COLOR DETECTION IN DERMOSCOPIC IMAGES OF PIGMENTED SKIN LESIONS THROUGH COMPUTER VISION TECHNIQUES

Bachelor's Thesis

University Carlos III of Madrid

Author: Irene Hernández Serrano.

Tutor: Javier López Labraca.

Audiovisual Systems Engineering.

February 2017.

INDEX

INDEX OF FIGURES

INDEX OF TABLES

INDEX OF EQUATIONS

1. INTRODUCTION. PROJECT MOTIVATION AND OBJECTIVES

This thesis offers an insight into skin cancer detection, focusing on the extraction of distinct features (color, namely) from potential melanoma lesions. The following document provides an outlook of melanoma analysis, as well as experimental results based on Matlab implementations.

The relevance of the work carried out throughout this project resides in the specificity of the study: color is a key characteristic in melanoma inspection. It is usually linked to pattern analysis but seldom the sole object of research. Most lines of work in the field of skin cancer diagnosis associate color with other features such as texture, shape, asymmetry or pattern of the lesion.

Studies cement this belief regarding the vital significance of color, as the number of colors in a lesion happens to be the most significant biomarker for determining malignancy [1].
Different image processing techniques will be applied to build statistical models that shape the outcome of the prospective diagnosis.
The purpose of the project is the development of an assisting tool able to detect the most prevalent colors in skin pigmented lesions, in order to give a probabilistic result. The strength of this idea lies in the resemblance to actual medical procedures; dermatologists examine color to diagnose melanoma. Simulating medical proceedings is a burgeoning trend in CAD systems [2] because it renders the advancements in this field more likely to be accepted by the medical community.

An additional motivation comes from real-life statistics: skin cancer is, by far, the most frequent type of cancer. Moreover, although melanoma is the least common form of skin cancer at only around 1% of all cases, the majority of deaths related to skin cancer are due to melanoma [3]. Furthermore, the rate of melanoma occurrence is particularly high in Spain and has significantly increased in the last decade [4], hence the importance of reliable diagnosis that is not exclusively contingent on the specialist's subjective judgment.

2. REGULATORY FRAMEWORK

The legislative context relevant to the work carried out for this project concerns data protection laws. In Spain, such a regulation law is typified in the Organic Law on Personal Data (*Ley Orgánica 15/1999 de Protección de Datos de Carácter Personal, LOPD*). This Organic Law ensures and protects the liberties and fundamental rights in the matter of personal data processing [5].
Within the aforementioned text, cession of data is defined in article 3.i as any disclosure of data to a person other than the data subject. The terms for the communication or assignment of personal information are established in article 11, according to which personal data can only be used by third parties for objectives directly connected to the conceded functions.
The fulfillment of data protection legislations is guaranteed, nationally, by a public institution known as AEGP (Agencia Española de Protección de Datos).

On an European level, common EU rules have been established to secure personal data anywhere in the European Union. The European Commission presented a reform in this area, mostly to promote the Digital Single Market, a key goal for the Commission that aims to merge European markets into one. The Regulation and Directive that constitute the reform shall be effectively implemented in all pertinent countries in 2018, by mid-year at the latest [6].

For the set of data used in this project, the PH^2 database, consent is mandatory: every individual whose lesion is featured in the collection of dermoscopic images had to sign an agreement with those responsible for the creation of the dataset.

As disclosed on the official website for the PH^2 database, the available data can be used for research and educational reasons, but not for commercial or redistribution operations [7].

3. PROJECT SCHEDULING

The project began in June of 2016 and concluded on February of 2017.

During the initialization of the project, after introduction of its general idea, the very beginning involved reading documentation provided by the tutor about the subject, to learn the related concepts and current state-of-the-art. This was followed by a meeting to establish the scope of the project. Next, familiarization with the database was necessary, which entailed the first part of Matlab code development, as to create a workspace of accessible information.
Once experiments started, the organization of tasks was greatly influenced by the progress in functionality implementations using Matlab.

Then, the segmentation phase ensued. K-means was the first implemented algorithm, starting with a fixed K parameter and elemental settings, progressively increasing the complexity of the exercise. After that, focus was shifted to investigate the pre-processing step.
Next, two other segmentation procedures (N-cuts and mean shift) became the task at hand, continuing with feature extraction afterwards.
At that point, the data was ready to be classified. The classification endeavor was divided into training and testing. First, a multivariate Gaussian model was used. Once it was working, the classifier evolved and a Gaussian Mixture model was utilized instead. A toy or sample data set with well-known results was used to test the correct performance of the classifier.
Last but not least, performance evaluation was studied. It consisted of reviewing the different measures such as balanced accuracy, confusion matrix, F-score and receiver operating curves.
Having collected a great deal of data, the project reached experimental conclusion; it was time to undertake the writing of the present thesis.
All of these stages were closely monitored by the tutor, who was always available to solve doubts and propose new lines of progress. Regular meetings were scheduled according to what was required.

	Topic	Start date	End date	Duration (days)
1	Introduction	10/06/2016	10/06/2016	1
2	Documentation	11/06/2016	15/06/2016	5
3	Scope meeting	16/06/2016	16/06/2016	1
4	Database import	17/06/2016	23/06/2016	7
5	Basic K-means segmentation	17/06/2016	20/06/2016	4
6	Improved K-means segmentation	21/06/2016	14/07/2016	24
7	Source normalization	15/07/2016	21/07/2016	7
8	N-cuts segmentation	22/07/2016	04/08/2016	14
9	Feature extraction	05/08/2016	25/08/2016	11
10	Multivariate Gaussian model training	26/08/2016	29/08/2016	4
11	Multivariate Gaussian model testing	30/08/2016	31/08/2016	2
12	Debugging	01/09/2016	10/09/2016	10
13	Gaussian Mixture model training	11/09/2016	27/09/2016	17
14	Gaussian Mixture model testing	28/09/2016	2/10/2016	5
15	Feature extraction corrections	3/10/2016	11/10/2016	9
16	Performance assessment I	12/10/2016	17/10/2016	6
17	Sample dataset tests	18/10/2016	1/11/2016	15
18	Performance assessment II	2/11/2016	20/11/2016	19
19	Mean-shift segmentation	21/11/2016	10/12/2016	20
20	Performance assessment III	11/12/2016	18/12/2016	8
21	Final debugging, plots and tests	18/12/2016	9/01/2017	23
22	Bachelor's thesis	19/01/2017	22/02/2017	37

Table 1. Project task breakdown.

