifiers	106	26.4		73.6	Average accuracy (%): 94.6	

These results show promising achievements in the field of automatic melanoma recognition. In fact, it can be seen that most works reported achievements in terms of sensitivity superior to 80%, which is comparable to the sensitivity reported for experienced dermatologists. In these works, high values of specificity were also achieved, which are of interest for lowering the number of unnecessary excisions that often result from the clinical evaluation of lesions. It is also interesting to observe that in almost all the works considered the number of features retained for the best classification performances was low, not exceeding 21 in all except one of the reviewed works. This is a strong indication that extracting the largest number of features for classification does not guarantee the best classification performance as most classifiers will perform best on a limited subset of the features available that carry the relevant information.

Although promising, it is important to highlight that for most works a limited set of dermoscopy images was available, thus failing to represent the wide variety of pigmented skin lesions that exists and are often presented in a clinical setting. This can lead to results that are overoptimistic and not valuable in practice. However, it should also be noted that these works compared the output of the classification experiments directly to the explicit diagnostic made by histology and not to the dermatologists' opinion, which makes the evaluation more difficult since it is based on the limited amount of information available in the 2D images.

As referred, some of the results reviewed are related to classification using features extracted from specific dermoscopic structures only. Stoecker and colleagues (W. V. Stoecker et al., 2011) found the best results for a combination of five color features and six texture features, indicating a significant contribution from both measures for the differentiation of skin lesions based on granularity. The best results were obtained when calculating the texture features with a pixel distance of 6, and the classification was performed with a standard back-propagation neural network. They have achieved an AUC of 0.964, having showed a very effective differentiation between the wide variety of benign and malignant lesions used.

Shrestha and colleagues (Shrestha et al., 2010) investigated the possibility of using statistical texture descriptors alone extracted from APN and non-APN areas to discriminate between melanoma and benign lesions. Pixel distances of 6, 12, 20, 30 and 40 were experimented for the extraction of texture features and six different classifiers were tested with the resulting feature vectors. The best results were obtained for GLCM computed for a pixel distance of 20, resulting in average classification accuracy from the six classifiers of 94.55%. They concluded that using only the correlation average measured from the APN areas, the results were similar (95.4%) and that the texture discrimination is critically dependent on the pixel distance used in the texture analysis. The results obtained allowed assuming that texture analysis from the APN areas can lead to good discrimination between malignant and dysplastic nevi.

It is important to refer that even though the results presented for discrimination based on dermoscopic structures (M. E. Celebi et al., 2008; Shrestha et al., 2010; W. V. Stoecker et al., 2011) are very encouraging, as they have showed high correct diagnostic rate even in challenging *in situ* melanoma cases (Shrestha et al., 2010; W. V. Stoecker et al., 2011), they face clear limitations, namely the manual detection of the differential structures and also limited number of images and variety of skin lesion cases analyzed.

The aforementioned results are related to the binary classification between melanoma and benign lesion. As has been referred, some of these studies (Y. Cheng et al., 2008; T. Fikrle & K. Pizinger, 2007; H. Ganster et al., 2001) have also tried the differentiation between the three classes, namely common nevi, dysplastic nevi and melanoma. The best results obtained were: in (H. Ganster et al., 2001), using 24-NN classifier trained with 270 lesions and tested on all 5363 images, 73% melanoma correctly classified, 53% dysplastic nevi and 59% common nevi; in (T. Fikrle & K. Pizinger, 2007), two classification experiments were performed, the first differing between malignant (melanomas) and benign (all others), where they achieved 91.3% sensitivity and 90.7% specificity, and the second with the benign group consisting of only atypical nevi, having achieved the same sensitivity but 80.7% specificity; and in (Y. Cheng et al., 2008), using a multilayer perceptron (MLP) artificial neural network with sigmoid neurons, they achieved 86.4% correct melanoma classification, 56.6% dysplastic nevi and 72.2% common nevi. These results prove that often dysplastic nevi can't be individually differentiated probably due to significant overlapping of attributes with melanomas'.

2.3. Summary

This chapter presented a review of the literature used for the development of this project, and was divided in two main sections. The first, dedicated to the biological perspective on the pigmented skin lesions, intended to give the reader an idea of how these lesions are originated and the characteristics they normally present. This section focused mainly on aspects related to the melanoma, namely how it constitutes a threat to public health, and the techniques used to differentiate it from other benign lesions in clinical practice. The second part of this chapter was dedicated to exposing the main steps that constitute the standard pipeline of a CAD system in dermatology, and to present the main computational methods used in the reviewed literature to address each step, emphasizing the feature extraction, selection and classification. It is thought that the works reviewed were up to date and showed several different approaches to the problem studied, and were therefore representative of the state of the art for this field of research.

The first section presented simple theoretical concepts to understanding how the pigmented skin lesions are originated, and also a summary of the characteristics they exhibit on the skin's surface. The most important ideas to retain from this section are the wide diversity of pigmented skin lesions that exists, and especially the phenotypic similarities that malignant melanomas in early stages of development share with other benign melanocytic lesions. The benign melanocytic lesions are extremely frequent, and one person may have several hundreds of common nevi, all with different aspect and characteristic attributes, and so the identification of the appearance of a new or suspicious lesion is often not easy, but they should always be evaluated by a professional. However, even for the most experienced dermatologists, assessing a correct diagnostic for these lesions is not trivial and failing to do so can lead to, on one hand, to the unnecessary excision of a harmless benign lesion, or on the other, to the overlooking of a malignant lesion, which may evolve and metastasize to other parts of the body, stage where no treatment is available and can therefore lead to death of the patient. Although the introduction of dermoscopy and the scoring algorithms allowed increasing the amount of information that could be used to detect the malignancy of a lesion, the evaluation remains subjective and in the hands of practitioners unfamiliar with

the imaging technique it may even lower their diagnostic accuracy by visual inspection (Kittler, Pehamberger, Wolff, & Binder, 2002). This problem is aggravated further in under developed regions, where access to specialized health professionals is limited. These are the main reasons that motivate the research for the implementation of a CAD system in dermatology, which should be able to reduce the mortality and morbidity associated with the melanomas, and to extend the diagnostic ability of professional dermatologists to general practitioners.

The second section presented some of the efforts made so far to address the implementation of such a system applied to dermoscopy images. As it was not the main focus of this work, only few examples of methods commonly used for preprocessing the images and automatically segment the lesion region were provided. A description of the several features explored in the literature studied was also given in this section, and these were often inspired on the diagnostic cues used by dermatologists. Most works included the extraction of a combination of shape features, to characterize the degree of asymmetry and irregularity of shape and borders of the lesion; color features, to describe the colors and their distribution across the lesion; and texture features, to account for the presence of dermoscopic structures that may differentiate melanomas from the benign lesions. It was found that few of the works reviewed explored the use of single categories of descriptors to characterize the lesions, an approach that is addressed in this work. Additionally to the extraction of features from the whole lesion region, some authors explored the influence of color and texture features measured from manually delineated dermoscopic structures on benign and malignant lesions, and reported a significant contribution from considering information from these structures alone for the differentiation of challenging atypical nevi from melanomas.

A description of the general framework for selecting valuable features for classification was also provided in this section. One common limitation found on the reviewed literature was the absence of information about this step and its results, especially in the works that use a large number of features measured from the whole lesion region. Several works considered similar features, but only reported the classification results obtained using a combination of features selected automatically by the selection methods, without providing information about the value of the features used in the experiments. Therefore, research was still lacking in determining those that can provide the most discriminative potential for melanoma recognition. One strong motivation for the development of this work was to tackle the latter, by characterizing and identifying the value of several features for this context of application.

Regarding the classification step, several different approaches have been explored, and promising results have been achieved so far, matching or even outperforming the average diagnostic accuracy reported for experienced dermatologists. These studies are very encouraging, as most have used automatic segmentation of lesions, validating the possibility of a complete CAD system. However, most faced the common limitation of having limited number of samples, translating into low variability between the lesions available and limiting the generalization ability of the results obtained to a practical setting. The experiments using features extracted from the dermoscopic structures also proved to have significant value, emphasizing the value of directing research towards automatic detection of these structures.

The methods adopted in this project to achieve the proposed goals are described in the following chapter.

Chapter 3

Methodology

The present chapter describes the methods chosen during the work to address the proposed problem of identifying relevant features for the detection of malignant skin cancer, melanoma. To do this, it was divided in six subchapters:

- *Image Datasets*, in which the source of the dermoscopic images used is presented, as well as the number of benign and malignant lesions they contain;
- *Software*, introducing the software packages used for implementing the desired methods;
- *Overview of the proposed system*, highlighting the sequence of processing steps applied;
- *Masking of the lesion images*, describing the process of separating the region of interest (lesion) from the surrounding (healthy skin);
- *Feature Extraction* presents the features selected for extraction from the dermoscopic images, the mathematical methods used to obtain them and the rationale behind their choice;
- *Feature Selection and Classification* exhibits the various methods experimented for the selection of relevant features, and summarizes the classifiers used to assess their significance;

3.1. Image Datasets

For this work, two different databases were used, adding up to a total of 300 images. Both consist of pigmented skin lesions acquired using a digital dermoscope in a clinical setting, and include benign lesions and malignant melanomas. All images used were accompanied with its histopathology diagnostic, reference used as the ground-truth to evaluate the performance of the implemented classifiers, and also the manual segmentation of the lesion from the healthy skin, performed by an experienced dermatologist.

Most part of the work was dedicated to the first dataset, consisting of a set of 100 dermoscopy images, from which 29 were malignant melanomas, most in early stages of development, and 71 were benign lesions, common nevi and dysplastic nevi. These images

were obtained from the EDRA Interactive Atlas of Dermoscopy (Argenziano et al., 2002), and the dermatology practices of Dr. Ashfaq Marghoob (New York, NY), Dr. Harold Rabinovitz (Plantation, FL) and Dr. Scott Menzies (Sydney, Australia). These are 24-bit RGB color images with dimensions ranging from 577 x 397 pixels to 1921 x 1285 pixels (Emre Celebi et al., 2008). The reason to focus the work on this dataset was that, although small numbered, it contained challenging malignant lesions and therefore good results in the classification of this database were expected to lead to good results in others.

The second database, PH^2 database from faculty of science of Oporto's university (FCUP), was used mostly to help validate the results obtained from the first. Its images were obtained at the Dermatology Service of Hospital Pedro Hispano (Matosinhos, Portugal) under the same conditions through *Tübinger Mole-Analyzer* system (developed at the University of Tuebingen, Germany) using a magnification of 20x. These are 8-bit RGB color images with a resolution of 768x560 pixels. This image database contains a total of 200 dermoscopic images of melanocytic lesions, including 80 common nevi, 80 atypical nevi and 40 melanomas.

One evident limitation of both datasets is the lack of diversity of pigmented skin lesions, especially melanomas. The small number of images does not allow accounting for the wide variety of malignant and benign lesions found in everyday clinical practice, and hence the classification models created will not be capable of identifying and differentiating every type of pigmented skin lesions, but still it is believed that the datasets used were representative of the problem at hand and therefore would allow drawing useful conclusions on the subject.

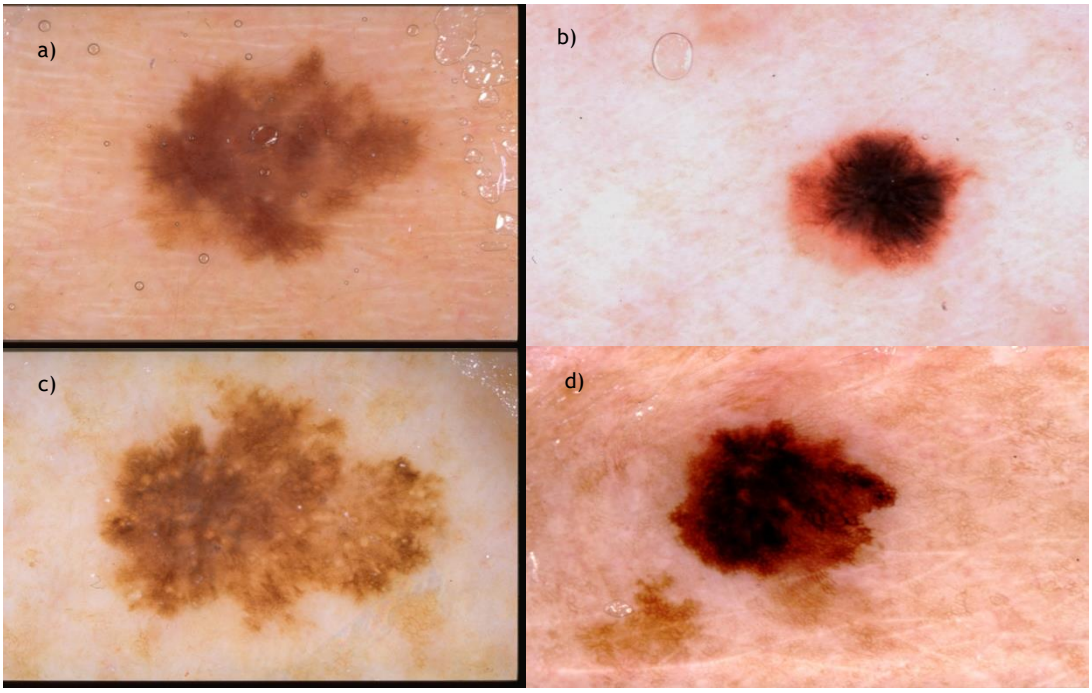


Figure 3.1 - Examples of challenging lesions of the first dataset for classification: a) and b) (top row) are atypical benign lesions. c) and d) are melanomas, similar in shape and color to a) and b), respectively.

To justify the difficulty present in the 100 images' dataset, see Fig. 3.1. On the top row two examples of atypical benign nevi are presented, and below them are the two examples of melanomas that were considered to be the most similar to the previous. In images a) and c) it is possible to see similar brown color and shape asymmetry, having the visual diagnostic probably to be upon the presence of dermoscopic structures, namely dots and pigment network which should be characterized by texture, the asymmetrical distribution of color

throughout the lesion, and maybe the irregularity of the border. On the other pair of images, b) and d), shape, symmetry and colors present are very similar, as well as the pattern of visible textures, however the color distribution seems to be less even on the melanoma case. These subtle differences are difficult to detect in a computerized analysis and therefore pose a significant challenge for automatic classification. It seems to be a constant issue across this dataset, and so it was chosen to apply the feature extraction methods and investigate its relevance.

On the other hand, the PH^2 dataset presents a wide variety of benign lesions, both common nevus and atypical nevus (dysplastic, intradermal and blue nevus) but most malignant melanomas it contains appear to be in a more advanced stage of development (see *Fig. 3.2 c)* and *d)*), with large areas of intensified black, signs of ulceration (*Fig. 3.2 c)*), and clear presence of blue-whitish veil (*Fig. 3.2 d)*). Although this characteristics can appear in other skin pigmented skin lesions, in this database are mostly specific to melanomas, and therefore present significant visual differences from the benign lesions, which should facilitate the correct classification of these lesions. Nonetheless, some benign lesions should pose a challenge for automatic detection, as they show large dark black areas and the presence of gray whitish areas (see *Fig. 3.2. a)* and *b)*, respectively) and also often have asymmetric shape and irregular borders. One additional limitation of this dataset is that some lesions do not fit completely over the image frame, as can be seen in *Fig. 3.2 b), c)* and *d),* and consequently the lesion border considered for those images was the image border, which deteriorates the extraction of some features, as is discussed later. Because of the aforementioned reasons, this image database was mostly used to validate the results obtained for the initial 100 images.

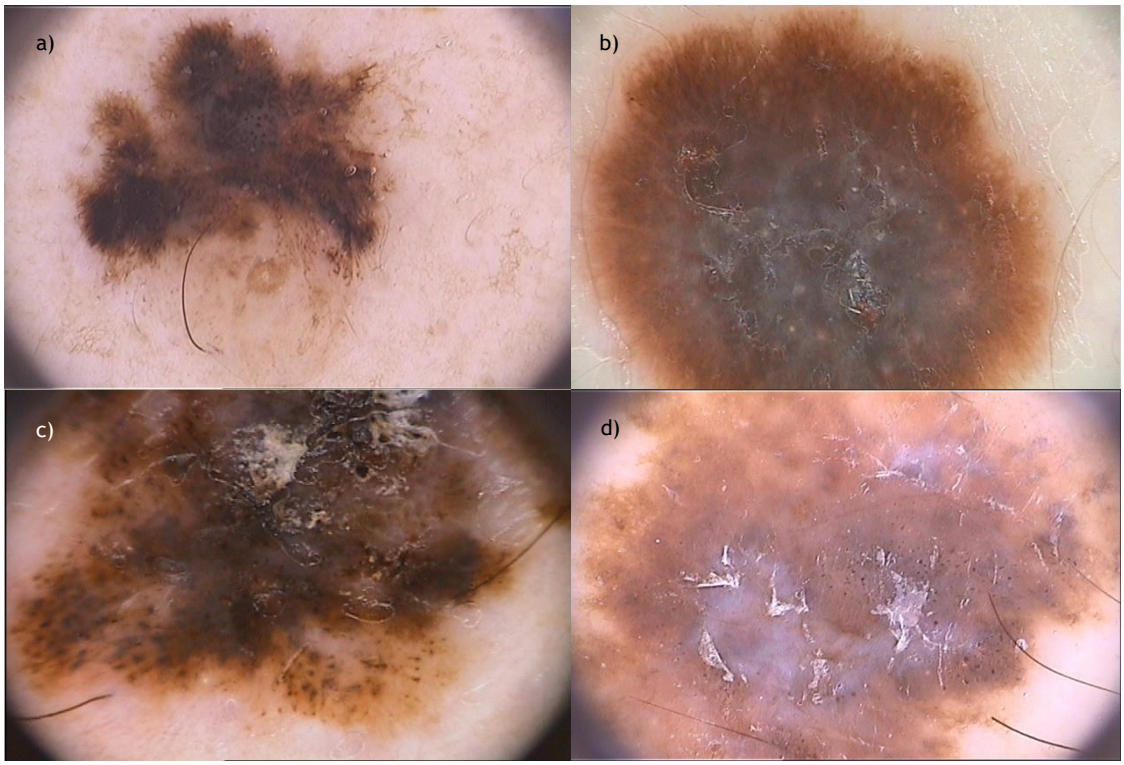


Figure 3.2 - Examples of lesions from the PH^2 dataset: a) and b) represent atypical benign lesions whose diagnostic is difficult; c) and d) represent developed melanomas with distinct features from the available benign lesions.

3.2. Software tools

For the development of the presented work two main software packages were used, MATLAB version R2013a® (developed by Mathworks Inc.,Natick, Massachusetts, United States) and WEKA version 3.6 (Waikato Environment for Knowledge Analysis, developed at the university of Waikato, New Zealand) (Mark Hall et al., 2009).

MATLAB is a numerical computing environment which allows matrix manipulation, plotting of functions, data analysis, and algorithm implementation. MATLAB can be used for a range of applications, including signal processing and communications, image and video processing, computational biology, etc. The amount of available functions dedicated to image processing makes it a viable option for this area of application. It stores most images as matrices, in which each element corresponds to a single pixel in the digital image. This software was used in this work for extracting the studied features from the images, and also to aid in implementing automatic feature selection and classification routines.

WEKA first release was in April 2000 (Mark Hall et al., 2009), and was developed to answer the need of a unified workbench that would allow researchers easy access to state-of-the-art techniques in machine learning and data mining. The most recent version available of this software is the 3.6, which was used during this project. It was implemented in JAVA programming language, and works under an intuitive graphical user interface (see *Fig. 3.3*) and the latest version provides a comprehensive collection of data pre-processing, feature selection, classification and clustering algorithms. It permits quickly trying and comparing different machine learning methods on a given dataset. In this work, it was used to test several combinations of feature selection and classification algorithms on features extracted from the pigmented skin lesions with MATLAB, providing useful information regarding the choice of the attributes.

As can be seen in *Figure 3.3*, its simplest GUI has six tabs on the top of the window: the *preprocess* tab, where data can be visualized and filtered to prepare it for further analysis; the *classify* tab, with algorithms available to implement supervised classification, which requires a nominal attribute to be provided containing the label of each image; the *cluster* tab, which allows to apply unsupervised classification routines, such as the KNN algorithm, where no label known *a priori* is required; the *associate* tab has functions that permit deriving the underlying association rules within the data; the *select attributes* tab contains the algorithms to perform feature selection, where one must choose an attribute evaluator (for example the *Pearson's correlation attribute evaluator)* and a search method, defining how the feature space should be investigated; and finally the *visualize* tab, which allows visualizing the data graphically with relation to every pair of features.

One additional software package worth mentioning in this section is the LibSVM (developed by Chih-Chung Chang and Chi-Jen Lin in Taiwan)(Chang & Lin, 2011). It is an integrated software for applying different SVM formulations for classification. There are extensions available for this package to work under both MATLAB and WEKA environment. It was used to apply support vector classification to the generated data, as discussed further in *section 3.6*.

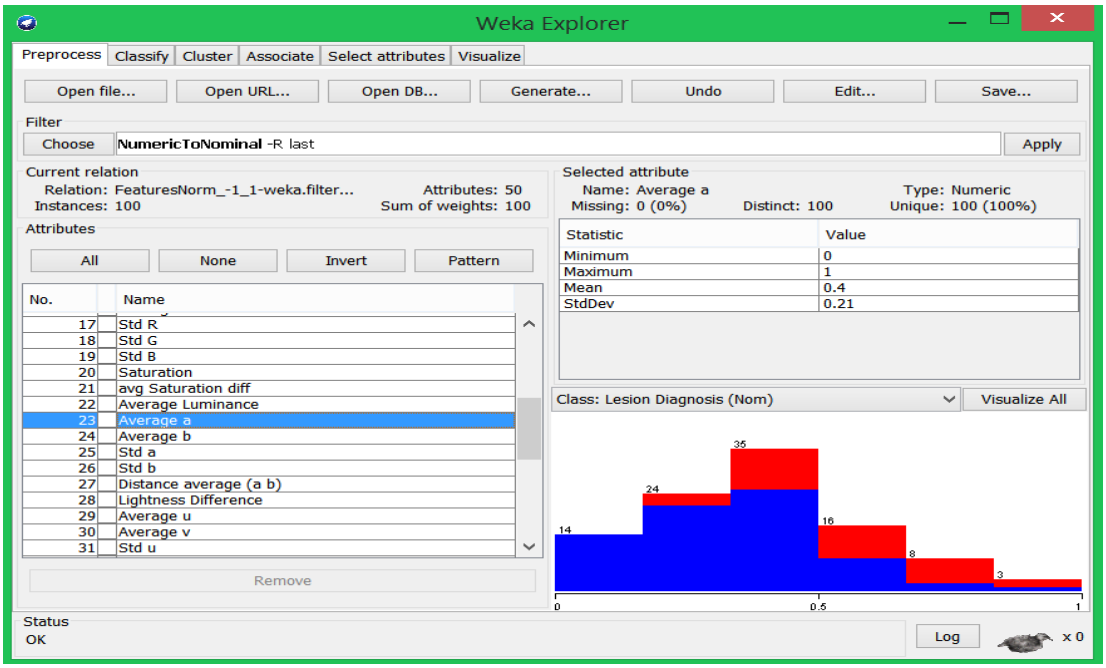


Figure 3.3 - WEKA graphical user interface – data pre-processing tab (for data general statistics visualization and transformation). In the example, the lower right corner, the distribution of the average a* (from the L*a*b color space) attribute is shown (blue – benign lesions; red-melanoma)

3.3. Overview of the Adopted Approach

The goal of this project was to identify combinations of relevant features that lead to an accurate detection of malignant melanomas. In order to do this, it was focused solely on the steps of feature extraction, selection and classification of a typical CAD system. The scheme shown in *Fig. 3.4* exemplifies the steps adopted for this task. The scheme is merely representative and shows the implementation for three input images.

The first step was to separate in each dermoscopic image the lesion from the surrounding healthy skin, in order to allow extracting features related to that region alone. This was done using the results of manual segmentation by an experienced dermatologist, and is further discussed in *section 3.4*. This was followed by the step of feature extraction (see *section 3.5)*, in which features of just the lesion region and also the healthy skin were extracted for every available image. Although expected to be helpful in differentiating the lesions, the selected features were then submitted to feature selection algorithms, in order to reduce the presence of irrelevant features and allow maximizing the classification efficiency. In this work both filter (disregarding the performance of a machine learning algorithm) and wrapper (features selected based on the performance of a learning scheme) selection methods were experimented. In *Fig. 3.4* the example of applying a filter method can be seen, where features 2 and 4 are discarded for not providing significant differentiation between the three examples before using a classification algorithm. The last step performed was evaluating the performance of classification algorithms using the most effective subsets of the initially extracted features, allowing to infer the relevance of each to the classification of melanoma. The performance of the experimented classification algorithms was assessed by comparing their output to the *a priori* knowledge of the diagnostic of each lesion.

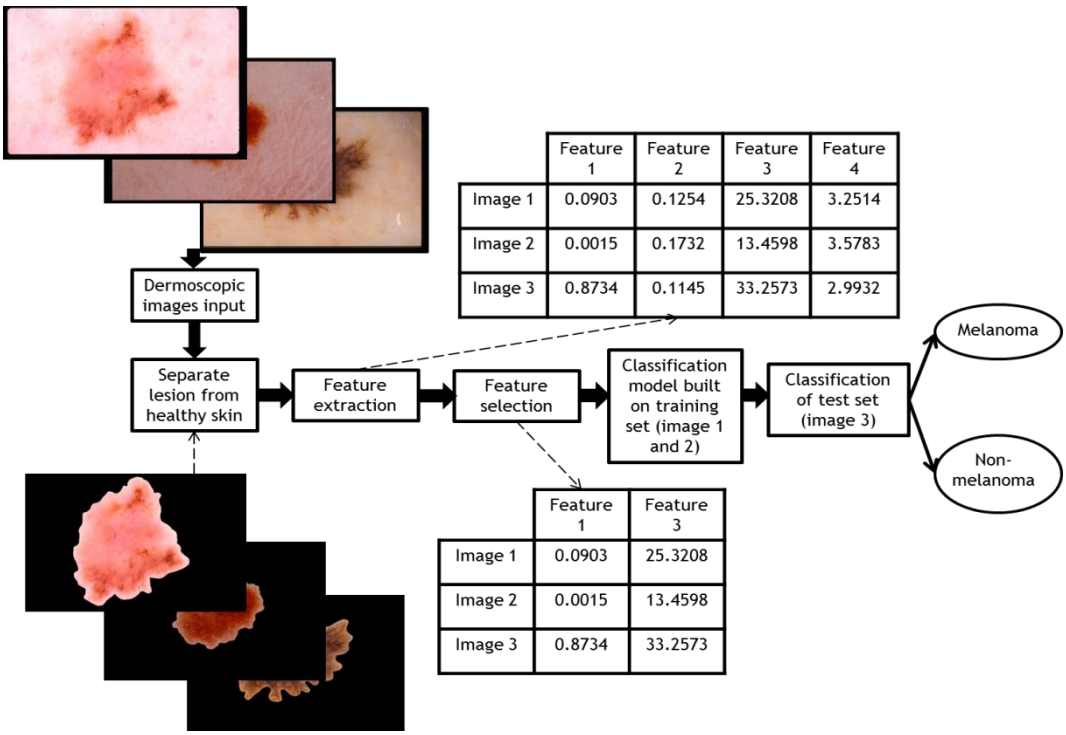


Figure 3.4 - Schematic overview exemplifying most of the steps implemented in the project, from the input of the pigmented skin lesion image, to the decision output of a classifier algorithm.

3.4. Masking of the lesion images

Masking the lesion images consists of separating the region of interest, the pigmented skin lesion, from the surrounding healthy skin. This is made possible through the results of an adequate segmentation, either manual or automatic. To avoid errors that could be introduced by the use of automatically detected borders, the results of manual segmentation of the lesions by experienced dermatologists were used. This allows separating the problem of feature extraction from the problem of automated border detection, hence producing more informative results concerning the objective proposed. Although, as a fully automatic system is desired for it to be better accepted in clinical practice, it would be important to compare the results obtained using manual segmentation to the results that could be obtained following automatic segmentation.

Figure 3.5 - Masking of a pigmented skin lesion image: a) binary mask; b) original RGB image; c) result of the masking procedure.

The manual segmentation result for each image was accessible as a binary mask (see *Fig. 3.5 a))*, where the pixels belonging to the lesion were set to 1, and the remaining to 0. Since this mask has the same size as the corresponding RGB lesion image (*Fig. 3.5 b)*), multiplying each pixel value from the RGB image by its matching pixel in the binary mask, results in 0 for background (skin) pixels, and the original value for pixels inside the lesion (*Fig. 3.5 c)*).

3.5. Feature Extraction

In this section, the features extracted from the pigmented skin lesions, methods used for obtaining them, and the reasons behind their selection will be presented. From studying the visual aspects of the lesions in the datasets used, as well as the reviewed literature, it was expected that a combination of shape, color and texture features should perform better in the detection of malignant lesions than just using a single category. Therefore, a significant group of descriptors of each category was extracted from the lesions and used in a lesion classification routine. The performance of classification achieved when using the three categories was also compared to the performance results using descriptors of only one group (described further in *section 3.6*).

The purpose of the feature extraction step is to transform the large information present in a lesion image into individual numeric attributes that quantify important aspects of it that are expected to be highly correlated to the *a priori* known diagnostic. Hence, each image is transformed into an $(n+1)$-dimensional vector, where n is the number of extracted attributes, to which the *a priori* known diagnostic (label $\in$ {-1,1}, -1 corresponding to benign lesion, 1 to melanoma) of the respective image is added. The resulting feature vectors of m

images are then combined to form a $m \times (n+1)$ matrix, which is used to assess the performance of classification algorithms. The features extracted should be (Bishop, 2006):

- *Robust*, meaning they should be invariant to translation, orientation, scale and illumination, providing good differentiation even when facing noise and artefacts;
- *Discriminating,* the range of values for objects in different classes should be different and preferably be well separated and non-overlapping;
- *Reliable,* all objects of the same class should have similar values;
- *Independent* or uncorrelated from each other, so that no redundant information is added.

This section is divided into three subsections, similarly to the division made in the previous chapter (*section 3.3*), where each of the descriptors extracted from the respective category will be described.

3.5.1. Shape Features

As already mentioned, one important aspect dermatologists take into consideration when assessing the malignancy of a pigmented skin lesion is its shape. Alarming shape factors that can indicate a lesion to be suspicious include larger size than typical benign lesions, irregular borders with abrupt cut-offs, asymmetric shape, and unorganized overall form. Some of the descriptors commonly used have already been reviewed in the previous chapter. Since no color information is required for the extraction of these features, only the binary mask image of each lesion was used for this step, in which the nonzero pixels (white) belong to the lesion and the rest to the surrounding skin. The purpose of extracting shape features is to account for the *A* (asymmetry) and *B* (border) parameters of the ABCD rule of dermoscopy. In this work, a total of twelve shape related features was selected, including seven moment invariants, four simple shape descriptors, and one asymmetry measure:

- *Hu's seven moment-based invariants*

The moment-based invariants have been first proposed by Ming-Kuei Hu (Hu, 1962) in 1962 under the framework of the theory of algebraic invariants, and since then have been target of much attention by the pattern recognition community, especially for object recognition applications. Its basic idea is to describe objects by a set of features which are not sensitive to translation, scaling, rotation and which provide enough discriination power to distinguish among objects from different classes. They are derived from the raw image geometric moments, $m_{p,q}(I)$, of a digital image I, as summarized below.

Considering a digital image $I(x,y)$, with discrete coordinates, where I is the assigned pixel value at the position (x,y), either 0 or 1 when working with the binary image, the raw image geometric moment of order $p+q$, $m_{p,q}$, is calculated by *equation 3.1*.

$$\boldsymbol{m_{p,q} = \sum_x \sum_y x^p y^q I(x,y)} \tag{3.1}$$

These raw moments can be used to derive simple image properties, such as area, $m_{0,0}$, and the coordinates of the shape centroid, x_c and y_c, through $(x_c, y_c) = (m_{1,0}/m_{0,0}, m_{0,1}/m_{0,0})$. Raw moments, however, are not translation invariant, reason why usually the central

moments are considered instead, $\mu_{p,q}$, in which a correction over the object's centroid is used, as follows:

$$\mu_{p,q} = \sum_x \sum_y (x - x_c)^p (y - y_c)^q I(x,y) \tag{3.2}$$

which are translation invariant. In order to achieve scaling invariance, which permits coping with the different sizes that the lesions may present due to different distances at which the dermatoscope was used, these moments can be normalized (see *equation 3.3*).

$$\eta_{p,q} = \frac{\mu_{p,q}}{\mu_{0,0}^{\left(1+\frac{p+q}{2}\right)}} \tag{3.3}$$

Since the purpose is to group similar objects, or lesions in this context, and they can be arbitrarily oriented in an image, it is important that the measures used are not only invariant to translation and scaling, but also rotation. The set of seven Hu's rotation invariant moments, from H_1 to H_7, were calculated using the normalized image moments (*equation 3.3)* following the equations shown below (Hu, 1962):

$$H_1 = \eta_{2,0} + \eta_{0,2} \tag{3.4}$$

$$H_2 = (\eta_{2,0} + \eta_{0,2})^2 + 4\eta_{1,1}^2$$

$$H_3 = (\eta_{3,0} + 3\eta_{1,2})^2 + (3\eta_{2,1} - \eta_{0,3})^2$$

$$H_4 = (\eta_{3,0} + \eta_{1,2})^2 + (\eta_{2,1} + \eta_{0,3})^2$$

$$H_5 = (\eta_{3,0} - 3\eta_{1,2})(\eta_{3,0} + \eta_{1,2})[(\eta_{3,0} + \eta_{1,2})^2 - 3(\eta_{2,1} - \eta_{0,3})^2] + (3\eta_{2,1} - \eta_{0,3})(\eta_{2,1} + \eta_{0,3})[3(\eta_{3,0} + \eta_{1,2})^2 - (\eta_{2,1} + \eta_{0,3})^2]$$

$$H_6 = (\eta_{2,0} + \eta_{0,2})[(\eta_{3,0} + \eta_{1,2})^2 - (\eta_{2,1} + \eta_{0,3})^2] + 4\eta_{1,1}(\eta_{3,0} + \eta_{1,2})(\eta_{2,1} + \eta_{0,3})$$

$$H_7 = (3\eta_{2,1} - \eta_{0,3})(\eta_{0,3} + \eta_{1,2})[(\eta_{0,3} + \eta_{1,2})^2 - 3(\eta_{2,1} + \eta_{0,3})^2] - (\eta_{3,0} - 3\eta_{1,2}) \times (\eta_{2,1} + \eta_{0,3})[3(\eta_{3,0} + \eta_{1,2})^2 - (\eta_{2,1} + \eta_{0,3})^2]$$

These features were selected because they have not been extensively studied before in the classification pigmented skin lesions, except by Barata et al. (Ruela, Barata, Mendonca, & Marques, 2013), in a work where the role of shape descriptors in the detection of melanoma was studied and these moment invariants have proven to be useful. It was expected that these features could capture the existing similarities between the irregularly shaped melanomas and between the mostly regularly shaped benign lesions.

- *Shape Compactness*

The shape compactness (Ruiz et al., 2011) is a simple shape descriptor widely used in many dermoscopic images classification problems. It is used as a measure of the border irregularity of lesions. It is independent of scale and orientation, and measures the relation between the objects shape to a circle with the same perimeter, which is the most compact geometric shape. This measure varies from 0 to 1, the latter corresponding to the perfect circle. Therefore, as melanomas tend to present very irregular shapes, and most common nevi show a perfectly symmetrical and often circular shape, this was thought to be a good indicator for malignancy of a lesion, and should be included in the set of shape descriptors.

The shape compactness index, C, was calculated as follows:

$$C = \frac{\text{Perimeter}^2}{4\pi\text{Area}} \tag{3.5}$$

In order to calculate this, one MATLAB image processing toolbox function was used, the *regionprops* function (Mathworks). The latter receives a binary image as input, and returns measurements for the selected properties from each object present in the image. The available properties to measure with this function include area, perimeter, object orientation, the bounding box (smallest rectangle containing the object region), the object solidity, etc. The binary masks for the dermoscopic images contained only a single region, and hence the perimeter and area were easily determined by this method.

- *Lengthening Index*

The lengthening index (Messadi et al., 2009) is used to measure the aspect ratio of a lesion, calculated by the ratio of the length of its major axis to the length of the minor axis. It describes the anisotropy degree of a lesion.

The principal axes of a given shape can be defined as the two segments of lines that cross each other orthogonally in the centroid of the object and represent the directions with zero cross-correlation (Peura & Iivarinen, 1997). Their lengths correspond to the eigenvalues, λ_1 and λ_2, of the covariance matrix, CC, of the object's contour (*equation 3.6*), and are calculated with *equations 3.7* and *3.8*.

$$CC = \frac{1}{N}\sum_{i=1}^{N}\begin{bmatrix} x_i - x_c \\ y_i - y_c \end{bmatrix} \times \begin{bmatrix} x_i - x_c \\ y_i - y_c \end{bmatrix}^T = \begin{bmatrix} \mu_{2,0} & \mu_{1,1} \\ \mu_{1,1} & \mu_{0,2} \end{bmatrix} \tag{3.6}$$

Where N is the number of points in the contour, (x_i, y_i) the coordinates of the ith contour point, (x_c, y_c) the coordinates of the object's centroid, and $\mu_{p,q}$ the $p+q$ geometric central moment, as calculated by *equation 3.2*.

$$\det(CC - \lambda_{1,2} I) = \det\begin{pmatrix} \mu_{2,0} - \lambda_{1,2} & \mu_{1,1} \\ \mu_{1,1} & \mu_{0,2} - \lambda_{1,2} \end{pmatrix} = (\mu_{2,0} - \lambda_{1,2})(\mu_{0,2} - \lambda_{1,2}) - {\mu_{1,1}}^2 = 0 \tag{3.7}$$

From the previous, the eigenvalues, λ_1 and λ_2, are determined as follows:

$$\begin{cases} \lambda_1 = \frac{1}{2}[\mu_{2,0}\mu_{0,2} + \sqrt{(\mu_{2,0} + \mu_{0,2})^2 - 4(\mu_{2,0}\mu_{0,2} - {\mu_{1,1}}^2)}] \\ \lambda_2 = \frac{1}{2}[\mu_{2,0}\mu_{0,2} - \sqrt{(\mu_{2,0} + \mu_{0,2})^2 - 4(\mu_{2,0}\mu_{0,2} - {\mu_{1,1}}^2)}] \end{cases} \tag{3.8}$$

Corresponding to the length of the major and minor axis of the object, respectively. The lengthening index was then calculated as the ratio between the two measures, λ_1/λ_2. A circle has a lengthening index of 1, because its major and minor axes have the same length, and the more elongated an object is, the higher its lengthening index. Melanomas, which often present ellipsoidal shapes, should have a lengthening index higher than the benign lesions, often circular in shape.

- *Solidity*

Solidity can also be used to measure the border irregularity of an object, similarly to the shape compactness measure, and is defined as the ratio of the object's area to its convex hull. The convex hull of an object is the smallest convex region that contains the whole object region. It describes the extent to which an object is convex or concave and was calculated using the *regionprops* (Mathworks, 2015b) function of Matlab.

In *Fig. 3.6* an example of the convex hull, in green, of a benign lesion (*3.6 a)*) and a melanoma (*3.6 b)*), and the resulting solidity is presented. It can be seen that the benign lesion presents smooth borders an almost fully convex shape. On the other hand, the melanoma has very irregular borders and therefore a lower ratio between the two areas.

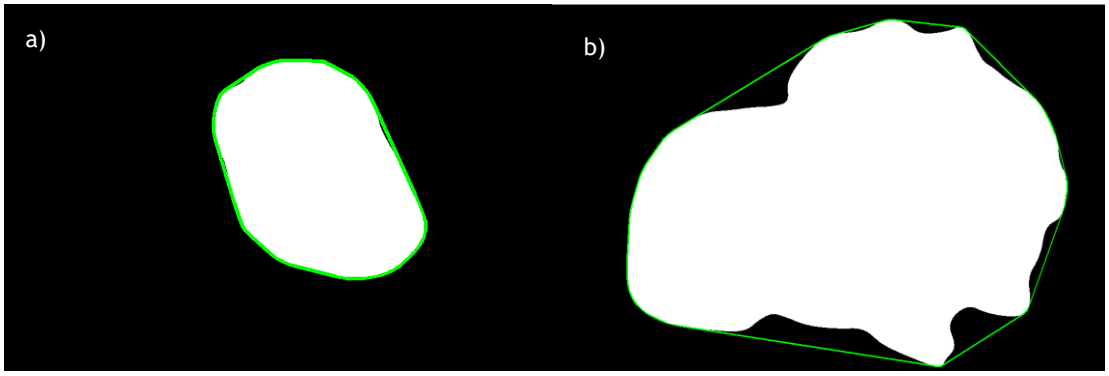


Figure 3.6 - Convex hull of a: a) benign lesion (solidity: 0.9859); b) melanoma (solidity: 0.9207).

- *Rectangularity*

The rectangularity represents how rectangular an object is, by measuring how much it fills its minimum bounding rectangle. It is calculated as the ratio of the area of the object region to the area of its minimum bounding rectangle. In this work, it was also calculated through the *regionprops* function, which returns the rectangularity value by using the property "Extent" as the input argument of the function together with the binary image, and also the parameters of the minimum rectangular bounding box, namely the coordinates of the upper left corner, width and height of the bounding box, through which it can be drawn (see *Fig. 3.7*). Overall melanomas are expected to present lower values of rectangularity than benign lesions, possibly providing a good parameter of differentiation.

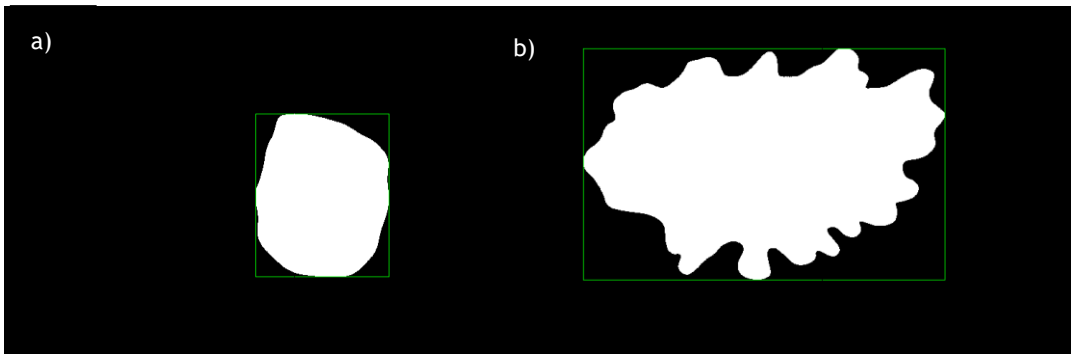


Figure 3.7 - Rectangularity index of a: a) benign lesion (rectangularity: 0.8167); b) melanoma (rectangularity: 0.6486)

- *Asymmetry Index*

The asymmetry index has the goal of quantifying how much asymmetric is a given shape, and is an important sign of a malignant lesion that should be taken into account in the visual inspection of a lesion. However, to translate this measure to a numerical quantity is difficult, and from the reviewed literature it was understood that there is not an universal measure to do this. In this work a method different from what was found in the literature was investigated, as described in the following paragraphs.

The first step was to determine the geometric center of the lesions' contour. The geometric center coordinates, (x_{gc}, y_{gc}), of an object can be calculated as the average of the contour points' coordinates, x_i and y_i, as presented in *equation 3.9*, where N is the number of countour points.

$$(x_{gc}, y_{gc}) \rightarrow x_{gc} = \frac{1}{N}\sum_{i=1}^{N} x_i \ , y_{gc} = \sum_{i=1}^{N} y_i \tag{3.9}$$

To find the contour points a function available in the image processing toolbox of Matlab was used, the *bwboundaries* [70]. This function receives a binary image as input, and returns the list of countour points' coordinates for each object present in the image as a $N \times 2$ matrix. Additional to the binary image, a 4 or 8 pixel connectivity needs to be specified to define how many neighbor pixels are considered when tracing the object boundary.

The next step was to select a starting random point from the contour, and use it to create an axis of symmetry, with the equation $y = mx + b$, where $m_{symAxis}$ is the slope of the line segment calculated by $m_{symAxis} = (y_{contour} - y_{gc})/(x_{contour} - x_{gc})$, and b, its intersection to the cartesian y axis, defined by $b_{symAxis} = y_{contour} - (m_{symAxis} \times x_{contour})$. This axis contains that contour point and the geometric center of the lesion, and should intersect the lesion one more time in the opposite side of the lesion. This intersection was calculated using a function available at *Mathworks file exchange* implemented by Douglas Schwarz (Schwarz), which allows an efficient computation of curve intersections.

After defining the symmetry axis (see the pink line in *Fig. 3.8 a)*) to be used, for each contour point between the initial point to the intersection of the defined axis with the contour on the opposite side, a line segment perpendicular to the symmetry axis, with slope $m_{perp}(i) = -1/m_{symAxis}$, that includes the ith contour point analysed is defined, having $b_{perp}(i) = y_c(i) - m_{perp} \times x_c(i)$. Using this line, it is simple to define the reflection pixel to the one being analysed over the established symmetry axis. First the $x_{symAxis}(i)$ and $y_{symAxis}(i)$ coordinates of the intersection of the symmetry axis to the perpendicular line are obtained using *equation 3.10,* and then the symmetric pixel's coordinates are determined, $x_{sym}(i)$ and $y_{sym}(i)$, using *equation 3.11.*

$$x_{symAxis}(i) = \frac{(b_{symAxis} - b_{perp}(i))}{(m_{perp}(i) - m_{symAxis})} \tag{3.10}$$

$$y_{symAxis}(i) = m_{symAxis} * x_{symAxis}(i) + b_{symAxis}$$

$$x_{sym}(i) = 2 \times x_{symAxis}(i) - x_c(i) \quad y_{sym}(i) = 2 \times y_{symAxis}(i) - x_c(i) \tag{3.11}$$

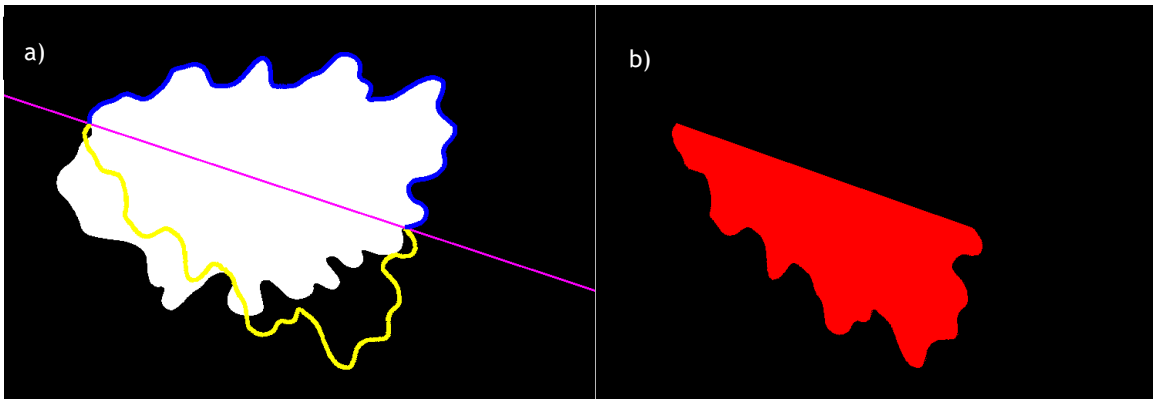


Figure 3.8 - Steps for determining the Asymmetry Index of a lesion: a) construction of the symmetric contour (in yellow) of the original contour (in blue) over an axis of symmetry (in pink); b) filling of the symmetric region (in red).

The resulting group of points $(x_{sym}(:), y_{sym}(:))$ are the contour points of the symmetric half obtained over an initially defined symmetry axis (the yellow line in *Fig. 3.8 a)*). The last step for obtaining the asymmetry measure is to fill the region enclosed by the symmetry axis and the symmetric contour (see the red region *Fig. 3.8 b)*), and divide the area of the non-matching region (the combined area of the blue and red regions in *Fig. 3.9*) by the area of the original region on the same side of the symmetry axis (the area of the green plus the blue region in *Fig. 3.9*).

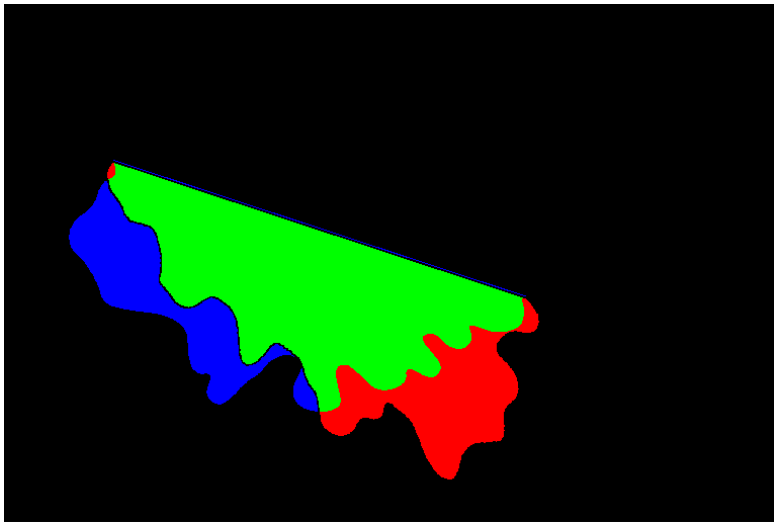


Figure 3.9 - Overlapping of the symmetric region over the original region (in green) by the symmetry axis (shown in *Fig. 3.8 a)*). In blue, the non-overlapping original lesion area can be seen, and in red the non-overlapping symmetric region.

For each lesion, the initial symmetry axis was iteratively rotated clockwise by 10º until reaching 180º of rotation, and the previously described process was repeated. In the end, the minimum value found for the 18 indexes calculated was used as the asymmetry index of the lesion. For benign lesions, this value was expected to be very small, as the diference between the overlapping areas should be minimal most of the times, and larger for melanomas, due to the expected asymmetry over all the considered axes.

It should be noted that this measure was the only based solely on the border pixels of the lesions. Therefore it was the most deteriorated when working with the PH^2 image dataset, because the image border was used as the contour, which sometimes lead to meaningless results for this feature in those images.

3.5.2. Color Features

Color assessment is essential in the detection of malignant melanoma, and most of the dermoscopy scoring systems include color as a diagnostic criterion. The most important color factors for considering a lesion as suspicious are the presence of multiple colors across the lesion, and uneven distribution of color. As previously referred, the color exhibited by a nevus is dependent on the localization of melanin in the skin. Since melanomas origin from abnormal growths of melanocytes, the melanin agglomeration origins from different skin layers, therefore leading to the formation of a nevus presenting various colors, while the benign lesions often present an even distribution of single colors. Since it consists of a specific trait of the malignant melanomas, most of the proposed CAD systems for dermatology have included color information in their lesion analysis.

In the proposed system, the masked lesion images were used for extracting color features from the whole lesion region. Additionally, the relation of some color parameters inside the legion region with the same from the surrounding skin was also considered. This relation was included because it is expected that the smoothness of transition between the lesion to the surrounding skin may carry important diagnostic information. Three color spaces were considered for this evaluation: RGB, CIE $L^*a^*b^*$ and CIE $L^*u^*v^*$, and a total of 19 features were measured. A color space is a method by which we can specify, create and visualise color. These spaces are often three-dimensional, hence defining a color through a set of three parameters. The way these coordinates relate to each other and the color each triplet characterises is what is specific to each color space.

The most common color space used to store color information is the RGB, an additive color system based on tri-chromatic theory, where every color is defined by a level of red, green and blue components. However it has the disadvantage of not separating the chromaticity (which conveys the color information) from the luminance (or lightness). The same color level (same red, green and blue components) under different illumination conditions is perceived as a different color, which can lead to inaccurate results in automatic detection systems that compare images acquired under different illumination settings based only on color from this space.

The CIE has defined a system that classifies a color according to the human visual system, and allows to use the color information more effectively. The CIE color standard is based on imaginary primary colors X, Y and Z, also called the *tristimulus values* (Ford & Roberts, 1998) which do not exist physically. They are virtual primary colors that have been devised so that all colors which can be perceived by the human eye lie within this color space. The XYZ system is based on the response curves of the three color receptors of the human eye's, and its values are determined from the RGB values by color matching functions, for which the values are dependent of the color system of the output device. This color space is often converted to the CIE chromaticity diagram, by projecting the three-dimensional XYZ space to the $X + Y + Z = 1$ plane, in which the coordinates are usually called xy, and are derived using *equation 3.12*. They are called the chromaticity coordinates, and always add up to one, meaning that z can always be expressed in terms of x and y, and hence only x and y are required to specify any color. However, since this is a projection of the 3D space, each point in xy corresponds to many points in the original space. The missing information is the luminance component of color Y, so a color can be described by its xyY coordinates.

$$x = \frac{X}{(X+Y+Z)},\ y = \frac{Y}{(X+Y+Z)},\ z = \frac{Z}{(X+Y+Z)} \qquad (3.12)$$

To determine the exact color matching functions, the chromaticity coordinates of the white point in the color system of the output device must be known. With them, the Y component can be determined and the tristimulus values can be calculated (Hunt & Pointer, 2011). The CIE $L^*a^*b^*$ and $L^*u^*v^*$ color spaces are based directly on the CIE XYZ in an attempt to linearize the perceptibility of unit vector color differences, and are known as uniform color models. This coloring information is referrenced to the lightness of the white point (Y_0) of the output system and its components are derived using equations *3.13* and *3.14*. To perform the conversion between colorspaces in Matlab, a function available in the *Mathworks fileexchange repository*, *ColorSpace*, developed by Pascal Getreuer (Getreuer, 2010) was used. It uses a string as input naming the desired conversion (example *'sRGB->Lab'*), and outputs the standard RGB image transformed to the defined output color coordinates. In these color spaces, the lightness component is present in channel *L* (ranging

from 0 to 100) and is separated from the color expression. In $L^*u^*v^*$, the (u^*, v^*) coordinates correspond to the chromaticity coordinates, representing the position of the color in the uniform chromaticity scale. In $L^*a^*b^*$, the a^* coordinate represents the position of the color between magenta (positive direction) and green (negative direction); while the b^* coordinate the position of the color between yellow and blue (Z. Ma & J. Tavares, 2015). The main advantages of using these color spaces are that they separate the chromaticity from luminance, thus allowing to compensate for the uneven illumination that is present between images obtained under different conditions, and also the fact they they are perceptually uniform, meaning that long (short) distances within the color space should correspond to large (small) perceived distances between colors, which permits a more consistent matching of color information between lesions.

$$\begin{cases} L^* = 116\left({}^{Y}/_{Y_0}\right)^{\frac{1}{3}} - 16 & if\ \frac{Y}{Y_0} > 0.008856 \\ L^* = 903.3\ \left({}^{Y}/_{Y_0}\right) & if\ \frac{Y}{Y_0} \leq 0.008856 \\ a^* = 500\left[f\left({}^{X}/_{X_0}\right) - f\left({}^{Y}/_{Y_0}\right)\right] & \\ b^* = 200\left[f\left({}^{Y}/_{Y_0}\right) - f\left({}^{X}/_{X_0}\right)\right] & \end{cases} \tag{3.13}$$

$$\begin{cases} L^* = 116\left({}^{Y}/_{Y_0}\right)^{\frac{1}{3}} - 16 & if\ \frac{Y}{Y_0} > 0.008856 \\ L^* = 903.3\ \left({}^{Y}/_{Y_0}\right) & if\ \frac{Y}{Y_0} \leq 0.008856 \\ u^* = 13L^*[U(X,Y,Z) - U(X_0,Y_0,Z_0)] & \\ v^* = 13L^*[V(X,Y,Z) - V(X_0,Y_0,Z_0)] & \end{cases} \tag{3.14}$$

with

$$\begin{cases} f(U) = U^{\frac{1}{3}} & if\ U > 0.008856 \\ f(U) = 7.787U + {}^{16}/_{116} & if\ U \leq 0.008856 \end{cases}$$

and

$$U(X,Y,Z) = \frac{4X}{X+15Y+3Z},\ V(X,Y,Z) = \frac{9Y}{X+15Y+3Z}$$

The features extracted from the color spaces considered, namely the *RGB*, the $L^*a^*b^*$ *and* $L^*u^*v^*$ are summarized below:

- *Average and Standard deviation in R, G and B channel*

For each lesion, the average and standard deviation of the red, green and blue channel was determined. This measures are used to quantify the color variegation in each of these channels. To calculate the average, μ_i, and standard deviation, σ_i, of each color channel, i, in a lesion image, Img, each pixel belonging to the lesion (where the binary mask has value 1) is analyzed, and *equations 3.15* and *3.16* are applied.

$$For\ i = 1{:}3\ \rightarrow\ \mu_i = \frac{1}{N}\sum_{k=1}^{N} Img(r(k), c(k), i) \tag{3.15}$$

Where N is the total number of pixels in the lesion, r and c are the row and column of the kth pixel belonging to the lesion, and $Img(r(k), c(k), i)$ is the value of pixel k from image Img in channel i (1 – Red; 2 – Green; 3 – Blue).

$$For\ i = 1{:}3\ \rightarrow\ \sigma_i = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(Img(r(k), c(k), i) - \mu_i)^2} \quad (3.16)$$

- *Average Lesion Saturation*

To describe saturation, it is important to define colorfulness first. Colorfulness is the visual sensation according to which the perceived color of an area appears to be more or less chromatic (Hunt & Pointer, 2011). Saturation, based on the previous, can be defined as the colorfulness of a color relative to its own brightness, meaning that the saturation of a color is determined by a combination of light intensity and how much it is distributed across the spectrum of different wavelengths (Hunt & Pointer, 2011). It is an intuitive concept based on the human's perception of color. A saturated color corresponds to a color created by a narrow band of wavelengths at high intensity. This concept can be adopted to the analysis of pigmented skin lesions, because the presence of multiple colors inside the lesion should lead to an overall less saturated color than benign lesions that often exhibit single colors. The color saturation, S, of each pixel k belonging to the lesion was calculated using *equation 3.17,* by dividing the highest value of the three color channels by the sum of their values. The average saturation inside the lesion was determined using *equation 3.15*, with the exception that since the calculation uses the information from the 3 channels, it was not repeated 3 times to cycle through the channels, and instead of the pixel value for each channel, the pixel saturation is used.

$$S(k) = \frac{\max[Img(r(k),c(k),1),Img(r(k),c(k),2),Img(r(k),c(k),3)]}{Img(r(k),c(k),1)+Img(r(k),c(k),2)+Img(r(k),c(k),3)} \quad (3.17)$$

- *Difference between the Average Lesion Saturation and Average Skin Saturation*

For the saturation parameter, the relation between the lesion and the surrounding skin was considered. It was determined by the absolute difference between the average saturation of the skin lesion and the average saturation of the neighboring surrounding skin. This feature was included to verify if the presence of a malignant melanoma may affect the color properties of surrounding healthy skin.

In order to determine the average saturation of the surrounding skin, a band of 8 pixels width was considered all around the pigmented lesion. To define this band, for each pixel in the mask image that belongs to the background, p_{skin} (where the mask image is valued 0), the minimum distance between it and every contour point, $p_{contour}$, is determined (the distance, $dist$, between two pixels can be calculated using *equation 3.18*), and if it is less than 8 pixels, the pixel index is marked as belonging to the desired region. The resulting band can be seen in *Fig. 3.10* in green. The average saturation across this region was calculated as described previously, and the value of the absolute difference between its value and the average lesion saturation was used as a feature.

$$dist(p_{skin}, p_{contour}) = \sqrt{(x_{skin} - x_{contour})^2 + (y_{skin} - y_{contour})^2} \quad (3.18)$$

where (x_{skin}, y_{skin}) are the 2D coordinates of the background pixel, and $(x_{contour}, y_{contour})$ the coordinates of each contour pixel.

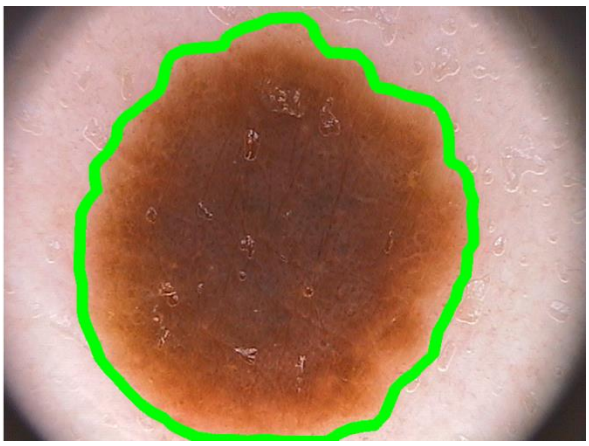


Figure 3.10 - The area considered for computing the features of the skin surrounding the lesion.

- *Average Lesion Lightness*

Similarly to color saturation, the lightness of a color also shares close relation to a broader definition, brightness. The brightness is the human sensation by which an area exhibits more or less light, whereas the lightness is often referred to as the brightness of an area judged relative to a reference white or highly transmitting area in the scene (Hunt & Pointer, 2011). This parameter is determined in $L^*a^*b^*$ and $L^*u^*v^*$ and its average from the lesion region was included in the samples' feature vector. To determine it, the value of the first channel of the $L^*a^*b^*$ coordinates for every pixel belonging the the lesion region was used in *equation 3.15*.

- *Difference between the Average Lesion Lightness and Average Skin Lightness*

The method applied to determine the absolute difference between the average lesion and skin saturation was applied with the lightness parameter. The average lightness from the pixels surrounding the skin lesion in a band of 8 pixels width (green region in *Fig. 3.10*) was calculated, and the absolute difference between it and the measured average lightness inside the lesion was determined. The goal was to capture the effect the lesions have on surrounding tissue, to further study how this can help differentiating between them.

- *Average and Standard deviation in* a^*, b^*, u^* *and* v^* *coordinates*

In order to quantify the variegation of color within the lesions, the colorimetric and chromaticity coordinates from the $L^*a^*b^*$ and $L^*u^*v^*$ color spaces were also considered. Following the transformation of the original RGB images into the CIE's color spaces, the method used to calculate the average and standard deviation of the R,G and B value was repeated (*equations 3.15* and *3.16*), this time using only the second and third channel of each pixel. This was expected to capture the differences in absolute color existing between the lesions acquired under uneven illumination conditions that could otherwise go undetected.

- *Distance between average (*$\mathrm{a}^*, \mathrm{b}^*$*) and (*$\mathrm{u}^*, \mathrm{v}^*$*) from the lesion and surrounding lesion*

The last color descriptors considered were the relation between the colorimetric coordinates, *(*$\mathrm{a}^*, \mathrm{b}^*$*)* and chromaticity coordinates *(*$\mathrm{u}^*, \mathrm{v}^*$*),* found for the considered lesion and for its surrounding healthy skin in the same band considered for the previous relations. The averaged 2D coordinates characterize the average absolute color of a region. These coordinates were calculated for the surrounding skin, and their relation was determined in

terms of the distance between the two points in the 2D plane. This was done using *equation 3.18*, where instead of (x_{skin}, y_{skin}) and $(x_{contour}, y_{contour})$, the coordinates used were the average of (a^*, b^*) and (u^*, v^*), namely $\mathrm{dist}((a_{skin}{}^*, b_{skin}{}^*), (a_{lesion}{}^*, b_{lesion}{}^*))$ and $\mathrm{dist}((u_{skin}{}^*, v_{skin}{}^*), (u_{lesion}{}^*, v_{lesion}{}^*))$, respectively. Due to deterioration melanomas may cause in the surrounding skin, this relation was expected to highlight differences between the benign and malignant lesions that can help distinguishing them in the automatic classification.

3.5.3. Texture Features

The definition of texture is closely related to the human sense of touch, and it is how he defines the surface of a certain object, for example as being rough or soft, smooth, corrugated, granulated, dense, uniform etc. In a 2D image, the texture of the objects result in varying brightness and color properties that can be identified in small patches and related to its physical properties. The ability to numerically describe the textures properties found in a digital image can provide significant information about an object, which makes it a widely used technique in the field of computer vision.

Many texture analysis methods have been proposed, and they are often categorised into four different groups: structural, statistical, model-based and signal processing methods (Materka & Strzelecki, 1998). The structural approaches attempt to describe an image texture by defining a set of microtextures in a hierarchy of spatial arrangements that categorise the overall macrotexture, which provides a good symbolic description of an image, but often struggles to deal with natural textures because of the variability of both micro and macro-structures present. The statistical methods represent texture by measuring the distributions and relationships between pixels in a gray level image, and computing neighbor pixel statistics. Model-based texture analysis represents an image using fractal or stochastic models, for which the parameters are estimated and then used for image analysis. Signal processing approaches uses a bank of filters to represent an image in a space whose co-ordinate system has an interpretation that is closely related to the characteristics of a texture. For the latter, Fourier transform, Gabor filters and wavelet transforms are the most commonly used.

In the context of dermoscopy images, the analysis of textures is used in an attempt to detect anatomical structures that dermatologists consider for the indication of malignancy in a lesion, like atypical pigment network, irregular vascularization, structureless areas, branched streaks, dots and globules. These factors account for the *D*, dermoscopic structures, in the ABCD rule of dermoscopy, and their presence is expected to generate characteristic texture maps that should help in differentiating between the skin lesions. For this work, second order GLCM based statistical texture descriptors were employed.

In order to compute the GLCM from an image, it should first be converted to grayscale. Transforming an RGB image to grayscale is done by assigning to each pixel its gray value, G, the value of a weighted sum of its three components (see *equation 3.19*) and corresponds to eliminating the hue and saturation information while retaining the luminance. To perform the transformation, the Matlab function *rgb2gray* (Mathworks, 2015c) was used.

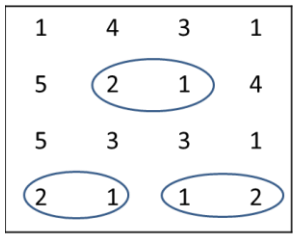
$$G = 0.2989 \times R + 0.5870 \times G + 0.1140 \times B \tag{3.19}$$

The GLCM was proposed by Haralick et al. in 1973 (Haralick, Shanmugam, & Dinstein, 1973). In this matrix the relative frequencies of gray level pixels or transitions of gray levels

between pixels are stored. The order of the GLCM measures are defined by how many pixels are considered to count the number of combinations occurred. Second order GLCM texture calculations are the most commonly used, and consider the relationship between groups of two pixels in the original image. This matrix can then be used to compute several statistics, which are used as texture descriptors. Matlab has a readily available function to create a GLCM from an input gray-level image, the *graycomatrix* function (Mathworks, 2015a). To do it, four parameters should be defined, namely the limits to apply to the gray values in the image; the number of gray levels, which represent the number of bins the band defined by the gray level limits should be divided into, an offset which specifies the direction and distance to consider between the pixel of interest and its neighbor, and a boolean to define if the ordering of values in the pixel pairs should (if set to false) or should not (if set to true) be considered.

To exemplify the creation of a gray-level co-ocurrence matrix, see *Fig. 3.10*. For each position of the GLCM, the row represents the value of the pixel of interest, and the column the neighbor value considered; $GLCM(row, column)$ value is the number of times the transition from row to column occurs in the image, for a given offset. The matrix on the left represents the input gray-level image. In this case, the parameters mentioned above would be:

- [1 5] for gray level limits, this represents the minimum and maximum gray level that the original image values were scaled to;
- 5, the number of gray level bins into which the previous band is divided into (1, 2, 3, 4, 5);
- [0 1] specifies the offset. This is used as a vector in the 2D image space (0 rows, 1 column in the positive direction) and this case represents a horizontal, 0º, vector with unit length, going from left to right. It specifies which pixel is searched for as neighbor of the pixel of interest. Therefore, 1 is usually referred to as the pixel distance, D, (distance between the considered pixels). To specify other angles of search, the offset should be set to: [-D D], for 45º; [-D 0] for 90º; [-D -D] for 135º, etc;
- Symmetrical set to true. This option defines that the search is made in both ways of a search offset, for example, when using the offset defined for this case, when counting the number of times the value 2 is adjacent to value 1, both $1 \rightarrow 2$ (left to right) and $2 \leftarrow 1$ (right to left) are considered, resulting in a symmetric matrix (position $(1,2)$ and $(2,1)$ of the matrix will present the same count). If symmetric was set to false, these combinations were considered individually.



GLCM	1	2	3	4	5
1	1	3	2	2	0
2	3	0	0	0	1
3	2	0	1	1	1
4	2	0	1	0	0
5	0	1	1	0	0

Figure 3.11 - Example of GLCM computation: the matrix on the left represents a gray level image limited to 5 gray levels between [1 5] used as input; on the right is the resulting GLCM, using a pixel distance of 1 and a horizontal offset ([0 1]). The highlighted pairs of pixels show the transitions between levels 1 and 2 and where they are positioned in the GLCM.

Before calculating the texture measures, this matrix must be normalized, so that each cell contains a probability of occurrence of each transition, $P_{i,j}$, instead of the number of times it occurs, $C_{i,j}$, where i and j represent the rows and columns of the matrix, respectively. To do it, the number of counts of each transition needs to be divided by the total number of transitions (see *equation 3.20*).

$$P_{i,j} = \frac{C_{i,j}}{\sum_{i,j=1}^{N} C_{i,j}} \tag{3.20}$$

Table 3.1 - Summary of the equations used to compute the GLCM texture statistics.

GLCM Feature	Equation
Contrast	$\sum_{i,j=1}^{N} P_{i,j}(i-j)^2$
Dissimilarity	$\sum_{i,j=1}^{N} P_{i,j}\lvert i-j\rvert$
Inverse Diff. Moment	$\sum_{i,j=1}^{N} \frac{P_{i,j}}{1+(i-j)^2}$
Energy	$\sqrt{\sum_{i,j=1}^{N} {P_{i,j}}^2}$
Entropy	$\sum_{i,j=1}^{N} P_{i,j}(-\ln(P_{i,j}))$
Maximum Probability	$Max(P_{i,j})$
Homogeneity	$\sum_{i,j=1}^{N} {P_{i,j}}^2$
Correlation	$\sum_{i,j=1}^{N} P_{i,j}\left[\frac{(i-\mu_i)(j-\mu_j)}{\sqrt{(\sigma_i^2)(\sigma_i^2)}}\right]$ where $\mu_{i,j}$ is the average probability of row i and j, respectively, and $\sigma_{i,j}^2$ the variance

After obtaining the normalized symmetric GLCM, texture descriptors are computed. For this work, eight gray-level shift invariant features were considered (Clausi, 2002): *Contrast*, *Correlation*, *Dissimilarity*, *Energy*, *Entropy*, *Homogeneity*, *Maximum probability*, *Inverse difference moment*. These allow a characterisation of texture that is robust to linear shifts in the illumination intensity, and the equations for obtaining them are summarized in *table 3.1*. The texture measures result from a weighted sum of the transitions probabilities, where the

weights enhance the importance of specific transitions (for example the direction or the size of the gap), resulting in the characterization of specific properties. These are usually grouped with relation to the aspect measured in the following groups: the contrast group, to which *contrast, dissimilarity* and *inverse difference moment* belong, emphasizes the transitions that are away from the diagonal (where the gaps between gray levels are higher); measures related to the orderliness, to which the *energy, entropy, maximum probabilit*y and *homogeneity* belong, define how regular (orderly) the pixel values are within the image, and its weights are constructed related to how many times a given pair occurs; and descriptive statistics, for which only the *correlation* was used, which measures the linear dependency of gray levels on those of neighboring pixels.

In order to compute these statistics for the lesion region only, the background pixels had to be converted to NaN. It was done by assigning the value NaN for every pixel in the mask image that was black (value 0). The GLCM computing function used automatically ignores pairs of pixels where one of them is not assigned any value. For these statistics to be significant, it is important that the GLCM is reasonably dense. When using the full dynamic range (8-bit data), the image has 256 gray levels, which results in a matrix with 65536 cells, and some gray level transitions would be non-existent. This leads to a sparse matrix, for which the approximation for the probability distribution is bad, and the discrimination power of some statistics derived from it are greatly reduced (Clausi, 2002). A lower number of gray levels also improves the computational efficiency and reduces the effects of noise in the image. The number of gray levels used in this work was 50. The pixel distance considered for all measurements was 2. For the features to be rotation invariant, the symmetric GLCM was computed in four directions ({0º, 45º, 90º and 135º}), and the 8 statistics were computed for each of them. The resulting 16 texture features kept for classification were the average and the range of each computed statistic across the four directions.

The methods described in this section were used to describe each image in the database as a feature vector, containing 48 numerical measures corresponding to the referred features. To this vector the actual lesion diagnosis known *a priori* was added so that the performance of classification could be assessed. Gathering all feature vectors together, two feature matrices were created, one concerning to the first database studied, with 100 samples, and the second to the 200 images from PH^2 dataset. Before feeding these for the subsequent steps, the data was normalized (see *equation 3.21*). Some features' ranges were several orders of magnitude larger than others, so normalization was required for every feature to take on a value from 0 to 1. Without this pre-processing step, the latter could introduce errors in the classification models built, as in some classification schemes they would favour the variables with wider ranges, which could lead to losing significant information.

$$X_{i,0\ to\ 1} = \frac{X_i - X_{min}}{X_{max} - X_{min}} \tag{3.21}$$

Where X_i is the value of a certain feature X for the image i, $X_{i,0\ to\ 1}$ is the value for the feature in the same image but on the scale of 0 to 1, X_{min} is the minimum value of the feature X within the samples considered, and X_{max} its maximum value.

3.6. Feature Selection and Classification

This section describes the work done with the previously selected features to assess their relevance in the context of the automatic classification of dermoscopic images. In the first part, methods applied to assess the relevance of features in the context of the datasets used will be presented, followed by the methods applied to find the best performing feature combinations. In the two image databases considered, benign lesions were classified into more accurate sub-categories (namely into atypical nevi and common nevi); but given that binary classification is easier than multiclass classification, in this work the sub-classes were ignored and a skin lesion was classified as either a melanoma or non-melanoma.

Although the features extracted from the images were all expected to convey important information for the distinction between melanoma and benign lesions, it was likely that the use of all features would not lead to the best classification performance. In this work the individual discriminative potential of the features selected, as well as their potential when used in combination were studied. In order to achieve this, some *dimensionality reduction* methods were experimented. The *feature construction* methods, described in *section 2.2.4*, were thought to be inadequate for the addressed problem. Although being capable of improving the classification performance and lowering the computational requirements, the subset of features they create result from combinations of the original, so they carry no physical meaning and are hard to link to the original, which makes the determination of the significant features impossible. These methods are more suited for applications where a set of features is already known to be relevant, but the high dimensionality deteriorates the prediction and generalization capabilities of the classifiers. Therefore, only *feature selection* methods were considered.

With the exception of the *exhaustive search,* described further in this section, the evaluation of the *feature selection* algorithms was solely performed in the WEKA software. Previously to this step, the normalized feature values extracted from the available images were saved in an excel file as a feature matrix, where each row corresponded to an image, and the values in the various columns represented the values of the considered features for it, terminating with its respective label. In order to be compatible with the WEKA software, the excel workbook format had to be converted to comma-separated value (CSV) format, which is capable of holding the values from an excel worksheet in text format, with each line of text corresponding to a row from the worksheet and the values of the columns for each row separated by commas. This format is compatible with WEKA and was used to save the feature matrices for both datasets studied in separate files.

Since the study was not only focused in determining the value of the features for the datasets studied, but also to find classifiers that could perform well in this field, the feature selection methods were implemented in the *classify* tab of the WEKA interface, instead of the *select attributes* tab only. The *attribute selected classifier* algorithm is one of the available options of the *classify* tab in the WEKA explorer, and allows automatically performing classification trials with a determined classifier upon a subset of features resultant from the application of a chosen feature selection algorithm. The work's purpose was not to develop new classification algorithms, but to explore the potential of existing algorithms in the field of pigmented skin lesions using the selected features, and hence four classifiers (Naïve Bayes – NB; multilayer perceptron – MLP; J48 decision tree – J48; support vector machine – LibSVM) were chosen and used in their default configuration in WEKA, with exception of the SVM for which the parameters had to be previously tuned for it to return

reasonable results. The choice of these classifiers was made in order to include different approaches to classification and understand how differently from each other they could perform in this context. Three *feature ranking* methods and two *feature subset evaluation* were employed for preliminary evaluation of the feature space.

The *feature ranking* methods considered were the *Pearson's correlation attribute evaluation*, the *ReliefF*, and the *information gain attribute ranking* algorithms. The procedure followed to obtain the results for these algorithms was similar. In the *classify* tab of the WEKA explorer, the *grid-search* algorithm was used. This allows searching within a defined range of values, for a maximum of two parameters, for the values that maximize a certain performance measure of a given classifier, selected from the available options. One limitation found when using this method was that sensitivity was not available as the evaluation measure. The weighted AUC was chose instead, since it balances the contribution of sensitivity and specificity to the result and hence should lead to results that present high values for both the measures. The *attribute selected classifier* was selected as the classifier for the *grid-search*, experimenting the three *ranking* algorithms in combination with each of the four classifiers to experiment. The parameter fixed for all three methods in the *grid-search* was the number of features from the ranked list to use in the classification, and was set to range from 1 feature only, to using the whole set of 48 features.

For the *feature subset evaluation,* one *filter* approach, the *correlation-based feature selection* (CFS) method, and one *wrapper*, the *wrapper subset evaluation*, were employed. The use of these methods requires choosing a search strategy to define how the subset tested is changed at each step. For this work, a greedy hill-climbing searching in the forward direction was chose, starting with an empty subset and incrementally adding the feature that maximizes the evaluator defined criteria. The searching was set to terminate if no improvement in the output result was found after five node expansions. For the CFS method, since no classifier is involved in the evaluation, the subset of features it outputs for each dataset was experimented directly with the four classifiers. For the *wrapper* method, the output of the four classifiers is used to condition the search, leading to the selection of a different subset of features for each classifier.

A brief description of each of these methods is now provided:

- *Pearson's Correlation Attribute Evaluation*

This method defines the worth of an attribute by measuring the Pearson's correlation coefficient between it and the class label. The correlation coefficient is used to measure how well two variables relate, or how much a variation in the value of one is accompanied by a variation in the value of the other. Its value varies from zero (no correlation) to 1 (perfect correlation) and can be positive (increase in the value of one followed by an increase of value the other) or negative (increase followed by decrease or vice-versa).

Considering a set of m examples consisting of n features and one output label, where $x_{k,i}$ ($k = 1,..,m$ and $i = 1,..,n$) is used to represent the value of feature i from sample k, and y_k is used to represent the label of sample k, the Pearson's correlation coefficient of each feature, $R(i)$, can be calculated (see *equation 3.22*) (Tang et al., 2014) dividing the covariance between the feature and the output by the square-root of the product of the variance of each.

$$R(i) = \frac{\sum_{k=1}^{m}(x_{k,i}-\bar{x}_i)(y_k-\bar{y})}{\sqrt{\sum_{k=1}^{m}(x_{k,i}-\bar{x}_i)^2 \sum_{k=1}^{m}(y_k-\bar{y})^2}} \tag{3.22}$$

The output of this method is the list of ranked features according to the value of R, representing the order by which the features are the most correlated to the target labels.

- *Information Gain Attribute Ranking*

The *information gain attribute ranking* is a simple and widely used feature selection method. It evaluates the worth of an attribute by measuring the information gain with respect to the class. First the entropy of a class Y before (*equation 3.23*) and after (*equation 3.24*) observing an attribute X is measured. Entropy is the measure of disorder or unpredictability used for discrete variables. The amount by which the entropy of the class decreases reflects the additional information about the class provided by the attribute and is called information gain (Tang et al., 2014). Then, *equation 3.25* is used to assign each attribute X_i a score based on the information gain $IG(i)$, between itself and the class.

$$H(Y) = -\sum_{y \in Y} p(y) log_2(p(y)) \tag{3.23}$$

$$H(Y|X) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) log_2(p(y|x)) \tag{3.24}$$

The higher ranked features from the output list are the features that can best separate the input instances into homogeneous groups.

$$IG(i) = H(Y) - H(Y|X_i) = H(X_i) - H(X_i|Y) = H(X_i) + H(Y) - H(X_i, C) \tag{3.25}$$

- *ReliefF Attribute Evaluation*

This algorithm is an extension of the Relief algorithm, introduced by Kira et al. (Kira & Rendell, 1992). The relief algorithm works by randomly selecting an example from the data and locating its nearest neighbour according to each feature i from the same and opposite class, using a distance measure $d()$. The values of this feature for the nearest neighbors are compared to the selected example and then used to update its score $S(i)$ (see *equation 3.26*) (Kira & Rendell, 1992). In the equation for computing the score of each feature, M_k represents the values of feature i for the nearest neighbors to x_k with the class label and H_k represents the values of the feature for the nearest neighbors with different class label. The idea is that a useful attribute should differentiate well between different classes and have approximate values for instances of the same class. The process is repeated until a specified number of examples N to analyze. Since the number of examples available in this work is not exaggerated, the algorithm was set to sample all instances.

$$S(i) = \frac{1}{2} \sum_{k=1}^{N} d(x_{k,i} - x_{M_k,i}) - d(x_{k,i} - x_{H_k,i}) \tag{3.26}$$

The ReliefF (Kononenko & Simec, 1995) was proposed later and was designed to handle noise and multi-class data sets. It smoothes the effects of noise in the data by averaging the contribution of k nearest neighbors from the same and opposite class instead of using just the nearest neighbour. This is an additional parameter that must be defined, the number of neighbors to consider in the analysis, and the default value on WEKA implementation is 10.

- *Correlation-based Feature Selection*

The CFS (MA Hall, 2000) evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. The worth, $W(s)$, of a subset with k features is determined using *equation 3.27*, where

the nominator gives an indication of how predictive a group of features are, $\overline{r_{y,x}}$ is the average feature-class correlation; and the denominator indicator of how much redundancy there is among the considered subset, $\overline{r_{x,x}}$ is the average feature inter-correlation. CFS uses the symmetric uncertainty measure SU (see *equation 3.28*) to estimate the correlation measure between two attributes, X_1 and X_2.

$$W(s) = \frac{k\overline{r_{y,x}}}{\sqrt{(k+k(k-1)\overline{r_{x,x}}}} \tag{3.27}$$

$$SU = 2 \times [\frac{H(X_1)+H(X_2)-H(X_1,X_2)}{H(X_1)+H(X_2)}] \tag{3.28}$$

This method is capable of discarding irrelevant features, since they will have low correlation to the class; and discarding redundant features, because they will have high correlation with one or more of the other features. However, it treats attributes independently, and hence is incapable of identifying strongly interacting features (MA Hall, 2000).

For the heuristic search strategy, the *best-first* search algorithm was used. It explores a graph by expanding the most promising node chosen according to the previous criteria in each step, until the best result is found that does not improve for a defined number of expansions. It was implemented starting with an empty subset of features (forward direction), and the stopping criteria was the default in the WEKA implementation, to stop after 5 successive node expansions that did not improve the previous result.

- *Wrapper subset evaluation*

In the *wrapper subset evaluation*, the value of each generated subset of features is measured according to the performance of a classifier designed on them. Cross-validation is used to provide an estimate of the accuracy when using only the attributes in each subset. In this implementation, 10-fold cross validation was used, and the generation of each feature subset was performed using the *best-first* search algorithm previously described.

This method has the advantage of considering the effects of the selected feature subset on the performance of a specific induction algorithm, since the optimal set should depend on the specific biases it introduces. It was expected to perform better than the previous methods because of the interaction between the search and the learning scheme, but at the cost of significantly higher computational time, because for each generated subset, 10 models of the predictor are built to evaluate the cross-validation accuracy.

- *Exhaustive Search*

All of the above mentioned methods are usually employed as alternatives to performing an *exhaustive search*. In an *exhaustive search*, all possible feature combinations are evaluated. Such a search demands a considerable running time, since for a number of features N, 2^N combinations have to be evaluated, which can make the classification experiments unfeasible to be carried within a reasonable time. For the 48 features that were extracted for example, it would already represent experimenting 281 hundred billion combinations, which should take years to complete. However, this method could guarantee finding the optimum solution within the feature space for a given dataset, and allow extracting significant information about the contribution of features for the classification results that can be achieved. Since the results of the preliminary *feature selection* methods used were unsatisfactory (as presented in *chapter 4*), it was decided to implement an

approach to the *exhaustive search*. In order to accelerate the process of the search and avoid evaluating 2^{48} combinations, proper strategies were adopted, which are discussed further in this section.

The learning scheme adopted for this search was the LibSVM [65] implementation of support vector machines (SVM). SVM's are state-of-the art large margin classifiers that excel at performing binary classification tasks both in terms of time and efficiency. In the field of automatic classification of pigmented skin lesions, they have also been shown to outperform or at least perform as good as other algorithms (Stephan Dreiseitl et al., 2001; Torre, Caputo, & Tommasi, 2010).

A short introduction to the SVM classification is provided in (Cristianini & Shawe-Taylor, 2000), and some important concepts are presented here. Consider a set of n training data points $\{x_i, y_i\} \in R^n$, where x_i is a feature vector and $y_i \in \{-1,1\}$ is the output label. Supposing that there is a hyperplane $w \times x + b = 0$ that can separate the positive from the negative samples, then the optimal hyperplane is that which has the maximum distance to the closest points of opposite labels in the training set. The optimal values for w and b can be found by solving a constrained minimization problem:

$$f(x) = sign\left(\sum_{i=1}^{n} \alpha_i y_i w.x + b\right) \tag{3.29}$$

Where α_i and b are found during training. Most of the α_i's take the value of zero, while those x_i with non-zero coefficient are called the *support vectors*. When the two classes are not linearly separable, the Lagrange multipliers have to take on an upper bound $\alpha_i \leq C, i = 1, \dots, m$, where C determines a trade-off between the margin maximization and the error minimization. To generalize the equation for non-linear decision functions, a mapping function, ϕ, has to be introduced, to map the input data points to a high-dimensional space where the data points of the two classes can be linearly separable. The nonlinear decision functions can be constructed by assuming there exists a kernel $K(x_i, y_i) = \phi(x_i).\phi(y_i)$, which replaces the inner product $w.x$ in the previous equation. Popular kernel functions include the polynomial (*equation 3.30)* and the Gaussian radial basis function (*equation 3.31)* kernels.

$$K(x_i, y_i) = (\gamma^* x_i. y_i)^d \tag{3.30}$$
$$K(x_i, y_i) = \exp(-\gamma^* |x_i - y_i|^2) \tag{3.31}$$

The first step for implementing the SVM classification was to choose the kernel function to use. The radial basis function (RBF) kernel was chosen, based on the fact that a linear division surface could not perform well with the existing samples, and in terms of nonlinear kernels, the RBF is often chosen (M. E. Celebi et al., 2007) due to its stability and the little need for hyperparameter tuning, since there are only two parameters that should be adjusted, namely the cost or penalty C and the gamma γ. The C parameter defines a trade-off between misclassification of training examples against the simplicity of the decision surface (Cristianini & Shawe-Taylor, 2000). A low value of C makes the decision surface smooth, while a high C aims at classifying all training examples correctly by giving the model freedom to select more samples as support vectors. A high value of C often leads to a less generalizable model, since the model is too overfit to the training data used. The value of the parameter γ represents the kernel width (Cristianini & Shawe-Taylor, 2000), and defines how far the influence of a single training example reaches, where low values represent long

distances and high values represent short distances. If gamma is too large, the radius of the area of influence of the support vectors often only includes the support vector itself and leads to an unavoidable overfitting; and when it is very small, the resulting model is too constrained and cannot capture the complexity of the data, behaving similarly to a linear model.

In order to fix the C and γ to use in further classification trials, the performance of the SVM classifier with a RBF kernel when using the whole feature set was considered. Initially, following the procedure suggested in [85], exponentially growing sequences of values for the $C \in \{2^{-5}, 2^{-3}, 2^{-1}, \ldots., 2^{15}\}$ and $\gamma \in \{2^{-15}, 2^{-13}, 2^{-11}, \ldots, 2^{3}\}$ parameters of the kernel were tested. This was done by selecting the *grid-search* method available in the WEKA *classify* tab, using the libSVM with the RBF kernel as the classifier, and selecting its *cost* and *gamma* as the varying parameters in the *grid-search*, across the defined ranges. Using this method, the classification performance using each possible pair of parameters was evaluated, and the output of the algorithm revealed the group of parameters that lead to the best performance according to a determined evaluation measure, which was defined to be the weighted AUC. After finding values within the initial range that lead to reasonable performance, a narrower search was performed around their values, until no further improvement was found. Reasonable results were found using $C = 140$ and $\gamma = 0.08$, so these were the parameters kept for every classification trial using the SVM classifier. The choice of the parameters for a SVM classifier is an empirical process and for each combination of features tried it was likely that there was a pair of parameters that resulted in improved performance, but it should also lead to overoptimistic results which was undesired, so they were kept fixed in order for the resulting models to be consistent across every experiment.

The classification of the feature combinations was performed in Matlab, using the libSVM (Chang & Lin, 2011) package. Because the data available was limited, a leave-one-out approach to classification was used. This is a straightforward method that allows using the most out of limited data. To perform the leave-one-out classification on a set of m samples, for each combination of features and iteratively for each sample, the remaining $m - 1$ samples are used as the training data to create a classification model using the *svm_train* function, which is then applied to the analysed sample to produce an output, using the *svm_predict* function. Both of these functions are included in the standard libSVM package for Matlab. For each combination of features, the vector of the correct diagnostic of the images is compared to the vector containing the output of the *svm_predict* function for each sample, and the performance measures are determined. The registered measures were the number of wrongly classified instances, determined by counting the number of samples whose actual diagnostic differed from the algorithm's output for it, the sensitivity (see *equation 2.20*), quantifying the rate of correct detection of melanoma, and the specificity (see *equation 2.21*), the rate of correct detection of the benign lesions.

In order to perform an *exhaustive search,* the following steps can be used (for the example, consider $N = 10$):

1. For $i = 1{:}\, 2^{10} - 1$;
2. Convert i to binary form, with N digits
   $\rightarrow (1 = 0000000001)|(2 = 0000000010)|\ldots|(2^{10} = 1111111111)$
3. Create a vector $selections$ with 1 row and N columns and insert each digit from the previous in each column;

4. Create a temporary matrix from the $m \times n$ complete set, containing only the columns where the vector $selections$ has value 1;
5. Perform leave-one-out cross-validation classification on the temporary matrix.

To go around the limitations introduced by the number of features, the search was implemented in several steps, starting the search by forcing features of related concept or expected importance to be used together to decrease the total number of features to consider in the search, and splitting them up in a stepwise manner over the best results found in each previous step. One disadvantage that could be introduced by using this strategy is that by limiting the further search to the best results from the previous step, it was possible that splitting the feature groups used in a specific combination would not improve the results obtained since the overall combination was already the best possible when considering those groups.

The initial search was performed on 25 features, which corresponded to evaluating 33.5 million combinations of features. Using the binary representation, each of the 25 digits was associated to the feature/group of features to select from the original set of 48. When analyzing the binary representation of each number from 1 to $2^{25} - 1$, the index of the digits that had value 1 defined the feature/group of features to integrate in the combination of features for classification in each iteration. The feature groups considered are summarized in the list below:

- The seven *Hu's invariant moments*;
- *Average lesion lightness* with the *difference of average lightness* between lesion and surrounding skin;
- *Average lesion saturation* with the *difference of average saturation* between lesion and surrounding skin;
- *Distance of average (a*,b*)* and *(u*,v*)* between lesion and surrounding skin;
- *Averages* with *standard deviations* of each *color channel* considered;
- *Averages* with *ranges* of each *texture feature*.

The groups above were only considered to make the complete search a viable option on the situation considered. They were formed based on their concept similarity and expected importance, which is the case for every group with exception of the second last presented that also represented two physically dependent variables. The concept of the standard deviation carries no meaning when used without the average, and so it should not be used in a combination where the average is not present as well.

For each iteration evaluated, the combination of features used, the number of total misclassifications, sensitivity and specificity values achieved were saved in separate lines of a text file, for them to be available for further experiments. After completion of the first group of iterations the combinations of features that obtained the highest sensitivity, and the combinations that resulted in the lowest number of wrongly classified instances were considered to continue the search. The continuing of the search was to splitting the grouped features step by step, to assess its importance individually for each of these results. For each of the top results considered. Analyzing the combinations considered, if a group of features was being used, they were split, according to the following:

1. *Seven Hu's Invariant Moments:* because in the first step of the search they were only used all together, they were not part of any of the top results (discussed further in *section 4.3*). Although, since it was expected that they would have value if used alone, all possible combinations of these seven features was experimented with each of the selected top results. If their sensitivity or number of wrongly classified instances improved, they were saved; if not, the original combination of features was kept as a top result;
2. *Distances of average* (a^*, b^*) and (u^*, v^*) between lesion and surrounding skin: if the group of these features was selected, the combination using only one or the other was experimented, and saved if any improved the initial result that considered both;
3. *Differences of lightness* and *saturation* between the lesion and the skin: same as in 2.;
4. *Averages'* and *ranges'* of texture properties: same as in 2.;
5. *Averages'* and *standard deviations'* of color properties: for these grouped features, only the possibility of using the average alone was considered versus the use of the two together.

The exhaustive search as suggested was first performed until the first step of the previous list only for the first 100 dermoscopic images considered, due to time limitations. After it, the best combinations found were applied to the feature matrix extracted from the 200 images of the PH^2 dataset in an attempt to understand if the best performing feature combinations on for the first dataset could perform well with unseen lesions. The rest of the steps were applied to both datasets, which corresponded to evaluating several million combinations and allowed finding the best performing feature combinations for each dataset and how they differ between them.

With the exception of step 1, all of the remaining were applied only if the feature group appeared in the original best combination selected. Although this strategy was expected to fail due to optimum solution that was already obtained in a specific combination, as discussed in the following chapter, in some situations, the splitting of the feature groups lead to an improvement of the overall result.

In addition to the search performed in combinations of features belonging to the three category of descriptors studied, an exhaustive search was also made in each group individually, in order to highlight the difference between the discriminative power of using the shape, color and texture descriptors alone and using them in combination.

3.7 Summary

In the present chapter, the computational methods adopted in this project were described. The work was divided in two main stages, namely the feature extraction stage, and the evaluation of the features' value and their prediction performance on the classification experiments. For the feature extraction stage, the MATLAB software was used, a programming environment that excels at matrix manipulation and contains several mature algorithms for image processing tasks, for which it was thought to be adequate for this application. In order to evaluate the worth of the extracted attributes, and determine their

discriminative potential, the WEKA and MATLAB were used. The WEKA was used because it is open source software, with an intuitive graphical user interface that permits easily applying a wide variety of state-of-the-art selection and classification algorithms to the extracted data, allowing to study its value in the context of application. In this stage, MATLAB was used to perform the complete search through the feature space, applying the LibSVM package to evaluate the predictive performance of the combinations generated using SVM classification.

For this study two dermoscopy image datasets were used. The first was obtained in three different private dermatology practices and was composed of 100 lesions, from which 29 were melanomas. One important characteristic of this dataset are the early stages at which the melanomas are presented and the numerous atypical benign lesions hard to differentiate from the malignant, which was important to evaluate the robustness of the features selected. It should also be highlighted that as this images were obtained from different hospitals, the acquisition conditions varied significantly, which also constitutes an important challenge for the automatic classification. The disadvantage of this dataset is that the images were acquired by scanning dermoscopy slides, which often introduced blurring in the images and caused the extraction of some features to be inaccurate. Regarding the second dataset of images, it contained 200 lesions from which 40 were melanomas. These were acquired at the dermatology service of hospital Pedro Hispano all under the same conditions. The advantage of this dataset is that it is publicly available at FCUP's PH^2 database with a significant amount of information available about each lesion, and can therefore be used to benchmark results with other studies using it. One important aspect of this dataset is that the melanomas are at later stages of development, showing clear distinctive attributes from the benign lesions available, which makes this dataset expectedly easier to classify in an automatic setting. One additional limitation highlighted for these images is that for most melanomas the manual segmentation provided was inaccurate because they did not entirely fit in the dermatoscope capture area, which can also affect the extraction of features from them.

The methods used to extract the features from the available lesions were individually presented in this chapter. For each lesion the result of the manual segmentation of the lesion region performed by an experienced dermatologist was available, which allowed focusing the feature extraction stage on this region, and avoid the influence that errors associated with automatic segmentation of the lesions could introduce. The features extracted were grouped in three main categories, namely shape, color and texture, which were motivated by previous research and the aspects considered in visual evaluation of melanocytic lesions. The shape features selected were used to quantify the irregularity and asymmetry of shape and border of the lesions, and with the exception of the asymmetry index were all region-based, as the use of contour-based shape features is highly dependent on accurate contour definition. The color features were used to define the content of colour of the lesions, and the distribution of color within the lesion. In this category of descriptors, some attributes were included that measured the transition of color between the lesion region and the surrounding skin, as it was believed it could carry important diagnostic value. Texture features were determined in an attempt to identify differences that the presence of dermoscopic structures specific to malignant lesions introduced in their structural pattern. The features were extracted using simple computational methods, but were thought to be representative of the main differentiating aspects between benign and malignant lesions, and would be beneficial if

proved valuable in this context, as they can be accurately measured from different acquisition settings. Although most features considered in this work had already been used in previous works, the combination used in the end was unique, and their individual contribution to the task has not yet been studied.

In order to address the latter, feature selection methods were applied, and their results evaluated using different classification algorithms. These methods were described in the previous section, and their presentation was divided in three subsections. First, for a preliminary evaluation using WEKA, *feature ranking* methods were applied, which perform an individual assessment of an attribute's worth by measuring their discriminative potential between the target classes according to a specified evaluation criteria. These methods allow comparing the value of different attributes in the context of the dataset used. Second, and still for a preliminary evaluation of the features' value, *subset evaluators* were used, which evaluate subsets of features from the whole set considered according to specific evaluation criteria or to the performance of a specific machine learning algorithm. These methods have the advantage of considering the features in the context of others, thus possibly including features that are irrelevant if used individually, but valuable when grouped. Both these approaches for selecting features were tested with four machine learning algorithms found in the literature to be suited for binary classification tasks, as is the one considered in this work. In order to deal with the limited amount of data available, all classification experiments were performed using the leave-one-out strategy, which consists of keeping one sample for testing the output of a classification model, and the remaining for creating the model to be tested, and repeating the process for every sample available.

Additionally to the use of the feature selection methods, an empirical search through the feature space was also used. This strategy involved experimenting all possible combinations of features for classification, which although being very demanding computationally, was thought to be able to convey relevant information about the attributes that can best perform in the context of each dataset considered. In this approach, for each combination of features generated, a SVM classifier was used to assess its predictive performance. The SVM classifier was implemented using the LibSVM package for MATLAB, which allows easily creating different formulations of this classification strategy for evaluating the available data. To condition the creation of the classification model through this approach, a preliminary step of parameter tuning was performed using the whole feature set calculated from dataset 1. The best parameters found from this were then used in every experiment with this classifier, which was done in order to avoid overoptimistic results that could result from adjusting the parameters to each trial. During this exhaustive search, the results of every combination tested were saved, which allowed creating a large amount of information about which combinations of features can perform the best for the automatic classification, and which features contribute the most for it. This strategy was also used to evaluate the performance of classification for single categories of descriptors, to help determining which are more helpful for the task of automatic melanoma recognition.

The results obtained using the methods previously described in this chapter together with their discussion are presented in the following chapter.

Chapter 4

Results and Discussion

In this chapter, the results obtained from the adopted methodology are presented and discussed. The problems related to the feature extraction will be made clear in conjunction with the results found in the classification steps. The comparison of the different classification methods following the various feature selection strategies was based upon three performance measures, namely the number of wrongly classified images, the sensitivity and specificity. These allow directly inferring about the overall classification accuracy, the rate of melanoma detection, and the rate of detection of benign lesions, respectively. Across every result obtained, focus was given to results that achieved the lowest number of misclassifications, with high sensitivities. Although high rate of detection of melanomas is the most important aspect of the classification, it is also important that not too many benign lesions are considered to be melanomas, as in a practical setting it could lead to many unnecessary excisions.

In the first part of this chapter, the results of the application of *feature ranking* methods on the two image datasets are presented. This section intends to highlight the value of the considered features when evaluated individually, allowing to infer about their expected importance to the field of pigmented skin lesions. It also presents the classification results obtained using the features ordered by these methods, when applied to four different classifiers. Section *4.2* is dedicated to the results obtained using *feature subset evaluation* methods, presenting the sets of features selected and the classification performance when using them, showing the main differences between the evaluation of the features' worth individually and when considered in the context of others. The results presented in these sections also allow comparing the performance of the four classifiers considered, and helped selecting the classifier to use in the exhaustive search. The last section presents the best combinations of features and classification results found using the adopted exhaustive search. This step aims at showing the differences between the results that can be obtained in the classification task after using feature selection methods and after performing an empirical search through the feature space. The classification results using the best combinations of features found for the 100 images' dataset are presented, and also how they perform on the separate dataset of 200 images. The classification performance obtained when using descriptors of each category individually versus using them combined is also discussed in this section.

Throughout this chapter the dataset of 100 images is referenced as *dataset 1,* and the 200 images used for validating the generalization of the results obtained with the previous as *dataset 2*. Feature selection algorithms were first applied to the 100 images' dataset, from which the results lead to the decision of applying an exhaustive search. Since the time required to perform the exhaustive search was too long, the study of the discrimination potential of features in the second dataset was only made posteriorly to the former, and hence they are considered separately. Nonetheless, the results for both datasets are presented together in each section, since it allows highlighting the main differences present between the data available in each.

To facilitate the presentation in some of the results, a list of the abbreviations used to represent each of the features considered is presented in *table 4.1*.

Table 4.1 - List of the Acronyms used to represent the features.

Shape Features	Color Features	Texture Features
1st Hu's Invariant moment - **Hu1**	Average R - **aR**	Average of Contrast - **aCont**
2nd Hu's Invariant moment - **Hu2**	Standard deviation in R - **sR**	Range of Contrast - **rCont**
3rd Hu's Invariant moment - **Hu3**	Average G - **aG**	Average of Correlation - **aCorr**
4th Hu's Invariant moment - **Hu4**	Standard deviation in G - **sG**	Range of Correlation - **rCorr**
5th Hu's Invariant moment - **Hu5**	Average B - **aB**	Average of Dissimilarity - **aDiss**
6th Hu's Invariant moment - **Hu6**	Standard deviation in B - **sB**	Range of Dissimilarity - **rDiss**
7th Hu's Invariant moment - **Hu7**	Average lesion saturation - **Sat**	Average of Energy - **aEn**
Compactness - **C**	Difference of average saturation - **dSat**	Range of Energy - **rEn**
Rectangularity - **Rect**	Average lesion lightness - **aL***	Average of Entropy - **aEntro**
Solidity - **Sol**	Difference of average lightness - **dL***	Range of Entropy - **rEntro**
Lengthening Index - **LInd**	Average in a* - **aA***	Average of Homogeneity - **aHomo**
Asymmetry Index - **Asym**	Standard deviation in a* - **sA***	Range of Homogeneity - **rHomo**
	Average in b* - **aB***	Average of Maximum Probability - **aMax**
	Standard deviation in b* - **sB***	Range of Maximum Probability - **rMax**
	Distance of average (a*,b*) - **dAB**	Average of Inverse Difference Moment normalized - **aInv**
	Average in u* - **aU***	Range of Inverse Difference Moment normalized - **rInv**
	Standard deviation in u* - **sU***	
	Average in v* - **aV***	
	Standard deviation in v* - **sV***	
	Distance of average (u*,v*) - **dUV**	

4.1. Feature ranking methods

The *feature ranking* methods were applied as a preliminary step to assess the relevance of the selected features for the classification task at hand. As referred, the main advantage

of using these methods is its speed of computation, since they only evaluate the features individually according to specific criteria and rank them in order of the value that is obtained for each. On the other hand, they ignore the interaction between features and how combining them may improve classification performance with relation to a certain target. Because of the latter, they were expected to perform poorer than the more time consuming methods experimented.

The methods considered were the *Pearson's Correlation Coefficient* attribute evaluation; the *ReliefF* algorithm and the *Information gain* attribute evaluation. The *Pearson's correlation coefficient* is used to measure the degree of linear dependence that exists between each feature and the target classes. The higher ranked features according to this criterion are those whose values can better predict the class of the considered samples. The weight given to each feature by the *ReliefF* algorithm is calculated based on the distance of that feature between instances of the same class, and between instances of opposite classes, and the most valuable features according to this criterion are those that present small intra-class and large inter-class distances. The *information gain* is a measure used to evaluate the features according to the increase in information that each provides regarding the target classes, and is used to order the features by their ability to separate the input data into homogeneous groups. These methods, although conceptually different in how they perform the evaluation (relying on correlation, distance and information measures, respectively), served the same purpose, which was finding the most relevant features from the selected set to this context of application. Due to the significant differences between them, their results were expected to differ from each other in the ordering of features, but also to show similarities regarding the most significant, which should be those that consistently figure in higher ranks across the two image datasets considered. The calculations behind each method were previously described in *section 3.6*.

Table 4.2 - Partial list regarding the ranking of features according to the three feature ranking methods considered. The features are presented in order of the rank they obtained in each method (1st, 2nd, 3rd…15th).

Evaluation Criterion	Image Dataset	
	Dataset 1	**Dataset 2**
Pearson's Coefficient	aSat*, aA*, sB*, dUV, rHomo, aCorr, aU*, rEntro, sA*, dSat, rCorr, aB*, aR, aInv, aCont	*LInd, aV*, aB*, dSat, sV*, sB*, aU*, aL*, aR, sR, sU*, aA*, dL*, rCont, Rect*
ReliefF	*aInv, aCont, aU*, aDiss, aR, aSat, aHomo, rHomo, aA*, aCorr, rEntro, aB*, rInv, rCont, dSat, dUV*	*LInd, aV*, aB*, dSat, aR, aSat, aU*, aL*, dL*,sB*, sR, sV*, aB, aG, sU**
Information Gain	*aU*, aSat, aCorr, aR, aA*, aCont, rHomo, aInv, rEntro, aB*, aDiss, dSat, dUV, rCorr, sB**	*dSat, aV*, Lind, aB*, aU*, sV*, aSat, sB*, sR, sU*, Rect, rCont, aR, aA*, rInv*

In order to facilitate the visualization of the results of the three methods, it was decided to group them and divide their presentation and analysis in two steps. In the first step, the best results of the ranking of features obtained for the three algorithms are presented, in

table 4.2, with the goal of highlighting the differences and similarities obtained with them between the two datasets, and to conclude about the individual worth of the features in the context of each. It was not possible to present the ranked list of the 48 features returned by each method, because it would take too much space, and would be hard to visualize and understand. It was also thought it would not provide the most useful information, since in most situations, below a certain ranking position the features all had similar values close to zero and could not therefore be distinguished as being more or less relevant than the rest. To facilitate the analysis, only the top 15 features from each ranking are presented. The analysis of these results is now presented, regarding each category of descriptors individually.

The shape descriptors considered in this work were selected in an attempt to capture the irregular and asymmetric shape and the irregular borders that are characteristic of melanomas. In a dermoscopic evaluation of skin lesions, features regarding the irregular contours and asymmetry of shape are considered in the ABCD rule of dermoscopy and the CASH algorithm, while in other methods these are outweighed by the presence of atypical structures and color patterns. The overlooking of these characteristics in some of the methods is probably associated with the common observation of their presence in the atypical benign melanocytic lesions. In the datasets studied, the benign lesions often presented irregular borders and asymmetric shapes and some melanomas, especially in *dataset 1*, presented approximately circular shapes without significantly irregular borders. In *figure 4.1 a)* and *b),* examples of the manual segmentation of one benign and one malignant lesion are presented, respectively, to demonstrate the previous statement. Following these observations, it was expected that shape descriptors would not be the most reliable differentiators between the lesions if considered individually. This can be confirmed by looking at the results presented in table *4.2,* where it is possible to see that for *dataset 1* no shape descriptors figured in the top rankings of any of the three methods, while for *dataset 2,* the *lengthening index* and the *rectangularity* were present in most. However, it should be noted that in the latter, some images did not have a good manual segmentation and some lesions did not fit entirely in the image borders, and so the rectilinear borders of the image had to be used as the lesion's borders. This happened the most in the case of the larger lesions, which was essentially true for most melanomas, and in these situations the shape features must have been gravely deteriorated because of their dependence on the borders considered. This fact might have made some shape features to be misleadingly valuable in the differentiation between the lesions in this dataset, due to assuming distinctive values in those for which the image border was considered as the lesion's, which happened mostly for the malignant lesions. An example of this is presented in *figure 4.1 c),* showing the manual segmentation of a malignant lesion for which the image borders were considered as the lesion's. These observations do not indicate that shape descriptors are irrelevant for skin lesions differentiation, but that they are likely to provide a worse prediction if no additional information is considered.

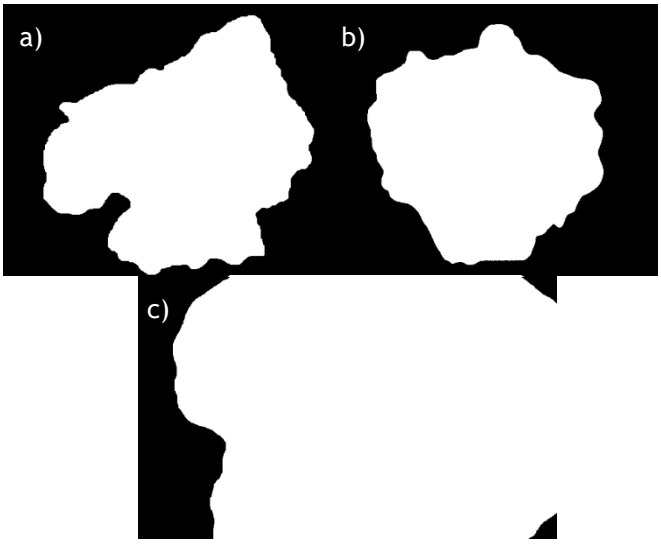


Figure 4.1 – Manual segmentation of lesions emphasizing the difficulty of differentiating between lesions based on shape measures only: a) irregular border of an atypical benign lesion; b) compact shape of a melanoma; c) border considered for a malignant lesion that did not fit the image window from PH^2 dataset.

From the reviewed literature, color information was expected to be indispensable for the diagnosis of melanoma. For all the presented scoring methods commonly used in dermatology, for example, color information is considered either directly by evaluating the number of colors and its distribution across the lesion (in the ABCD rule of dermoscopy and the CASH algorithm) or indirectly by assessing the presence of regression structures, blotches or the blue-whitish veil (in the Menzies method, the seven-point checklist and pattern analysis approaches). In this work, the average colors and its distribution across the lesion were evaluated through simple statistics computed from the RGB, $L^*a^*b^*$ and $L^*u^*v^*$ color spaces, namely through the average and the standard deviation, respectively. These measures should account for the differences existent between colors present in the benign and malignant lesions, and therefore provide important diagnostic cues for the automatic classification. Additionally, descriptors that considered the relation between the healthy skin and the lesion were also included, regarding the measures of color saturation, lightness and the absolute color described by the coordinates (a^*, b^*) and (u^*, v^*), in an attempt to quantify the transitions from the lesions to its periphery, which was also expected to carry significant diagnostic value. From the results obtained in the ranking methods, it was clear that this category of descriptors showed the most consistency between the two datasets studied. This was especially true for the simple statistics computed over the CIE color spaces, namely the *average* in a^*, b^* and u^* coordinates, and the standard deviation in b^*; and in the RGB color space, for the *average* of the red component, the *average lesion saturation* and the difference between the *average saturation* of the lesion and the surrounding skin. With few exceptions, these measures figured in the top rankings of all three methods, for both datasets, which suggests that these features should be the most reliable for distinguishing between benign and malignant lesions. Also, since it is the only category showing similar presence in both datasets, it indicates that from the features selected, the ones carrying the color information are the most valuable for classification across the two datasets. In fact, the vast majority of features highlighted by these methods for *dataset 2* were related to the color category, including the *average* and *standard deviation* in the v^* coordinate, the *standard deviation* of *red*, *average lesion lightness* and *difference* between the *average lesion* and *skin lightness* which were also emphasized by the three ranking criteria considered. The importance of color for characterizing melanomas in this dataset can be

highlighted by *figure 4.2 a),* a melanoma exhibiting multiple colors, including light to dark brown, black, the blue-gray in the blue whitish veil, and areas with no pigmentation, aspects not frequent in the benign lesions of this dataset. For *dataset 1*, the less relevance given to some color features could be associated to the lower resolutions that these images presented, since they were acquired from scanning dermoscopy slides and often presented significant blurring, which makes the color measures to be less accurate and therefore have lower influence on the results. In *figure 4.2 b)* it is possible to see an example of a lesion from *dataset 1* with intense blurring, and where color information is clearly affected.

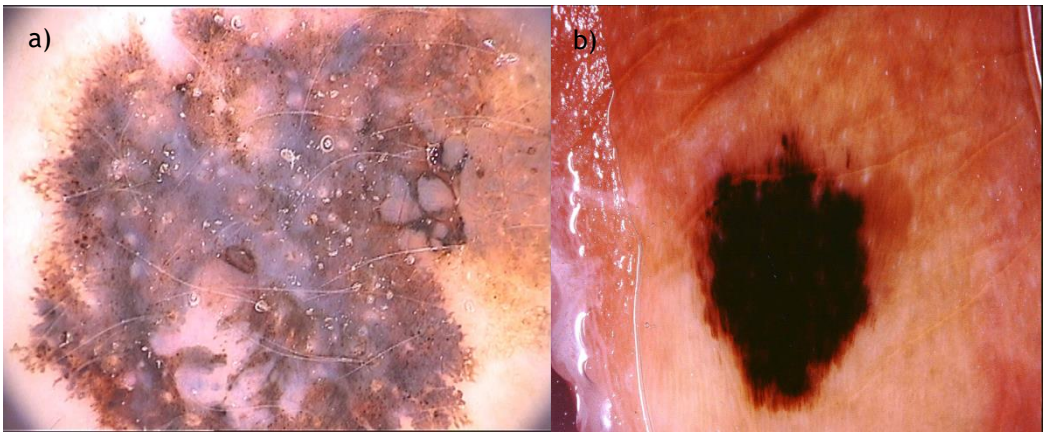


Figure 4.2 - Examples illustrating the presence of color in the lesions from the datasets studied: a) melanoma from PH^2 dataset exhibiting brown, dark brown, blue-whitish veil (blue/gray area) and areas with no pigmentation; b) lesion from *dataset 1* with intense blurring, and hence low resolution of color information.

Regarding the texture features, their extraction was based on previous works, and was used in an attempt to discriminate between the anatomic structures that dermatologists consider in all scoring algorithms used for dermoscopic evaluation of melanocytic lesions as important indicators of malignancy, like the presence of atypical pigment network, structureless areas, regression areas and branched streaks, for example. Examples of the presentation of some of these structures in the lesions considered for this work are presented in *figure 4.3 a)* and *b)* for a benign and a malignant lesion from *dataset 1,* respectively. It can be seen that for the case of the benign lesion, the pigment network is presented throughout the lesion with regularly distributed thick lines and uniformly spaced holes except for the area with hyperpigmentation; while for the malignant lesions it can be seen that the pigment network is only present in some regions, and is not consistent throughout them, the presence of a large strucuteless area in the middle of the lesion and also branched streaks throughout its periphery. The use of GLCM features extracted from manually delineated atypical structures has been shown (Shrestha et al., 2010; W. V. Stoecker et al., 2011) to be very effective in discriminating between benign and malignant lesions, however in an automatic setting this would require the implementation of complex segmentation techniques to first detect the presence of these structures. The assumption is that the texture measures selected, when calculated from the entire lesion region are capable of capturing significant differences between the structural organization in the two types of lesions, and therefore carry important diagnostic value. However, in the reviewed works that considered these features, no information was provided about which measures were more valuable, or if these features were selected in the best classification results achieved. From the preliminary evaluation performed in this work, it was found that the texture features selected showed significantly more relevance in the context of *dataset 1*. The measures highlighted by this

evaluation were the *average* and *range* of *Correlation*, the *average* and *range* of *Homogeneity*; the *average* of the *Inverse difference moment*, *Contrast* and *Dissimilarity* and the *range* of *Entropy*. For *dataset 2*, the measures present in the top rankings only included two measures not present in the results for the first dataset, namely the *range* of *Contrast* and of the *Inverse difference moment*. The measures of the *Energy* and *Maximum probability* are disregarded by these findings. These results suggest that while for *dataset 1*, most texture measures considered have significant discriminative potential between lesions; for *dataset 2* their relevance is outweighed by the use of color features. This observation may be due to the lesions in *dataset 2* not fitting entirely inside the image borders, which implicates that important information is missed because the structures present in the periphery of the lesions are not considered. Additionally, the computation of the GLCM ignores border pixels because these have no neighboring pixels in the respective direction, which aggravates the previous problem even further. This difference could also be justified by the small pixel distance used for the computation of texture features, which might not be suited for the often larger lesions found in *dataset 2*. However the use of small pixel distances was expected to be able to capture more accurate details due to the higher amount of information considered, but to be less robust to noise. Still, the texture features extracted from the images in *dataset 1,* which presented significantly more noise, proved to be relevant for this task and hence the pixel distance chosen for their computation seems to be appropriate.

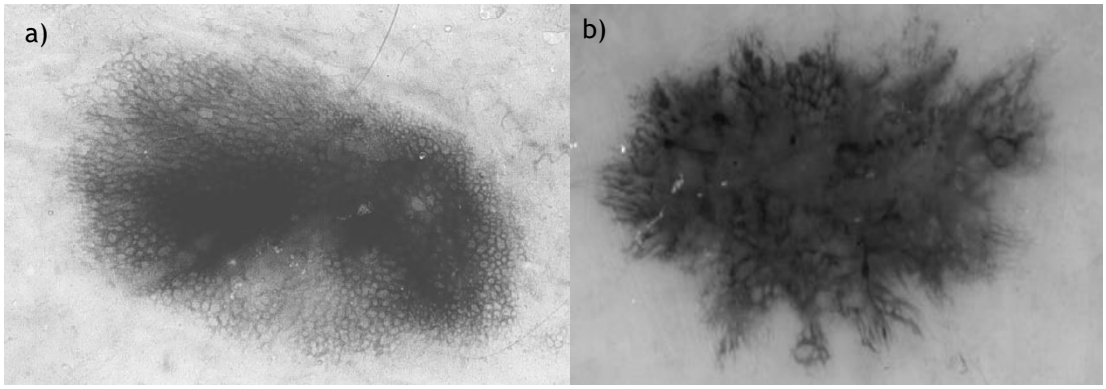


Figure 4.3 - Examples of differentiating texture aspects between benign and malignant lesions: a) regular pigment network and presence of blotches in a benign lesion; b) malignant lesion exhibiting branched streaks, dots, a structureless area and regions of irregular pigment network.

In order to evaluate the prediction performance that could be achieved following the filtering of features according to their relevance measured by the three criterion considered, four classifiers were tested using 1 to 48 features from each of the ranked lists. The best result that each was able to achieve for the images in *dataset 1* and *dataset 2* are presented in *table 4.3* and *table 4.4*, respectively. For each *ranking* method considered, the results obtained using the four classifiers are presented horizontally. For each classifier, four measures were kept from the best result obtained, namely the number of features selected (NS) from the ranked list, the number of wrongly classified images (NW) which also includes the number of missed melanomas and benign lesions inside parenthesis, the sensitivity (SN), presented as a percentage calculated as the ratio between the number of wrongly classified melanomas and the total number of melanomas in the dataset and specificity (SP), calculated as the ratio between the number of failed benign lesions and the total number of benign lesions available.

Table 4.3 – Best classification results achieved from the application of four classifiers after selection of features based on *feature ranking methods* on *dataset 1.*

Feature ranking method[1]	Classifier															
	Naïve Bayes (NB)				Multilayer Perceptron (MLP)				Support Vector Machine (SVM)				Decision tree (J48)			
	NS	NW[2]	SN (%)	SP (%)	NS	NW[2]	SN (%)	SP (%)	NS	NW[2]	SN (%)	SP (%)	NS	NW[2]	SN (%)	SP (%)
P.C.	25	23 (7/16)	76.00	77.50	26	27 (17/10)	41.40	85.90	43	18 (9/9)	69.00	87.32	2	27 (12/15)	58.60	78.90
RF	34	24 (5/19)	82.80	73.25	26	26 (15/11)	48.30	84.50	33	20 (10/10)	65.50	85.90	2	29 (15/14)	48.30	80.30
I.G.	16	25 (8/17)	72.00	76.10	19	30 (18/12)	37.90	83.10	25	21 (10/11)	68.90	84.50	20	34 (16/18)	44.80	74.60

[1] Criterion used to establish the ranking of features: *P.C.* – Pearson's Correlation Coefficient; *RF* – ReliefF algorithm; *I.G.* – Information gain;

[2] Number of wrongly classified images presented as T (M/NM): Total number of misclassifications (number of melanomas wrongly classified / number of benign lesions wrongly classified);

Table 4.4 - Best classification results achieved from the application of four classifiers after selection of features based on *feature ranking methods* on *dataset 2.*

Feature ranking method[1]	Classifier															
	Naïve Bayes (NB)				Multilayer Perceptron (MLP)				Support Vector Machine (SVM)				Decision tree (J48)			
	NS	NW[2]	SN (%)	SP (%)	NS	NW[2]	SN (%)	SP (%)	NS	NW[2]	SN (%)	SP (%)	NS	NW[2]	SN (%)	SP (%)
P.C.	10	28 (8/20)	80.00	87.50	30	17 (8/9)	80.00	94.40	3	18 (11/7)	72.50	95.60	2	24 (13/11)	67.50	93.10
RF	2	25 (9/16)	77.50	90.00	44	20 (9/11)	77.50	93.10	20	13 (10/3)	75.00	98.10	27	27 (16/11)	60.00	93.10
I.G.	2	35 (8/27)	80.00	83.10	44	20(10/10)	75.00	93.80	20	19 (10/9)	75.00	94.38	9	22 (11/11)	72.50	93.10

[1] Criterion used to establish the ranking of features: *P.C.* – Pearson's Correlation Coefficient; *RF* – ReliefF algorithm; *I.G.* – Information gain;

[2] Number of wrongly classified images presented as T (M/NM): Total number of misclassifications (number of melanomas wrongly classified / number of benign lesions wrongly classified);

From the results presented in *table 4.3* and *4.4*, it is possible to infer that *dataset 2* presents a smaller challenge for classification using the selected features. This can be observed by the lower number of misclassifications, higher sensitivities and specificities achieved by the four classifiers in almost every trial for this dataset, even though having twice more samples. Another indication for the previous observation is the lower number of features required to achieve the best results for every classifier with the exception of the artificial neural network considered, the MLP classifier. Two results that can be highlighted in this regard are the performance obtained with the Naïve Bayes classifier following selection according to the *information gain* criterion and with the support vector machine following selection with the *Pearson's correlation coefficient*. For the former, using only the *difference* between the *average lesion* and *skin saturation* and the *average* in v^*, the NB classifier achieved a sensitivity of 80% (8 out of 40 melanomas missed) and a specificity of 83.10% (27 out of 160 benign lesions missed); while for the SVM, using only the *lengthening index, average* in v^* and in b^*, this classifier achieved a sensitivity of 72.50% (11 melanomas missed) with a specificity of 95.60% (7 benign lesions missed). The high specificities achieved for this dataset were expected due to the significant unbalance that exists between the two classes (160 benign lesions and 40 melanomas), for which there is more data to train the models on from benign lesions. However, it was expected that this unbalance would also pose a challenge for the correct detection for the malignant lesions because of the scarcity of data related to them, but the highest sensitivities achieved are actually reasonable, which may be due to the advanced stage that the melanomas in this dataset present, hence showing significant differences in the characteristics measured. For the classification of *dataset 1,* the results were less promising, and the best results were often obtained considering a larger number of features from the ranked list, emphasizing the fact that the classification using this dataset was more difficult.

Overall, the worst performing classifier on both datasets was the decision tree algorithm, J48. The best result obtained for *dataset 1* with this classifier used only the two top ranked features from the *Pearson's coefficient* ranking, and achieved a sensitivity of 58.60% with a specificity of 78.90%. For *dataset 2*, it achieved a sensitivity of 72.50% with a specificity of 93.10% with the ranking according to the *information gain*, but was also the worst result considering the performance of the other classifiers. It is likely that the simple division rules this classifier attempts to produce are insufficient for the complex problem of differentiating between pigmented skin lesions when considering the features selected in this work. For the multilayer perceptron classifier, it can be seen that it consistently performed the worst when evaluating the samples in *dataset 1,* but was able to match or even outperform the other classifiers when considering the images in *dataset 2*. This classifier is the slowest to build a classification model, and showed reasonable performance only when considering the case where the two classes were better separated. Its poor performance regarding *dataset 1* should, however, be related to the complex cases it presented, for which further tuning of its parameters, namely the number of hidden layers and nodes to build the network, was likely to be required to enhance the results that could be achieved.

Both the support vector machine (SVM) and the naïve bayes (NB) algorithms have proven to be the most consistent performing across the two datasets. The former was capable of achieving higher overall accuracies (lower number of misclassifications) for every *feature ranking* method considered, while the latter consistently achieved higher sensitivities, failing less in the detection of melanomas for both cases. It can also be seen that both these

classifiers use more features to achieve the best result in *dataset 1* than in *dataset 2*, which also highlights the increased challenge that the first dataset poses for classification. For *dataset 1,* the best result obtained with the NB classifier used 34 features selected from the ranking returned by the *ReliefF* algorithm, and achieved 82.80% sensitivity (5 melanomas missed out of 29) and 73.25% specificity (19 benign lesions missed out of 71). Although achieving encouraging results regarding the rate of melanoma detection within the more difficult dataset, it corresponded to failing a quarter of the 100 images available in this dataset, which is unsatisfactory. The SVM achieved the best result of 69% sensitivity (9 melanomas missed) with 87.32% specificity (9 benign lesions missed) for this dataset following the selection of 43 features according to the *Pearson's correlation coefficient*. This result shows an improvement of overall accuracy from the former, but with a significant increase of the number of melanomas missed. Regarding the *dataset 2,* the NB was capable of achieving 80% sensitivity (8 melanomas missed) with 87.50% specificity (20 benign lesions missed) with 10 features selected by the *Pearson's correlation coefficient*, while the SVM performed the best with 20 features following selection by the *ReliefF* algorithm, obtaining 75% sensitivity (10 melanomas missed) and 98.10% specificity (3 benign lesions missed). The trend verified in the performance of the two classifiers using this dataset was similar to what was observed in the previous, but the SVM was capable of outperforming the NB in this case, since although failing 2 more melanomas than the latter in its best result, it was capable of correctly predicting almost every benign lesion.

The classification results achieved when performing the selection of features based on measures of their relevance were mediocre, especially when considering the results obtained using *dataset 1*. It should also be noted that, as previously referred, the datasets were evaluated separately, and this might have led to overoptimistic results due to inevitable similarities existing between the lesions present in each. This effect should be more pronounced for *dataset 2,* since all lesions were acquired in the same hospital, and should be another reason for the better results found for it. However, the results of these algorithms helped understanding some of the underlying characteristics of the data used, namely the worth of the selected attributes according to each dataset, and which were expected to be the most relevant in the context studied. Even though, it is important to refer that this only revealed an individual evaluation of each attribute, and hence is not surprising that the classification results that follow from it were inferior to what was desired. One important aspect that can be concluded from the analysis performed is the heterogeneity that exists between the two datasets experimented, verified in the significant differences between the ranked lists of features produced. To determine the effect of taking into consideration the relation between attributes, *feature subset evaluation* methods were investigated and are presented in the next section.

4.2. Feature subset evaluation

Following the *feature ranking* methods, two feature *subset evaluators* were tested. This step was added to the analysis with the purpose of verifying if they could lead to an improvement of the classification performances obtained thus far by evaluating the value of subsets of features instead of considering them individually. The use of these methods was expected to generate improvements from the classification results obtained previously and to highlight features that might have been classified as potentially irrelevant by the previous

criteria, but that carry important predictive value when considered in relation with other features. One filter method, the *correlation based feature subset evaluator*, and one wrapper method, the *wrapper subset evaluator*, were applied to the available data using the WEKA default configuration.

These methods avoid evaluating exhaustively every possible combination of features by using a strategy to search the feature space that alleviates the number of combinations that needs to be tested. In this work a best first strategy was used, which searches the space of attribute subsets by greedy hillclimbing augmented with a backtracking facility. The search was set in the forward direction, starting with an empty set of features, and incrementally adding the feature that maximizes the evaluation criteria. Since these methods need to evaluate several possible combinations at each step, they required more time to run than the *feature ranking* methods.

- *Correlation-based feature subset evaluator*

The first feature *subset evaluator* algorithm tested was the *correlation based feature selection*. This algorithm measures the worth of a subset by determining the inter-correlation between its features, and their correlation to the target class. According to this method, a good subset of features is one that contains features that are highly predictive of the target class, while being uncorrelated with each other. The output of the algorithm on the two datasets studied is provided in *table 4.5*. Since this algorithm does not consider the influence of a learning scheme, it is significantly faster to run than the *wrapper subset evaluation* method considered.

Table 4.5 - Best subset of features found for the two datasets studied according to the CFS evaluation algorithm

Dataset 1	Dataset 2
aR, aB, dSat, aA, sA*, aB*, sB*, aU*, dUV, aCont, aCorr, aDiss, rEntro*	*Rect, LInd, aB, aSat, dSat, aB*, sB*, aU*, aV*, sV*, rCont*

In the table above the group of features with the highest value according to the CFS algorithm for the two datasets are presented. It can be observed that the results obtained with this method do not deviate much from the top results found for the ranking of features presented previously. Between the two datasets, it can be observed that the lesion's color computed with simple statistics in the coordinates (a^*, b^*) and (u^*, v^*) and the level of *saturation* within the lesion and its difference to the surrounding skin were present in the best subsets selected. This fact corroborates the suggestion that these color features should to carry significant diagnostic value for the assessment of lesions' malignancy. Regarding the remaining features selected in the best subsets found with this algorithm, with the exception of the *average blue* of the lesion, present in both subsets, they were all part of the highlighted features with the *ranking* methods for each dataset. It can be seen that in the best subset found for *dataset 1,* no shape features were integrated and that major relevance was given to color and texture features, while for *dataset 2*, although being mostly composed of color features, the *lengthening index* and *rectangularity* for shape and *range of Contrast* for texture, highlighted by the previous analysis, were also included by this evaluation. This observation may be justified by the similarity that this method of evaluation shares with the *ranking* criterion used. The CFS algorithm, for each feature added to the subset, measures

the ratio between its correlation to the target class and its correlation with the remaining features of the subset, being this way able to discard redundant features, which the *ranking* methods were not capable of. According to the CFS evaluation, focus is given to the features that are highly predictive of the target class, which had been already determined by the *ranking* methods considered, but not redundant between themselves, resulting in a definite subset of features that is dependent on the search strategy employed. This evaluation was expected to lead to more significant subsets of features that should improve the classification performance. The low number of features present in both of the returned subsets of features should also be emphasized, especially for *dataset 1*, for which the previous observations suggested that a small set of features is unlikely to produce good classification results. The small subset of features obtained might indicate that there was some degree of redundancy between the features that was unaccounted for. However, achieving good classification results with the smallest possible subset of features would be ideal, since it would represent a reduction of the time required to perform the extraction of features from a lesion and also the computational requirements to perform its classification.

The results of applying the returned subsets of features for each dataset to the four classifiers studied can be seen in *table 4.6*.

Table 4.6 - Classification results from four different classifiers, following attribute subset selection using the correlation based feature selection algorithm.

Correlation based feature subset evaluation		Classifier[1]			
		NB	MLP	SVM	J48
Dataset 1	Wrong classifications (M/NM)[2]	24 (9/15)	28 (16/12)	19 (11/8)	34 (16/18)
	Sensitivity (%)	69.00	44.80	62.07	44.80
	Specificity (%)	79.00	83.10	88.70	74.60
Dataset 2	Wrong classifications (M/NM)[2]	23 (7/16)	17 (11/6)	19 (9/10)	28 (15/13)
	Sensitivity (%)	82.50	72.50	77.50	62.50
	Specificity (%)	90.00	96.30	93.80	91.90

[1] Classifiers: Naïve Bayes (NB); Multilayer Perceptron (MLP); Support Vector Machine (SVM); Decision Tree (J48)

[2] Results for this value are presented as T (M/NM): Total number of misclassifications (number of melanomas wrongly classified / number of benign lesions wrongly classified)

The overall behaviour of the four classifiers studied remained consistent with the previous findings. The J48 algorithm performed the worst from the classifiers considered, and the MLP although performing reasonably on *dataset 2,* achieved results similar to the J48 when considering *dataset 1*. The NB and SVM classifiers were the best performing among the four, the former in terms of sensitivity, and the latter in overall accuracy. A significant improvement was found for the use of the NB classifier on *dataset 2*, matching the highest specificity (90%) obtained in previous results and increasing the highest sensitivity (from 80% to 82.50%). Also a slight improvement was found for the application of SVM in this dataset in terms of sensitivity (75% to 77.50%), but the specificity was lower than the previous best found. Although this indicated good value for the application of this selection algorithm, the results obtained with *dataset 1* presented a significant decrease from the best predictive performances that had been observed previously. This finding must be associated to the small number of features included in the subset selected for this dataset, which was expected to be insufficient for achieving reasonable predictions.

- *Wrapper subset evaluator*

The only wrapper method experimented was the *wrapper subset evaluator*. The main difference between this method and the others used is that the wrapper methods perform the selection of a subset of features based on the output of a predetermined machine learning scheme. In this evaluation, a model has to be trained and tested for every increment to the attribute set, and so this was the most time consuming feature selection method applied. However, it has the advantage of determining a group of features that can best perform on a particular machine learning algorithm, and hence should yield better results overall by selecting the most appropriate combinations for each.

The results obtained by this method are summarized in *tables 4.7* and *4.8*.

Table 4.7 - Best subsets of features found following selection using the wrapper subset evaluator

Classifier	Dataset 1	Dataset 2
Naïve Bayes	*C, aSat, dSat, sA*, dAB, sV*, dUV*	*LInd, Asym, dSat, sB*, aV*, sU**
Multilayer Perceptron	*Hu1, Hu6, Hu7, C, S, aSat, sA*, aCorr, rCorr*	*Hu2, Hu7, Rect, LInd, dSat, sB*, aV*, sV**
Support Vector Machine	*Hu5, Hu7, C, sA*, dUV, aCorr, rCorr*	*C, Rect, LInd, dSat, sB*, aU**
J48 decision tree	*aB, sR, aV*, aCorr, rEn, aMax*	*Hu4, LInd, sR*

The results presented in *table 4.7* confirm that the combination of features best suited for one classifier can be entirely different from the best combination for another. It is also possible to see that, consistently with what was observed in the previous selection algorithms applied, the resulting combinations selected are significantly different between datasets. The most significant difference obtained using this method when compared to the previous is the increased importance given to shape features in both datasets. In fact, it can be seen that for both datasets, with the exception of the subsets obtained for the NB classifier, no subsets presented mainly color features, having a balance between the three categories of descriptors for *dataset 1,* and between shape and color for *dataset 2*. The most highlighted shape features that have been disregarded so far were the *Compactness,* especially for the first dataset, and the *7th Hu's invariant moment*. In addition to these, the presence of the *solidity,* the *asymmetry index* and the remaining *Hu's invariants* can also be observed. This is a strong indication that measures that are considered as potentially irrelevant by *ranking* methods should not be disregarded as they may provide useful information if considered with additional information. It may also be observed that texture features, especially the measure of *Correlation,* had a significant influence for the first dataset, while no measure from this category was found in the subsets obtained for the second. The small number of features that was included in every subset obtained should also be emphasized, as it might be associated with the risk that the search strategy employed has of getting stuck in a local optima, hence limiting the number of features that are kept in the resulting set.

Table 4.8 - Classification results from four different classifiers, following attribute subset selection using the wrapper subset evaluator algorithm.

Wrapper subset evaluator		Classifier[1]			
		NB	MLP	SVM	J48
Dataset 1	Wrong classifications (M/NM)[2]	26 (12/14)	29 (15/14)	25 (14/11)	32 (15/17)
	Sensitivity (%)	58.60	48.30	51.72	48.30
	Specificity (%)	80.30	80.30	84.50	76.10
Dataset 2	Wrong classifications (M/NM)[2]	22 (7/15)	21 (11/10)	17 (10/7)	35 (20/15)
	Sensitivity (%)	82.50	72.50	75.00	50.00
	Specificity (%)	90.60	93.80	95.60	90.60

Even though it was expected that this method should lead to the best subset of features for each classifier, translating into an improved prediction performance, it can be seen in *table 4.8* that this did not happen in most situations. Similarly to what was found with the CFS algorithm, a significant improvement from the results obtained with the Naïve Bayes classifier using the *feature ranking* methods was achieved. It was, in fact, slightly better than what was achieved with the CFS algorithm, failing one less benign lesion, and using only 6 features. The SVM was also able to match the best results found for it in the previous evaluations when considering *dataset 2*. These results suggest that a small number of features can be sufficient for achieving reasonable classification performance with this dataset. For *dataset 1,* however, the performance obtained was significantly worse than the previous best found for it considering all four classifiers. This observation, as highlighted in the results obtained with the CFS algorithm, must be related to the small number of features that this method selected for classification in this dataset, which is likely to be caused by the limitations of the search strategy used. With this search, the evaluator starts by considering the empty subset of features and tests the classifier output using each feature alone. The feature that leads to the best classification performance is then kept as the first feature of the subset. The subset is incremented gradually, adding at each step the feature that improves the most the classification result obtained with the previous fixed group. With this strategy, the first feature selected conditions the whole process and it can happen that the prediction rapidly deteriorates when other features are considered with it, hence reaching the stopping criterion with few additions of features and limiting the possibility of finding the best possible result. Other search strategies face similar limitations, which appear to be more aggravated when considering difficult datasets for classification.

The results obtained with the *feature subset evaluation* methods were unexpected, as the evaluation of the features value when grouped should be able to yield better results than when only based on their individual value. Although the CFS algorithm was very efficient regarding the time of computations, the *wrapper* methods took considerably more time to run than the remaining methods studied, for which significant improvements of the results would be required to justify its application. An improvement was observed in the results obtained for *dataset 2,* however a significant decline from the previous best results was found when considering *dataset 1*. The latter is likely to be associated with the increased difficulty that exists for the automatic classification of the samples in this dataset, which is deteriorated even further by the risk of getting stuck in a local maximum inherent to the search strategies available. Nonetheless, the results obtained with this type of evaluation reinforced the suggestion that the use of descriptors from shape, color and texture combined

is important for the problem of automatic classification of skin lesions, and suggested combinations of features that were suited for the context of each dataset.

The results obtained from the feature selection methods considered indicate that the use of *feature ranking* methods is appropriate in this context, since no consistent improvement was achieved in the classification trials following selection by the *subset evaluation* methods. Overall, the results indicated that there is significant value in the features selected and that good classification results can be achieved from them. However, as the limitations of the selection methods were highlighted, the best results found with these were unsatisfactory. To briefly summarize these results, the highest sensitivity (of 82.80%) for *dataset 1* was obtained using a Naïve Bayes classifier on 34 features derived from the ranked list obtained with the *ReliefF* algorithm, with a total of 24 misclassifications; and for *dataset 2* (of 82.50%) was also obtained with the NB, when using the subset of features resultant from the application of the CFS and the wrapper algorithms, within a total of 23 and 22 misclassifications, respectively. Regarding the number of misclassifications, the best result (of 18 missed lesions) for *dataset 1* was obtained using the SVM on a set of 43 features derived from the attributes ranked according to the correlation coefficient, with 69% sensitivity; and for *dataset 2* (of 13 missed lesions) was also obtained using the SVM, but on a set of 20 features retrieved from the *ReliefF* algorithm's output list, achieving 75% sensitivity. The above results indicated that these two machine learning algorithms were the most suited for the classification of the available datasets.

In order to improve the above results and study further the contribution of the proposed features to the problem of skin lesion classification, it was decided to perform an exhaustive search through the feature space for classification using the support vector machine package for Matlab, the LibSVM. The choice of this classifier was based on the results found in (S. Dreiseitl et al., 2001; Torre et al., 2010) and on the stability proved by this classifier across the results of the experimented feature selection methods, achieving superior overall accuracies over every algorithm. It is a very fast algorithm, best designed to deal with binary classification problems. The following section details the steps taken in this process and the results that were obtained with it.

4.3. Exhaustive Search

The classification results that were first obtained after applying the feature selection methods were rather unsatisfactory. This was especially true for the first dataset studied, containing the most difficult lesions, for which even though reasonable sensitivity results were found, the number of misclassifications was too high. It was thought that this should be due to the limitations of the feature selection algorithms used, and so a complete search through the feature space was performed. The main advantage of this strategy is the guarantee of finding the best possible combination of features for classification in a certain dataset. The disadvantages include the often unfeasible time of computation, and also the risk of leading to a combination of features that is very specific to the dataset experimented. Probably because of the latter, the proposed strategy had not yet been applied in the field of pigmented skin lesions. However by saving the combinations of features investigated and the results they produced, a significant amount of information could be created, and it was thought this could increase the knowledge about the features' discriminative potential, help

determining feature combinations that were useful for the problem of skin lesion detection, and also to achieve better classification results.

As previously referred in *section 3.6*, the first step of the search, using only 25 single features/feature groups, was only applied to the *dataset 1,* since it was the first dataset studied for this work, and due to time constraints, *dataset 2* was only used to validate some of the results obtained using the former. In order to accelerate the process of evaluating every combination using a group of 25 features, the iterations were divided in ten executable files, each running one tenth of the total combinations to evaluate. They were run separately in two computers, 5 executable files on each computer simultaneously, and it took approximately five days for this process to be completed. For each file, a text file was created in which the combination of features used, the number of total misclassifications, sensitivity and specificity values achieved were saved in separate lines for each trial.

Although the significant amount of time required to finish the first set of trials, the results obtained from them were already clearly superior to what had been found previously. Regarding the sensitivity, the best results achieved 82.76% with an average of 17 misclassifications; while for the number of misclassifications, the best results failed only a minimum of 12 lesions with an average sensitivity of 76%. This already represents a significant improvement from the results obtained using the SVM in the previous experiments for *dataset 1*, reaching the highest sensitivity that was found using the naïve bayes classifier with a lower number of misclassifications. These results motivated to continue the search by splitting the features that were grouped, in order to understand their individual value to classification, and see if the results could be improved even further.

In order to do the splitting of features, not all results were considered, since it would represent going through all the possible combinations in the end. From the ten text files created, the trials that achieved a number of misclassifications lower than 15 or a sensitivity higher than 80% (which happened in many results with more than 15 misclassifications) were kept separately for further analysis. This resulted in approximately five thousand trials that were considered for further processing. By this stage, however, one limitation of the strategy used became clear, namely the grouping of the seven *Hu's invariant moments*. Although important to reduce the time of computation required for this analysis, the use of the *invariant moments* all together was expectedly incapable of leading to the best results of classification, and this was the case with the five thousand top results extracted, since none of them integrated these features while grouped. Because of the latter, every possible combination of the *invariant moments*, a total of 128, was experimented with each of the top performing trials found previously, and if the results for a trial improved either in terms of the number of misclassifications or sensitivity, the best resulting combination adding these features would be saved instead of the original result. These computations took close to half a day to be completed and it was observed that for roughly 50% of the results considered, improved classification results could be obtained when considering part of the *invariant moments*, mainly in the number of misclassifications, leading to the best result achieved until this stage of 11 misclassifications, with 79.31% sensitivity, and reducing the average number of failed lesions in the results with the highest sensitivity to 15.5.

After completing the previous step, *dataset 2* started being considered in the evaluations. Although performing an exhaustive search in this dataset as well would have been ideal, this was impossible due to time constraints. It was mainly considered in an attempt to validate

the results found for the first dataset, since good performance of these results on a completely different lesion dataset would indicate reasonable generalization performance and help associating high predictive value for the given features. The performance results of the same SVM classifier, with the RBF kernel using the same parameters, were assessed for *dataset 2* using the best feature combinations found for *dataset 1* until this stage. Although around 60% of the results obtained were poor, with more than 25 misclassifications, it was possible to see very good results as well, including a minimum of 10 misclassifications, with 7 being melanomas (82.50% sensitivity), which was already superior to every result found previously for this dataset. These results were surprising considering the heterogeneity between the relevant features determined for each dataset, but were in line with the suggestion that this dataset presents lesions that are more easily differentiated by the automatic methods.

In order to finish the analysis and reveal more singular differences between the relevance of features to the classification of the lesions in the two datasets studied, the grouped features were split in order to search for possible improvements in the classification results. To do this, each feature combination for the best classification results found until then was used. At this stage, approximately five thousand combinations of features were considered for *dataset 1*, and approximately two thousand combinations of features were considered for *dataset 2*, corresponding to all combinations that lead to sensitivity results over 75%. Analyzing each combination of features kept in the top results, if a group of features was being used, this combination would be replicated according to the number of combinations that had to be experimented for that group of features (for example with the *average* and *range* of a texture measure, three combinations equal to the original at that point were created, adding the *average* to one, the *range* to another, and both to the last), repeating this process at each group of features that was encountered, multiplying the number of combinations to evaluate at each step. This resulted in evaluating several million more combinations of features that in the end lead to the results presented next.

After finishing the approximately complete search through the feature space, there were around thirteen thousand results with sensitivity above 80% for *dataset 1*, and thirty thousand for *dataset 2*. It was possible to achieve a maximum of 89.66% sensitivity in the first, within 14 misclassifications (4 missed melanomas); and a maximum of 90.00% sensitivity in the second, within 8 misclassifications (4 missed melanomas). Regarding the number of missed lesions, for the first dataset the best result found failed 9 lesions (with 86.21% sensitivity); and for the second a minimum of 7 lesions (with 87.50% sensitivity). These represent encouraging results to the field of automatic skin lesion classification and reinforce the idea that valuable features were selected to this application. In *figure 4.4*, the percentage of appearance of each feature in the classification results that failed less than 12 lesions in each dataset is summarized by a bar plot. These results obtained over 80% sensitivity in both datasets but do not consider the highest sensitivities found, which occurred often within a higher number of misclassifications. It is possible to observe that the bars in blue correspond to the percentage of appearance of each feature in *dataset 1*, while the red bars correspond to *dataset 2*. The horizontal axis contains the 48 features considered in this work, and the vertical axis represents the percentage of appearance of each feature. The results were ordered according to the percentage of appearance of features in the results for *dataset 1*, to highlight the significant differences that exist between the two datasets.

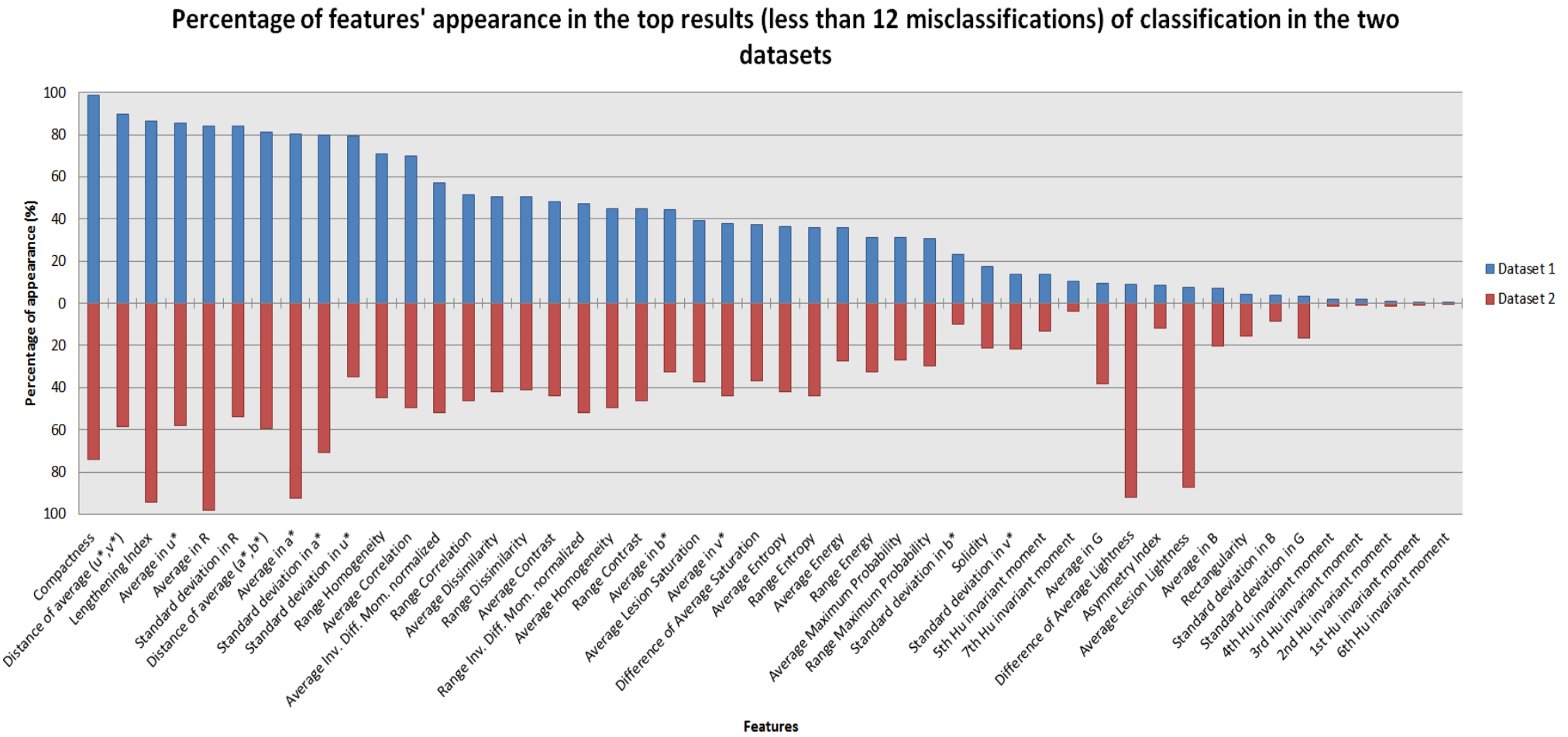


Figure 4.4 - Bar plot summarizing the appearance of each feature in the combinations of features found to achieve less than 12 misclassifications in the classification of both datasets studied. The blue bars are related to the appearance of features in *dataset 1*, the red bars to the appearance in *dataset 2*. The results are ordered according to the percentage of appearance of each feature in the results considered for *dataset 1*.

From *Figure 4.4*, it is possible to understand more clearly the contribution of the selected features to the classification in each of the datasets studied. The very high percentage of appearance of a given feature does not necessarily indicate that it has the highest predictive value, but that it can probably influence the classification results positively in the most number of diverse feature combinations, and should therefore be useful for the automatic classification of pigmented skin lesions. It is also important to refer that although some similarity between features' contribution across the two datasets was expected, since the combinations applied to *dataset 2* were derived from its good results in *dataset 1*, it is believed that the splitting of features as applied allowed achieving significant differences between the two, and allowed inferring about their contributions to each independently. However, this strategy also introduced a significant limitation, because if features were not present in the first best results for *dataset 1,* they would also not be present for *dataset 2,* which may lead to losing some useful information.

Regarding the contribution of shape descriptors, it becomes clear that the *lengthening* and *compactness* indexes were the most participant in the best results in both datasets. In fact, in the group of results considered, these measures were among the most frequently present from all the features considered. This observation highlights the limitations related to selecting features based on their individual value, according to which the contribution of shape was almost always overlooked.

The presence of the *lengthening index* in 94% of the top results of *dataset 2* was not surprising considering the results from the previous analysis, since it figured in the results of every *feature ranking* method and also in the *subset evaluators* used. However, its significant participation in the results of *dataset 1* was unexpected, since it had not been referenced as valuable by any of the *ranking* or *subset evaluators* considered. As to the *compactness index,* it proved to be a valuable measure in the classification of *dataset 1,* participating in 99% of its best results, and was therefore also considered in most results for *dataset 2*. Considering the poor segmentation available for the lesions in this dataset, especially for the melanomas, the *compactness* was not expected to be a valuable measure in this dataset, since many melanomas were considered to have an approximately circular shape. This was likely to be the reason for the significant less contribution this feature provided in this dataset. Nonetheless, due to the good values that were considered in this evaluation, it can be said that the compactness is an important indicator to take into consideration for the automatic detection of melanomas, even when they are in an early stage of development, which was the case for most in *dataset 1*.

The remaining shape descriptors appeared to be much less significant to the classification in the two datasets, not figuring in more than 30% of the top combinations obtained, which is the case for solidity, the third most present shape feature. In relation to the *Hu's invariants*, it can be observed that the 5^{th} and the 7^{th} were the most present, although not participating in much more than 10% of the results. Although this could be due to lack of predictive value of the *moment invariants* for the task considered, the way they were introduced in the combinations could also be the justification for these results, since using them separately from the beginning could have led to different combinations that would integrate them. It should also be noted that a much more significant presence of the *rectangularity* was expected for classification in *dataset 2,* since almost every method studied suggested high predictive value for it in this dataset. However, as can be observed

this measure figured in less than 5% of the best results of *dataset 1*, and so was likely to not be present in a large number of combinations experimented with *dataset 2*. Regarding the *asymmetry index* the results were also surprising considering the contribution of this attribute to the scoring methods used in dermatology. The preliminary evaluations performed showed that this measure did not carry significant predictive value, and was only included in the *subset* obtained for the NB classifier with the *wrapper evaluator* for *dataset 2*. The small influence of this measure could be related to the way it was calculated, but might also be related to the difficult lesions that are present in *dataset 1*, and was also expected to be significantly deteriorated in the lesions of *dataset 2*, due to the poor segmentation considered.

With respect to color, some of the results suggested by the use of the previously discussed methods were corroborated. Regarding the *RGB* color space, a very high participation of the *average* and *standard deviation* of the red component could be observed in the two datasets. The high predictive value for this color channel was also observed in *section 4.1* and suggests that information related to the red component of the *RGB* may be the most important to consider in the differentiation between lesions. The results also suggest that the measurements in the blue and green components play a less significant role in the classification of the available lesions for both datasets, especially for the lesions in the first. In fact for *dataset 2,* the *average lesion green* was present in close to 40% of the results considered while the *average blue* was present in almost 20%, and their *standard deviations* were kept in approximately half of these results, which confirms the prediction that color information is the most relevant for classification using this dataset. The *average lesion saturation*, and the *difference of average saturation* between the lesion and surrounding skin appeared to be the most consistent features between the two datasets, both figuring in approximately 40% of the observed results. The consistency of these features between the two datasets is in agreement to what had been previously seen in the results of most selection algorithms, and suggests that considering color information about the surrounding healthy skin should also be important for this task.

Regarding the measures considered for the $L^*a^*b^*$ and $L^*u^*v^*$ color spaces, the contribution of the simple statistics from the a^* and u^* colorimetric components appeared as the most considerable between the two datasets. The a^* coordinate defines the content of color within the magenta to green scale, which contains the most information regarding the red color. The latter, in agreement to what was observed for the results in the *RGB* space, should be the cause for the increased contribution of this measurement for the results obtained. Significant contribution to the classification results could also be observed for the *distance of the average* (a^*, b^*) and (u^*, v^*) coordinates between the lesion and the surrounding skin, especially in *dataset 1* (participating in approximately 80-85% of the results considered for the first and 60% for the second*)*, which reinforce the suggestion that the color information from the skin surrounding the lesion may carry important diagnostic value. The most surprising and discrepant results among the two datasets are related to the lightness component of these color spaces. While being among the most present features in the combinations considered for *dataset 2*, they were close to non-existent in the best combinations for the first. This is likely to indicate that the lesions of different classes from *dataset 1* present similar levels for these measures and therefore do not carry significant value for classification and distinction between the two. The previous observation may be

related to the lower resolution that these images presented, hence reducing the accuracy with which color information could be measured in these images. The information from the $L^*a^*b^*$ and $L^*u^*v^*$ presented the most relevance from the color features in both datasets, which is an important indication of the advantage of using these color features to deal with the uncontrolled imaging conditions with which different lesions are acquired.

As to the texture features, unlike suggested by the *ranking* methods applied, similar contributions were observed for both datasets. This could be related to their participation in previous results for *dataset 1,* but as was reported previously the classification results that were obtained with *dataset 2* were good, indicating that the texture features carry significant value in the context of this dataset as well. With exception of the *average of Correlation*, and *range of Homogeneity*, which seemed to represent significantly more value for classification in the first dataset, every texture feature experimented presented approximately the same contribution in both datasets, with participations ranging from 30% to 50% of the combinations considered. These results justify the importance of considering texture features in this application, and indicate that measures from the GLCM are suited for this task. Although it has been shown that texture descriptors obtained from the GLCM lead to the best classification performances when extracted from particular dermoscopic structures, the results obtained here suggest that they are also capable of capturing significant differences between lesions when the whole lesion region is considered.

Overall, the results presented suggest a significantly superior contribution from color features to the correct classification of the available lesions. However, they also indicate an undeniable contribution of shape and texture features, as all of the results considered presented some features from both these groups, which suggest that the use of multiple categories of descriptors is essential to achieve the best classification results. Clear differences could be observed between the contributions of the features considered for classification in each dataset, but also significant similarities between the most relevant for both. One important aspect that should be highlighted in the proposed analysis is the randomness of the feature combinations that can achieve reasonable results. As mentioned, the results presented in *figure 4.4* were obtained from approximately thirteen thousand results obtained for *dataset 1,* and thirty thousand for *dataset 2,* and they all achieved sensitivities equal or superior to 80%. This shows that there is a huge variety of combinations of features that can achieve these results in the datasets studied, and it is important to note that most of the features selected, with the few exceptions referred, showed an important contribution to them.

These results were, however, derived from a long list of classification trials, which although containing only reasonable classification results, above 80% sensitivity, they were selected based on the highest classification accuracies that were found, hence disregarding the few top sensitivity results obtained that were usually accompanied by a higher number of misclassifications, especially for *dataset 1*. These results were expected to convey more information about the features that in both datasets were the most related to the presence of malignancy in the lesions, and hence an analysis similar to the previous was performed and its results are presented in *Figure 4.5*. This corresponded to measuring the percentage of appearance of each feature in a total of 213 results for *dataset 1* and 585 for *dataset 2* that achieved sensitivities above 85% (5 or less missed melanomas in each dataset).

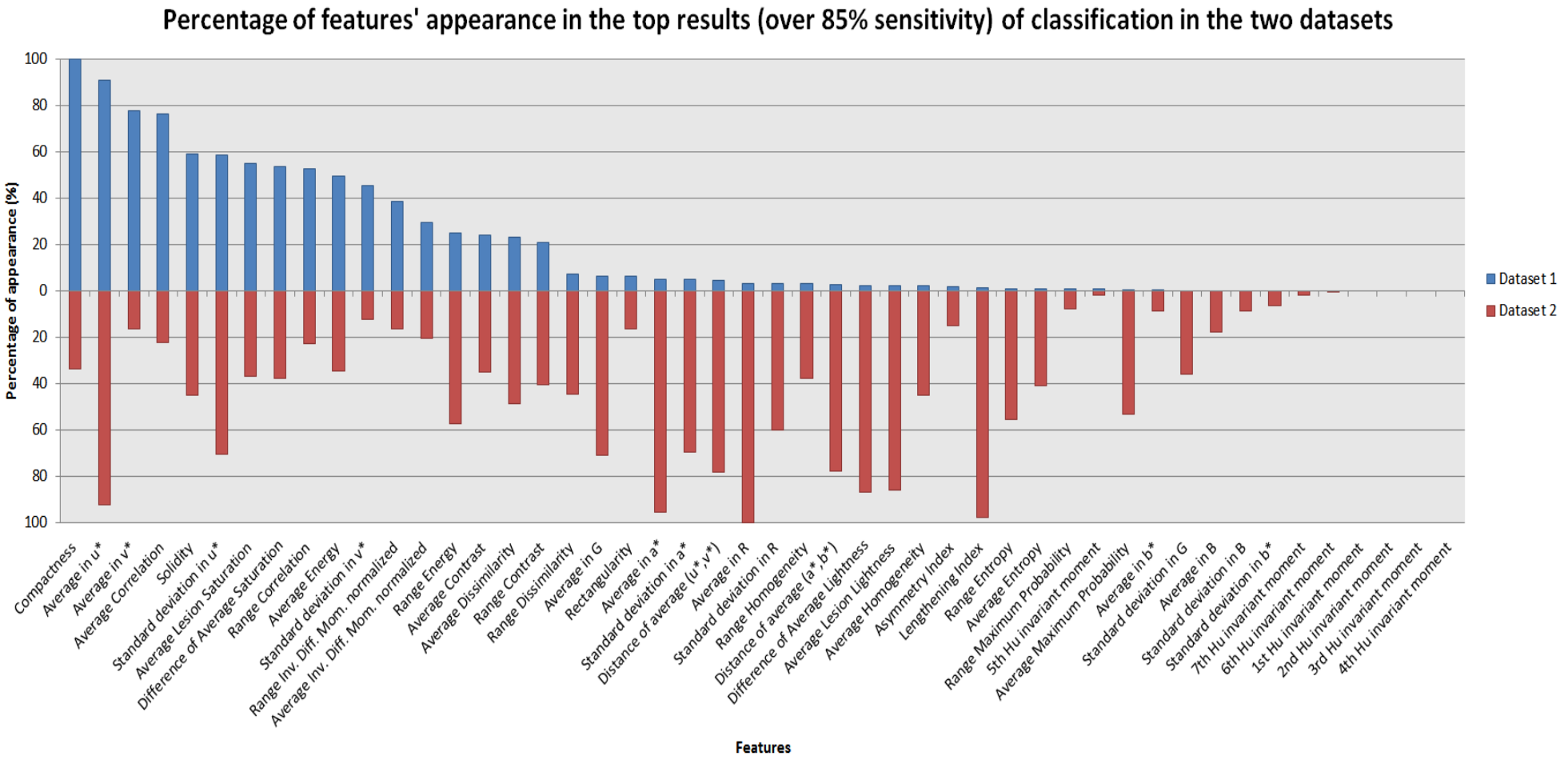


Figure 4.5 - Bar plot summarizing the appearance of each feature in the combinations of features found to achieve above 85% sensitivity in classification of both datasets studied. 213 results were considered for *dataset 1* and 585 for *dataset* 2. The blue bars are related to the first, while the red bars to the second dataset. The results are ordered according to the percentage of appearance of each feature in the results considered for *dataset 1*.

The most important fact that *Figure 4.5* transmits is the big discrepancy that exists between the features that lead to the best detection of melanoma using *dataset 1* and *dataset 2*. With respect to the shape descriptors, it was found that the most important for detecting a melanoma were the *compactness index* for the first dataset and the *lengthening index* for the second. This was expected from what has been presented previously, but it is interesting to see that both were part of practically all the results considered, corroborating the statement that simple shape descriptors provide significant value when considered in combination with additional information. On their counterparts however, it can be seen that the *compactness* participates in approximately 30% of the results for *dataset 2*, and the *lengthening index* figured in almost none of the best results for *dataset 1*, contrarily to what could be expected from *figure 4.4*.

Solidity should also be highlighted as a very important shape descriptor for the detection of melanoma, since it participated in close to 50% of the combinations considered for both datasets. This fact could easily be neglected by looking at the previous results presented, since the presence of *solidity* appeared to not be very significant in any of the datasets. As to the rest of the shape features considered, the results are consistent with what has been seen before. There was almost no contribution from the *Hu's invariant moments*, and the *rectangularity* and *asymmetry index* participated in just approximately 5% and 15% of the best combinations found for *dataset 1* and *2*, respectively.

Regarding color, there are some important observations worth noting. First, the clear presence of the *average* and *standard deviation* of the chromaticity coordinate, u^*, in over 90% and 60% of the results considered for both datasets, respectively. This is an important observation that makes it the color component expected to be the most reliable for detecting malignant melanomas. The v^* chromaticity coordinate also showed to be important for *dataset 1*, but was much less significant when considering the results of *dataset 2*. Regarding the colorimetric coordinates, a^* and b^*, it can be observed that while the latter appeared among the least significant color features for both datasets, the former appeared in almost every combination considered for the second dataset, and in almost none for the first. The measurements related to the *lightness* of the lesion and the surrounding skin appeared consistent with the previous results, figuring in most of the results for the second dataset but showing no contribution to the first dataset. The *distances of the average* (a^*, b^*) and (u^*, v^*) between lesion and skin, however, were among the most relevant for *dataset 2*, while being almost not present for *dataset 1*. This observation would also be neglected when considering *Figure 4.4*, where these measures appeared much more frequently for the first. Although not directly related to the highest sensitivity possible for *dataset 1*, this indicates that there is important value for these measures in both datasets.

With respect to the *RGB* coordinates, a significant difference can also be noted in the importance of the statistics of the *red* color component, being present in every one of the top results for the second set of lesions, and in close to none for the first. This observation reinforces the idea that it is closely related to the information transmitted the by a^*, and hence the similarity between the results obtained for both, disparate from the previous result. It should also be noted the significant presence of the *average* and *standard deviation* of the *green* channel for *dataset 2* only, having participated in 70% and 40% of the results considered for it, respectively. It can be seen that for *dataset 2*, 11 out of the 12 top ranked features by this analysis are related to color, which was expected according to what has been previously observed and highlight the difficulties suggested for the extraction of shape and

texture features in this dataset. Additionally, and in agreement to what had been suggested by the previous analysis, the *average lesion saturation* and the *difference* between the *average lesion saturation* and the *surrounding skin* appeared as the most consistent measures across the two datasets, figuring in 40% of the results for both datasets and proving to have significant value to this context of application.

In terms of texture descriptors, very significant disparities were also observed, which allowed assigning different textures' measurements to being the most related to the melanomas in each dataset studied. While for *dataset 1* only the *average* and *range* of *correlation*, *energy, inverse difference moment normalized* and *contrast* appeared in a significant amount of the top results; for *dataset 2*, it happened for the *average* and *range* of eight measurements considered, with their presence ranging from 25% to close to 60%. Considering what had been observed, texture features were likely to be overlooked for this dataset, while proving here to have an important contribution to the target classification.

The knowledge obtained up to this point indicated that a combination of descriptors from shape, color and texture categories should lead to improved classification performance. In order to confirm the previous statement, the method of the exhaustive search was applied to each of the available feature groups using the same SVM classification setting considered for the previous trials. The results are presented in *table 4.9*.

Table 4.9 - Evaluation of the classification performance of a SVM classifier with RBF kernel (parameters: C=140, γ=0.08) on the two lesion datasets using descriptors from each category alone and using them in combination.

		Feature Category			
		Shape	Color	Texture	Combined
Dataset 1	Wrong classifications (M/NM)[1]	24 (24/0)	20 (13/7)	21 (8/13)	9 (5/4)
	Sensitivity (%)	17.24	55.17	72.41	86.21
	Specificity (%)	100.00	90.14	81.69	92.96
Dataset 2	Wrong classifications (M/NM)[1]	21 (13/8)	13 (7/6)	23 (17/6)	8 (4/4)
	Sensitivity (%)	67.50	82.50	57.50	90.00
	Specificity (%)	95.00	96.25	96.25	97.50

[1] Results for this value are presented as T (M/NM): Total number of misclassifications (number of melanomas wrongly classified / number of benign lesions wrongly classified)

Analyzing the table presented above, the expected behaviour was confirmed. With the extracted features, it was found that using a combination of the shape, color and texture descriptors resulted in a significant improvement of the predictive ability. This is translated in the best result of 9 missed lesions for *dataset 1*, and 8 missed lesions for *dataset 2*, far better than what could be achieved using features from a single category.

Overall the results were significantly worse for *dataset 1*, and it can be seen that the worst performing category of features for this dataset is the shape. By visual inspection of the available lesion, the previous verification was expected because of the significantly irregular and asymmetric shape presented by most benign lesions in this dataset, together with the often regular and approximately circular shape presented for melanomas, for which the shape features were incapable of providing good differentiation if considered individually. Regarding the performance of texture, it can be seen that the use of these features alone allowed correctly detecting the most melanomas in this dataset, leading

however to the highest number of failed benign lesions as well. This indicates that texture is important to the detection of melanoma, but also that the measures selected capture patterns that are not only specific to melanoma, and can fail when presented to benign lesions with irregular textures, if no additional information is considered. The poor performance using only color descriptors in this dataset is assumed to be related to the low resolution of these measures and hence the inaccurate measurements for these features, as was described before.

Regarding the results for *dataset 2*, the observations are in agreement to what has been previously observed. In fact, the use of color features alone can achieve a very reasonable result in this dataset, confirming the high predictive value that has been suggested for them. In relation to the performance of the shape descriptors alone, it can be seen that they can provide significant prediction value. However, this is probably associated to the inaccurate manual borders that were used as the lesions', which for some measures, specifically the *lengthening index* and *rectangularity*, returned significantly different values between the two types of lesions and hence allowing to perform some distinction between them when considered individually. With respect to the texture features, the evaluations performed by the *ranking* methods considered did not assign significant predictive value for these features individually, so it was not surprising that the performance obtained when considering them alone was the worst from the categories considered.

The results presented so far in this section gave a clear idea of the heterogeneity that exists between the lesions of the two datasets used. This is emphasized the most by the results of *Figure 4.5*, where with the exception of very few features; huge discrepancies were observed for their percentage of appearance in the combinations that resulted in the least misclassifications of melanoma. This fact suggests that there might not be many combinations of features that can perform well in both the datasets studied in this work. To evaluate this premise, the lesions from *dataset 1* were classified using the combinations that achieved for *dataset 2* a sensitivity equal or superior to 85% or less than 10 misclassifications, which corresponded to approximately 3000 combinations; while the lesions from the latter were classified using the combinations that achieved a sensitivity superior to 80% or less than 11 misclassifications in *dataset 1*, corresponding to approximately 2500 combinations. The results for this evaluation are presented in *Figure 4.6*, in the form of a box plot that intends to demonstrate the distribution of the results obtained. On the left side of this plot, the distribution of the sensitivity measures when classifying the lesions in *dataset 2* using the best combinations from *dataset 1* is presented; and on the right side is the opposite. The green marks were used to represent average sensitivity for *dataset 2*, and the red marks the average sensitivity for *dataset 1*, using the combinations of features considered in each side of the plot. Each dark gray box represented defines the range where 50% of the results obtained are placed, and the dark line represents the median of these results. The black lines that go outwards from the box represent the values that fall outside its range and go until the maximum and minimum value observed.

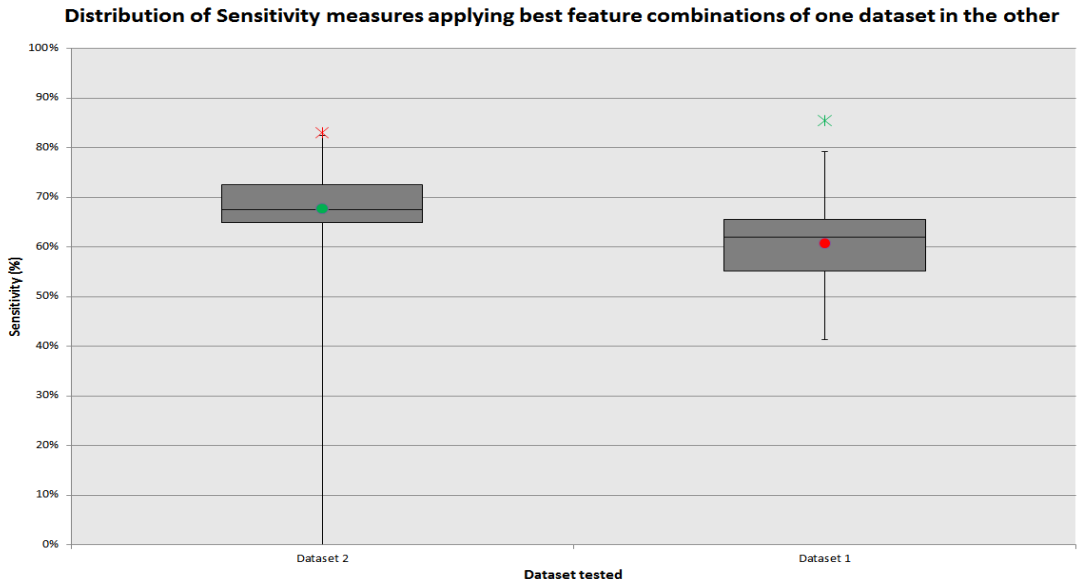


Figure 4.6 - Distribution of the sensitivity performance in the classification of each dataset using the best combinations of features found for the other.

In the results above, it was verified that using the combinations of features that achieved an average sensitivity of 83% when used with dataset 1 could only achieve an average of 68% in dataset 2. On the other hand, the results with an average of 85% sensitivity for *dataset 2* could only achieve an average of 61% when used for *dataset 1*. It can also be observed that the sensitivity values in the second dataset can go as low as 0% to as high as 82.50% when using the combinations of features that best perform on the first, while for the latter results range from approximately 41% to 80% when using the best combinations of features obtained from the second. Although it could be seen that a significant amount of the results achieved for *dataset 2* were below 40% sensitivity, the bulk of its results represented by the box show these can perform slightly better than the overall results using the first dataset, which corroborates the statement that the lesions from *dataset 2* pose a smaller challenge for automatic classification. The results above clearly show how the combinations of features that lead to the least number of missed melanomas in the classification of one dataset are not the combinations that work best in the other, which accentuates the heterogeneity that exists between these datasets and can also indicate that the features selected in this work cannot generalize well to the variety of existing pigmented skin lesions. However, it should also be noticed that the method used to search for the best performing feature combinations could have led to obtaining very specific results that are suited only in the context of the dataset used, and the verification for this should be the poor results that were obtained using the top combinations for one dataset on the alternative.

Although few, some combinations were found that achieve reasonable performance in both datasets, and the top four were summarized in *table 4.10*. The results included in this table represent four of the few combinations of features that are capable of achieving sensitivities equal or above 79%, corresponding to the failure of detection of no more than 7 melanomas out of the 29 available in *dataset 1*, and no more than 8 out of the 40 available in *dataset 2*. As was expected and can be seen, no result rejected the presence of a category of descriptors, and although color is in general more focused in these combinations, a significant weight is laid on shape and texture features as well, as these results intended to demonstrate. From the approximately 45 results found that satisfied the performance

requirements described, the most appearing features were the *lengthening* and *compactness indexes*; the *average* and *standard deviation* in a^*, u^* and *R*, the *distance of average* (u^*, v^*), the *average lesion saturation* and *difference* of *average saturation* and *average lightness* between the *lesion* and the *skin*; the *average* and *range* of *correlation* and the *average of the inverse difference moment normalized*. The only features that did not figure in any of these results were the *4th* and *6th Hu's Invariant moments*, and the *average* and *standard deviation* in *B*.

Table 4.10 - Summary of the combinations of features that achieved the best classification performance using a SVM classifier with RBF kernel (parameters: C=140, γ=0.08) in both lesion datasets studied.

Features selected			Dataset 1			Dataset 2		
Shape	**Color**	**Texture**	**NW (M/NM)[1]**	**SN (%)**	**SP (%)**	**W (M/NM)[2]**	**SN (%)**	**SP (%)**
Hu1, Hu2, Hu5, Hu7, C, S, Rect, LInd	**dL*, aU*, aV*, sU*, dUV**	**rCont, aCorr, rCorr**	10 (6/4)	82.76	92.96	13 (7/6)	82.50	96.25
Hu1, Hu7, C, LInd	**dL*, aU*, aV*, sU*, dUV**	**aCont, aCorr, rCorr**	12 (6/6)	82.76	90.14	12 (8/4)	80.00	97.50
C, Rect, LInd, Asym	**aR, sR, aSat, dSat, aA*, sA*, dAB, aU*, aV*, sU*, sV*, dUV**	**aCont, rCont, aCorr, rCorr, aInv**	13 (7/7)	79.31	90.14	9 (6/3)	85.00	98.13
S, LInd	**aR, aG, sR, sG, aSat, dSat, aL*, aA*, sA*, dL*, aU*, aV*, sU*, sV***	**aEntro, rEntro, aHomo, rHomo, rDiss, aInv**	13 (7/6)	79.31	90.14	10 (5/5)	87.50	96.88

[1] Number of wrongly classified images presented as T (M/NM): Total number of misclassifications (number of melanomas wrongly classified / number of benign lesions wrongly classified);

4.4. Summary

The main results presented in this chapter are summarized in this section. Three analysis steps were proposed to assess the discriminative potential of the features extracted from the pigmented skin lesions available in two image datasets, namely the *feature ranking* methods, the *subset evaluators,* and the *complete search* through the feature space. The biggest issue faced by evaluating the datasets separately was the heterogeneity between the lesions in both, as every method made clear. For the first dataset, every lesion was obtained from scanning dermoscopy slides, and often showed blurring, which mainly affected the color information that could be extracted from them. In the second dataset, many lesions did not fit entirely in the image borders, and hence the segmentation provided for them was poor, which mostly affected the shape and texture features, since both require information present in the border of the lesions.

With the *ranking* methods it was possible to understand that the color features showed the most consistency between datasets in terms of predictive value, having figured in the top

results for both. It was also made clear that the texture features selected were important for the first dataset studied, while showing a small contribution for the second. The lower value that was given to color features in the first dataset was expectedly related to the low resolution these images presented. The low value given to texture features and the high value given to some shape features in this dataset was associated with the poor manual segmentation provided for its lesions. The best classification results obtained using the features selected with the ranking methods showed that the lesions in *dataset 1* were significantly more challenging than the lesions in dataset 2, justified by the higher number of features required to achieve the best results in the first and the poorer results obtained overall.

The use of *subset evaluators* corroborated the suggestion that the lesions from *dataset 1* were significantly more challenging than *dataset 2*, and also indicated that there was significant value in the shape features selected, and that these should be considered in combination with the remaining. The use of these methods to determine a useful subset of features for classification allowed improving the results obtained with *dataset 2,* but led to a significant decline from the best results found with the selection trough ranking, due to the small number of features at which the search strategy terminated.

From the evaluation performed, it was concluded that the best performing *feature selection* algorithms were the *ReliefF* and the *CFS*.

To improve the classification results obtained with the previous methods, and to improve the knowledge about the contributions of the features used to the datasets studied, an approximation to a complete search through the feature space was used, applying an SVM classifier with a RBF-kernel with parameters fixed across all trials to avoid overoptimistic results. The best rate of melanoma detection found in this work was obtained using a combination of shape, color and texture descriptors.

For *dataset 1,* the best results obtained were:

- In terms of sensitivity, 89.66% (4 melanomas missed out of 29) with a total of 14 misclassifications, using 6 features;
- In terms of misclassifications, it was possible to achieve as low as 9 misclassifications, with 86.21% (5 melanomas missed out of the 29), using 17 features.

For *dataset 2*:

- In terms of sensitivity and misclassifications, 90% (4 melanomas missed out of 40) with a total of 8 misclassifications, using 20 features.

It is believed, however, that the optimal result for *dataset 2* might have been missed, since its results were derived from the best performing combinations on *dataset 1*, and the exhaustive search was not applied to this dataset. Nonetheless, the results that could be obtained were good and comparable to results reported by the works reviewed, and lead to a more thorough understanding of the features value to this context of application.

The exhaustive search as performed was considered to be suited for this problem, where a relatively small number of features was selected, but would be prohibitively time consuming if more features were considered.

Chapter 5

Conclusions and Future Perspectives

5.1. Final Conclusions

The use of computers to aid in the biopsy decision making of dermatologists and general practitioners in the analysis of pigmented skin lesions is expected to happen in the near future at a large scale. The results reported so far have been encouraging and showed that automatic algorithms may perform at the level of a dermatologist if the correct information is fed to it. These results were, however, obtained in limited amounts of data and their ability to generalize to unseen lesions is unknown. One common lack of information that was found in almost every research paper considered was the details of the feature selection process, and which features had, in the end, been the most contributing to the results obtained.

Motivated by previous research, this thesis presented a methodological approach to the classification of skin lesions, focusing on three main aspects of the process: the feature extraction, feature selection, and lesion classification. To avoid the errors that could be introduced by automatic segmentation algorithms, the approach proposed in this work was implemented using the manual segmentation results provided with the image datasets. For the first step, a significant group of features was selected, inspired by the scoring algorithms used in dermatology and the previous research on the automated classification of pigmented skin lesions, and extracted from the available lesions. These included shape, color and texture features, in an attempt to characterize the traits that are specific to the malignant skin cancer, namely the asymmetric shape with irregular borders, the presence of multiple colors and the presence of atypical structures. The second and third step were developed conjunctly, with the goal of determining the relevance of the features selected and how they could perform on the image datasets considered.

After conducting this study, it can be concluded that the selection of features for classification has a significant influence on the classification results that can be obtained. The use of a large amount of features does not necessarily translate into increased prediction performance, while the use of a small amount of features is expected to fail in most situations presented. The definition of features that will always work in a given application is very challenging and is dependent on the use of large amounts of examples and a cautious training and testing strategies for the machine learning algorithms.

In this work, two image datasets containing melanocytic lesions were studied. The first consisted of 100 dermoscopic images, containing 29 malignant melanomas and 71 benign lesions. There are two main aspects that should be referred about this dataset. First, the melanomas presented in this dataset were acquired in an early stage, and there was a significant amount of atypical benign nevi. The latter made this dataset to be very challenging regarding the implementation of an automatic classification routine, which was useful to draw important conclusions about the attributes that could be associated with early staged melanoma, when the prognostic of the patients is still very favorable. The second aspect that must be highlighted about this dataset is that the images were obtained from scanning dermoscopic slides, and many images presented significant blurring. This made the color information of these images to contain a lot of noise and hence the color features extracted to be inaccurate. This posed an additional challenge to this dataset and influenced the results that could be obtained.

The second dataset considered consisted of 200 images from the PH^2 database, an open image database containing 40 malignant lesions, 80 common nevi and 80 atypical nevi. One advantage of this database is the fact that it is available, and so the experiments carried out in this work can be used for comparison for the academic society, and to implement and test new features and classification algorithms. However, two limitations were also observed with this dataset. First, the melanomas present in this dataset appeared to be at late stages of development, and hence presented significant differences from the benign lesions available. The second aspect is that the manual segmentation provided is often poor, and many lesions do not fit entirely the image region, which significantly affects the extraction of shape features.

Due to the previous observations, it is clear that the lesions considered were very heterogeneous. The small size of the two image databases, the presence of melanomas on different stages and the limitations present in both groups motivated the use of these datasets separately, as it was predictable that the features could not get good results considering them together. Because of this, this study only focused on performing intra-database classification, applying cross-validation to make the best use of the limited data. However, this must also be referred as an important limitation of the study conducted, since the lesions considered in each classification trial were all from the same dataset, the inevitable similarities they present might have led to overoptimistic results. It also made the analysis of the results more difficult, since the best combinations of features found were significantly different for each and so it was harder to draw consistent conclusions.

Considering the previous limitations, it was observed that the features that could obtain good results with one dataset could completely fail with the other, which emphasizes the dependencies of the classification results on the image databases that are used for training and testing of the classifiers. Nevertheless, it was found that a significant group of features was frequently present in the best results achieved for both datasets, which should therefore have larger influence for the classification and be more useful for the problem of automatic classification of pigmented skin lesions.

Therefore, in spite of the limitations inherent to the adopted approach, this thesis explored one important issue that has been overlooked by the most literature related with the classification of skin lesions, which is the determination of features that are the most relevant in this context of application. The search methods implemented in this work allowed determining features, and also combinations of features that could lead to encouraging

classification results in both datasets, thus being expected to have significant value for automatic melanoma recognition. The first step of evaluation, using the ranking methods, allowed inferring that the category of features that carried the most relevance for classification in both datasets considered was the color, from which the following should be emphasized:

- The average color saturation obtained from the lesion region, and as the difference between the lesion region and the surrounding skin;
- The average and standard deviation of the red color component of the RGB image;
- The average and standard deviation of the (a^*, b^*), and (u^*, v^*) coordinates;

The steps taken further corroborated these assumptions, but also proved the importance of considering them in combination with some of the shape and texture features extracted, and also showed significant value for other color features:

- The compactness index;
- The lengthening index;
- The solidity;
- The distance between the average (a^*, b^*) and (u^*, v^*) obtained from the lesion and the surrounding skin;
- All of the texture features participated significantly in the best results found for one dataset or the other, and hence they were all considered to carry significant diagnostic value.

The information of the asymmetry as measured in this work, the average and standard deviation of blue and green, the Hu's invariant moments and the rectangularity were the features that were considered to have the less value from the initial group considered. The information considered about the lightness (L^*), although being present in the best results obtained with the lesions of dataset 2, did not figure in most of the results obtained with dataset 1 and hence was not highlighted in the group of the most relevant.

Finally, it is thought that the results obtained in this work can be used as guidelines for future developments in this area, and the features highlighted should be considered with priority when testing new classification algorithms or new features for this application.

5.2. Future Work

It is believed that the most important future work improvement would be to consider a larger and more diverse dataset of lesions to validate the results obtained. Although the features used in this work proved to have significant value for the recognition of lesion malignancy, the variety of lesions available was limited and hence these results might not generalize well to other datasets.

Another important improvement that could be applied to the present study would be to use inter-database cross-validation, allowing to make the most use out of the data available and to report the results on images obtained under the most diverse acquisition settings possible. This would implicate using lesions mixed from different datasets to train the

classification models and test them in lesions obtained from other datasets also, which should enhance the information obtained about the generalization ability of the features used.

It would also be important to evaluate the validity of the results obtained in automatically segmented images, to see if the loss of accuracy in the detection of the lesion's region leads to worse classification results.

One interesting approach that could be added to the present work would be to evaluate the performance of the simple features considered to a database of common digital cameras, for which if good results could be found, and considering the level of resolution that can be achieved with the smartphones' cameras, the patient self-examination could be significantly improved and affordable.

Regarding the exploration of new features and algorithms in the context of the skin lesion, it is believed that the exploration of segmentation algorithms to automate the detection of dermoscopic structures should be the most promising route to take for improving the diagnostic accuracy of the CAD systems in dermatology.

Bibliography

Abbasi, N. R., Shaw, H. M., Rigel, D. S., Friedman, R. J., McCarthy, W. H., Osman, I., . . . Polsky, D. (2004). Early diagnosis of cutaneous melanoma: revisiting the ABCD criteria. *JAMA, 292*(22), 2771-2776. doi:10.1001/jama.292.22.2771

Aberg, P., Nicander, I., Hansson, J., Geladi, P., Holmgren, U., & Ollmar, S. (2004). Skin cancer identification using multifrequency electrical impedance--a potential screening tool. *IEEE Trans Biomed Eng, 51*(12), 2097-2102. doi:10.1109/TBME.2004.836523

American Cancer Society. (2015a). Cancer Facts & Figures, 2015. Atlanta: American Cancer Society.

American Cancer Society. (2015b, 20/03/2015). What is Melanoma Skin cancer? Retrieved at July 25, 2015 from http://www.cancer.org/cancer/skincancer-melanoma/detailedguide/melanoma-skin-cancer-what-is-melanoma

Andreassi, L., Perotti, R., Rubegni, P., Burroni, M., Cevenini, G., Biagioli, M., . . . Barbini, P. (1999). Digital dermoscopy analysis for the differentiation of atypical nevi and early melanoma: a new quantitative semiology. *Archives of dermatology, 135*(12), 1459-1465.

Andreassi, L., Perotti, R., Rubegni, P., Burroni, M., Cevenini, G., Biagioli, M., . . . Barbini, P. (1999). Digital dermoscopy analysis for the differentiation of atypical nevi and early melanoma: a new quantitative semiology. *Arch Dermatol, 135*(12), 1459-1465. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10606050

Argenziano, G., Fabbrocini, G., Carli, P., De Giorgi, V., Sammarco, E., & Delfino, M. (1998). Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions. Comparison of the ABCD rule of dermatoscopy and a new 7-point checklist based on pattern analysis. *Arch Dermatol, 134*(12), 1563-1570. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9875194

Argenziano, G., Soyer, H. P., Chimenti, S., Talamini, R., Corona, R., Sera, F., . . . Kopf, A. W. (2003). Dermoscopy of pigmented skin lesions: results of a consensus meeting via the Internet. *J Am Acad Dermatol, 48*(5), 679-693. doi:10.1067/mjd.2003.281

Argenziano, G., Soyer, H. P., DeGiorgi, V., Piccolo, D., Delfino, M., & al., e. (2002). Dermoscopy: A Tutorial. Milan: Medical Publishing&NewMedia.

Barata, C., Ruela, M., Francisco, M., Mendonça, T., & Marques, J. S. (2014). Two systems for the detection of melanomas in dermoscopy images using texture and color features. *Systems Journal, IEEE, 8*(3), 965-979.

Bishop, C. M. (2006). *Pattern recognition and machine learning*: springer.

Braun, R., Gaide, O., Oliviero, M., Kopf, A., French, L., Saurat, J. H., & Rabinovitz, H. (2007). The significance of multiple blue-grey dots (granularity) for the dermoscopic diagnosis of melanoma. *British Journal of Dermatology, 157*(5), 907-913.

Braun, R. P., Rabinovitz, H. S., Oliviero, M., Kopf, A. W., & Saurat, J. H. (2005). Dermoscopy of pigmented skin lesions. *J Am Acad Dermatol, 52*(1), 109-121. doi:10.1016/j.jaad.2001.11.001

Celebi, M. E., Aslandogan, Y. A., & Bergstresser, P. R. (2005). *Unsupervised border detection of skin lesion images*. Paper presented at the Information Technology: Coding and Computing, 2005. ITCC 2005. International Conference on.

Celebi, M. E., Iyatomi, H., Stoecker, W. V., Moss, R. H., Rabinovitz, H. S., Argenziano, G., & Soyer, H. P. (2008). Automatic detection of blue-white veil and related structures in dermoscopy images. *Comput Med Imaging Graph, 32*(8), 670-677. doi:10.1016/j.compmedimag.2008.08.003

Celebi, M. E., Iyatomi, H., Stoecker, W. V., Moss, R. H., Rabinovitz, H. S., Argenziano, G., & Soyer, H. P. (2008). Automatic detection of blue-white veil and related structures in dermoscopy images. *Computerized Medical Imaging and Graphics, 32*(8), 670-677.

Celebi, M. E., Kingravi, H. A., Uddin, B., Iyatomi, H., Aslandogan, Y. A., Stoecker, W. V., & Moss, R. H. (2007). A methodological approach to the classification of dermoscopy images. *Computerized Medical Imaging and Graphics, 31*(6), 362-373.

Celebi, M. E., Kingravi, H. A., Uddin, B., Iyatomi, H., Aslandogan, Y. A., Stoecker, W. V., & Moss, R. H. (2007). A methodological approach to the classification of dermoscopy images. *Comput Med Imaging Graph, 31*(6), 362-373. doi:10.1016/j.compmedimag.2007.01.003

Celenk, M. (1990). A color clustering technique for image segmentation. *Computer Vision, Graphics, and Image Processing, 52*(2), 145-170.

Chan, H. P., Hadjiiski, L., Zhou, C., & Sahiner, B. (2008). Computer-aided diagnosis of lung cancer and pulmonary embolism in computed tomography-a review. *Acad Radiol, 15*(5), 535-555. doi:10.1016/j.acra.2008.01.014

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST), 2*(3), 27.

Cheng, Y., Swamisai, R., Umbaugh, S. E., Moss, R. H., Stoecker, W. V., Teegala, S., & Srinivasan, S. K. (2008). Skin lesion classification using relative color features. *Skin Res Technol, 14*(1), 53-64. doi:10.1111/j.1600-0846.2007.00261.x

Cheng, Y. I., Swamisai, R., Umbaugh, S. E., Moss, R. H., Stoecker, W. V., Teegala, S., & Srinivasan, S. K. (2008). Skin lesion classification using relative color features. *Skin Research and Technology, 14*(1), 53-64.

Clark, D. E. (1997). Computational methods for probabilistic decision trees. *Computers and biomedical research, 30*(1), 19-33.

Clausi, D. A. (2002). An analysis of co-occurrence texture statistics as a function of grey level quantization. *Canadian Journal of remote sensing, 28*(1), 45-62.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*: Cambridge university press.

Curiel-Lewandrowski, C. (2015). Risk factors for the development of melanoma. In P. T. (Ed.) (Ed.), *UpToDate*. UpToDate, Waltham, MA. (Accessed on August 10, 2015).

Dreiseitl, S., Ohno-Machado, L., Kittler, H., Vinterbo, S., Billhardt, H., & Binder, M. (2001). A comparison of machine learning methods for the diagnosis of pigmented skin lesions. *J Biomed Inform, 34*(1), 28-36. doi:10.1006/jbin.2001.1004

Dreiseitl, S., Ohno-Machado, L., Kittler, H., Vinterbo, S., Billhardt, H., & Binder, M. (2001). A comparison of machine learning methods for the diagnosis of pigmented skin lesions. *Journal of biomedical informatics, 34*(1), 28-36.

Elbaum, M., Kopf, A. W., Rabinovitz, H. S., Langley, R. G., Kamino, H., Mihm, M. C., Jr., . . . Wang, S. (2001). Automatic differentiation of melanoma from melanocytic nevi with multispectral digital dermoscopy: a feasibility study. *J Am Acad Dermatol, 44*(2), 207-218. doi:10.1067/mjd.2001.110395

Elgamal, M. (2013). Automatic skin cancer images classification. *IJACSA) International Journal of Advanced Computer Science and Applications, 4*(3).

Elter, M., & Horsch, A. (2009). CADx of mammographic masses and clustered microcalcifications: a review. *Med Phys, 36*(6), 2052-2068. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19610294

Emre Celebi, M., Kingravi, H. A., Iyatomi, H., Alp Aslandogan, Y., Stoecker, W. V., Moss, R. H., . . . Rabinovitz, H. S. (2008). Border detection in dermoscopy images using statistical region merging. *Skin Research and Technology, 14*(3), 347-353.

Emre Celebi, M., Wen, Q., Hwang, S., Iyatomi, H., & Schaefer, G. (2013). Lesion border detection in dermoscopy images using ensembles of thresholding methods. *Skin Research and Technology, 19*(1), e252-e258.

Erkol, B., Moss, R. H., Joe Stanley, R., Stoecker, W. V., & Hvatum, E. (2005). Automatic lesion boundary detection in dermoscopy images using gradient vector flow snakes. *Skin Research and Technology, 11*(1), 17-26.

Feit, N. E., Dusza, S. W., & Marghoob, A. A. (2004). Melanomas detected with the aid of total cutaneous photography. *Br J Dermatol, 150*(4), 706-714. doi:10.1111/j.0007-0963.2004.05892.x

Fikrle, T., & Pizinger, K. (2007). Digital computer analysis of dermatoscopical images of 260 melanocytic skin lesions; perimeter/area ratio for the differentiation between malignant melanomas and melanocytic nevi. *J Eur Acad Dermatol Venereol, 21*(1), 48-55. doi:10.1111/j.1468-3083.2006.01864.x

Fikrle, T., & Pizinger, K. (2007). Digital computer analysis of dermatoscopical images of 260 melanocytic skin lesions; perimeter/area ratio for the differentiation between malignant melanomas and melanocytic nevi. *Journal of the European Academy of Dermatology and Venereology, 21*(1), 48-55.

Ford, A., & Roberts, A. (1998). Colour space conversions. *Westminster University, London, 1998*, 1-31.

Friedman, R. J., Rigel, D. S., & Kopf, A. W. (1985). Early detection of malignant melanoma: the role of physician examination and self-examination of the skin. *CA Cancer J Clin, 35*(3), 130-151. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/3921200

Ganster, H., Pinz, A., Rohrer, R., Wildling, E., Binder, M., & Kittler, H. (2001). Automated melanoma recognition. *IEEE Trans Med Imaging, 20*(3), 233-239. doi:10.1109/42.918473

Ganster, H., Pinz, A., Röhrer, R., Wildling, E., Binder, M., & Kittler, H. (2001). Automated melanoma recognition. *Medical Imaging, IEEE Transactions on, 20*(3), 233-239.

Getreuer, P. (2010, 14 Jan, 2011). Colorspace Transformations. Retrieved at 15 August, 2015 from http://www.mathworks.com/matlabcentral/fileexchange/28790-colorspace-transformations

Gómez, D. D., Butakoff, C., Ersbøll, B. K., & Stoecker, W. (2008). Independent histogram pursuit for segmentation of skin lesions. *Biomedical Engineering, IEEE Transactions on, 55*(1), 157-161.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research, 3*, 1157-1182.

Hall, M. (2000). *ªCorrelation Based Feature Selection for Discrete and Numeric Class Machine Learning*. Paper presented at the ° Proc. 17th Int'l. Conf. Machine Learning.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter, 11*(1), 10-18.

Hall, P., Claridge, E., & Smith, J. M. (1995). Computer screening for early detection of melanoma—is there a future? *British Journal of Dermatology, 132*(3), 325-338.

Haralick, R. M., Shanmugam, K., & Dinstein, I. H. (1973). Textural features for image classification. *Systems, Man and Cybernetics, IEEE Transactions on*(6), 610-621.

Henning, J. S., Dusza, S. W., Wang, S. Q., Marghoob, A. A., Rabinovitz, H. S., Polsky, D., & Kopf, A. W. (2007). The CASH (color, architecture, symmetry, and homogeneity) algorithm for dermoscopy. *J Am Acad Dermatol, 56*(1), 45-52. doi:10.1016/j.jaad.2006.09.003

Hu, M.-K. (1962). Visual pattern recognition by moment invariants. *Information Theory, IRE Transactions on, 8*(2), 179-187.

Hunt, R. W. G., & Pointer, M. R. (2011). *Measuring colour*: John Wiley & Sons.

Japkowicz, N. (2000). *Learning from imbalanced data sets: a comparison of various strategies*. Paper presented at the AAAI workshop on learning from imbalanced data sets.

Kira, K., & Rendell, L. A. (1992). *The feature selection problem: Traditional methods and a new algorithm*. Paper presented at the AAAI.

Kittler, H., Pehamberger, H., Wolff, K., & Binder, M. (2002). Diagnostic accuracy of dermoscopy. *Lancet Oncol, 3*(3), 159-165. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11902502

Kononenko, I., & Simec, E. (1995). *Induction of decision trees using RELIEFF*. Paper presented at the Proceedings of the ISSEK94 Workshop on Mathematical and Statistical Methods in Artificial Intelligence.

Korotkov, K., & Garcia, R. (2012). Computerized analysis of pigmented skin lesions: a review. *Artif Intell Med, 56*(2), 69-90. doi:10.1016/j.artmed.2012.08.002

Loane, M. A., Gore, H. E., Corbett, R., Steele, K., Mathews, C., Bloomer, S. E., . . . Wootton, R. (1997). Effect of camera performance on diagnostic accuracy: preliminary results from the Northern Ireland arms of the UK Multicentre Teledermatology Trial. *J Telemed Telecare, 3*(2), 83-88. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9206278

Lorber, A., Wiltgen, M., Hofmann-Wellenhof, R., Koller, S., Weger, W., Ahlgrimm-Siess, V., . . . Gerger, A. (2009). Correlation of image analysis features and visual morphology in melanocytic skin tumours using in vivo confocal laser scanning microscopy. *Skin Res Technol, 15*(2), 237-241. doi:10.1111/j.1600-0846.2009.00361.x

Lorentzen, H., Weismann, K., Petersen, C. S., Larsen, F. G., Secher, L., & Skodt, V. (1999). Clinical and dermatoscopic diagnosis of malignant melanoma. Assessed by expert and non-expert groups. *Acta Derm Venereol, 79*(4), 301-304. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10429989

Ma, Z., & Tavares, J. (2015). A Novel Approach to Segment Skin Lesions in Dermoscopic Images Based on a Deformable Model. *IEEE J Biomed Health Inform*. doi:10.1109/JBHI.2015.2390032

Ma, Z., & Tavares, J. M. R. (2015). A Novel Approach to Segment Skin Lesions in Dermoscopic Images Based on a Deformable Model.

MacKie, R. M. (1990). Clinical recognition of early invasive malignant melanoma. *BMJ, 301*(6759), 1005-1006. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2249043

MacNeal, R. J. (2015). Structure and Function of the Skin. *Merck Manual Online*. Retrieved at July 28, 2015 from http://www.merckmanuals.com/home/skin-disorders/biology-of-the-skin/structure-and-function-of-the-skin

Maglogiannis, I., & Doukas, C. N. (2009). Overview of advanced computer vision systems for skin lesions characterization. *IEEE Trans Inf Technol Biomed, 13*(5), 721-733. doi:10.1109/TITB.2009.2017529

Maglogiannis, I., & Doukas, C. N. (2009). Overview of advanced computer vision systems for skin lesions characterization. *Information Technology in Biomedicine, IEEE Transactions on, 13*(5), 721-733.

Maglogiannis, I., Zafiropoulos, E., & Kyranoudis, C. (2006). Intelligent segmentation and classification of pigmented skin lesions in dermatological images *Advances in Artificial Intelligence* (pp. 214-223): Springer.

Marghoob, A., & Jaimes, N. (2015). Dermoscopic evaluation of skin lesions. In P. T. (Ed). (Ed.), *UpToDate*. UpToDate, Waltham, MA. (Accessed on August 10, 2015).

Materka, A., & Strzelecki, M. (1998). Texture analysis methods–a review. *Technical university of lodz, institute of electronics, COST B11 report, Brussels*, 9-11.

Mathworks, R. a. (2015a). "graycomatrix". Retrieved at 15 August, 2015 from http://www.mathworks.com/help/images/ref/graycomatrix.html

Mathworks, R. a. (2015b). "regionprops". Retrieved at 15 August, 2015 from http://www.mathworks.com/help/images/ref/regionprops.html

Mathworks, R. a. (2015c). "rgb2gray". Retrieved at 15 August, 2015 from http://www.mathworks.com/help/matlab/ref/rgb2gray.html

Menzies, S. W. (1999). Automated epiluminescence microscopy: human vs machine in the diagnosis of melanoma. *Arch Dermatol, 135*(12), 1538-1540. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10606065

Menzies, S. W., Crotty, K. A., Ingvar, C., & McCarthy, W. H. (2003). *An atlas of surface microscopy of pigmented skin lesions: dermoscopy*: McGraw-Hill Roseville.

Messadi, M., Bessaid, A., & Taleb-Ahmed, A. (2009). Extraction of specific parameters for skin tumour classification. *J Med Eng Technol, 33*(4), 288-295. doi:10.1080/03091900802451315

Nachbar, F., Stolz, W., Merkle, T., Cognetta, A. B., Vogt, T., Landthaler, M., . . . Plewig, G. (1994). The ABCD rule of dermatoscopy. High prospective value in the diagnosis of

doubtful melanocytic skin lesions. *J Am Acad Dermatol, 30*(4), 551-559. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8157780

National Cancer Institute. (2015). SEER Cancer Statistics Factsheets: Melanoma of the Skin. Retrieved at August 10, 2015 from http://seer.cancer.gov/statfacts/html/melan.html

Ng, V. T., Fung, B. Y., & Lee, T. K. (2005). Determining the asymmetry of skin lesion with fuzzy borders. *Comput Biol Med, 35*(2), 103-120. doi:10.1016/j.compbiomed.2003.11.004

Ng, V. T., Fung, B. Y., & Lee, T. K. (2005). Determining the asymmetry of skin lesion with fuzzy borders. *Computers in biology and medicine, 35*(2), 103-120.

Oakley, A. (2008, 10 November, 2014). Benign melanocytic lesions. *Common skin lesions.* Retrieved at 28 July, 2015 from http://www.dermnetnz.org/doctors/lesions/melanocytic.html

Oates, T., & Jensen, D. (1998). *Large Datasets Lead to Overly Complex Models: An Explanation and a Solution.* Paper presented at the KDD.

Ojala, T., Pietikäinen, M., & Mäenpää, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 24*(7), 971-987.

Oliveira, R. B., Tavares, J. M. R., Marranghello, N., & Pereira, A. S. (2013). An Approach to Edge Detection in Images of Skin Lesions by Chan-Vese Model.

Pellacani, G., Cesinaro, A. M., & Seidenari, S. (2005). Reflectance-mode confocal microscopy of pigmented skin lesions--improvement in melanoma diagnostic specificity. *J Am Acad Dermatol, 53*(6), 979-985. doi:10.1016/j.jaad.2005.08.022

Peura, M., & Iivarinen, J. (1997). *Efficiency of simple shape descriptors.* Paper presented at the Proceedings of the third international workshop on visual form.

Pleiss, C., Risse, J. H., Biersack, H. J., & Bender, H. (2007). Role of FDG-PET in the assessment of survival prognosis in melanoma. *Cancer Biother Radiopharm, 22*(6), 740-747. doi:10.1089/cbr.2006.356

Premkumar, A., Marincola, F., Taubenberger, J., Chow, C., Venzon, D., & Schwartzentruber, D. (1996). Metastatic melanoma: correlation of MRI characteristics and histopathology. *J Magn Reson Imaging, 6*(1), 190-194. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8851427

Rigel, D. S. (1992). The dilemma of the dysplastic nevus. *Annals of plastic surgery, 28*(1), 9-10.

Rogers, H. W., Weinstock, M. A., Harris, A. R., Hinckley, M. R., Feldman, S. R., Fleischer, A. B., & Coldiron, B. M. (2010). Incidence estimate of nonmelanoma skin cancer in the United States, 2006. *Arch Dermatol, 146*(3), 283-287. doi:10.1001/archdermatol.2010.19

Ruela, M., Barata, C., Mendonca, T., & Marques, J. S. (2013). *On the role of shape in the detection of melanomas.* Paper presented at the Image and Signal Processing and Analysis (ISPA), 2013 8th International Symposium on.

Ruiz, D., Berenguer, V., Soriano, A., & SáNchez, B. (2011). A decision support system for the diagnosis of melanoma: A comparative approach. *Expert Systems with Applications, 38*(12), 15217-15223.

Schaffer, J. V., & Bolognia, J. L. (2014a). Acquired melanocytic nevi (moles). In P. T. (Ed.). (Ed.), *UpToDate*. UpToDate, Waltham, MA. (Accessed on August 10, 2015).

Schaffer, J. V., & Bolognia, J. L. (2014b). Congenital melanocytic nevi. In P. T. (Ed). (Ed.), *UpToDate*. UpToDate, Waltham, MA. (Accessed on August 10, 2015).

Schmid, P. (1999). *Lesion detection in dermatoscopic images using anisotropic diffusion and morphological flooding.* Paper presented at the Image Processing, 1999. ICIP 99. Proceedings. 1999 International Conference on.

Schwarz, D. (2010). "Fast and Robust Curve Intersections". Retrieved at 15 August, 2015 from http://www.mathworks.com/matlabcentral/fileexchange/11837-fast-and-robust-curve-intersections/content/intersections.m

Sezgin, M. (2004). Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic imaging, 13*(1), 146-168.

Shrestha, B., Bishop, J., Kam, K., Chen, X., Moss, R. H., Stoecker, W. V., . . . Soyer, H. P. (2010). Detection of atypical texture features in early malignant melanoma. *Skin Res Technol, 16*(1), 60-65. doi:10.1111/j.1600-0846.2009.00402.x

Skrovseth, S., Schopf, T. R., Thon, K., Zortea, M., Geilhufe, M., Mollersen, K., . . . Godtliebsen, F. (2010). *A computer aided diagnostic system for malignant melanomas*. Paper presented at the 2010 3rd International Symposium on Applied Sciences in Biomedical and Communication Technologies (ISABEL 2010).

Soyer, H. P., Argenziano, G., Chimenti, S., & Ruocco, V. (2001). Dermoscopy of pigmented skin lesions. *Eur J Dermatol, 11*(3), 270-276; quiz 277. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11358742

Stoecker, W. V., Wronkiewiecz, M., Chowdhury, R., Stanley, R. J., Xu, J., Bangert, A., . . . Oliviero, M. (2011). Detection of granularity in dermoscopy images of malignant melanoma using color and texture features. *Computerized Medical Imaging and Graphics, 35*(2), 144-147.

Stoecker, W. V., Wronkiewiecz, M., Chowdhury, R., Stanley, R. J., Xu, J., Bangert, A., . . . Drugge, R. (2011). Detection of granularity in dermoscopy images of malignant melanoma using color and texture features. *Comput Med Imaging Graph, 35*(2), 144-147. doi:10.1016/j.compmedimag.2010.09.005

Swetter, S., & Geller, A. C. (2014). Skin examination and clinical features of melanoma. In P. T. (Ed.) (Ed.), *UpToDate*. UpToDate, Waltham, MA. (Accessed on August 10, 2015).

Tang, J., Alelyani, S., & Liu, H. (2014). Feature selection for classification: A review. *Data Classification: Algorithms and Applications*, 37.

Torre, E. L., Caputo, B., & Tommasi, T. (2010). Learning methods for melanoma recognition. *International Journal of Imaging Systems and Technology, 20*(4), 316-322.

Umbaugh, S. E., Moss, R. H., & Stoecker, W. V. (1991). Applying artificial intelligence to the identification of variegated coloring in skin tumors. *Engineering in Medicine and Biology Magazine, IEEE, 10*(4), 57-62.

Vennila, G. S., Suresh, L. P., & Shunmuganathan, K. (2012). *Dermoscopic image segmentation and classification using machine learning algorithms*. Paper presented at the Computing, Electronics and Electrical Technologies (ICCEET), 2012 International Conference on.

Yuan, X., Yang, Z., Zouridakis, G., & Mullani, N. (2006). SVM-based texture classification and application to early melanoma detection. *Conf Proc IEEE Eng Med Biol Soc, 1*, 4775-4778. doi:10.1109/IEMBS.2006.260056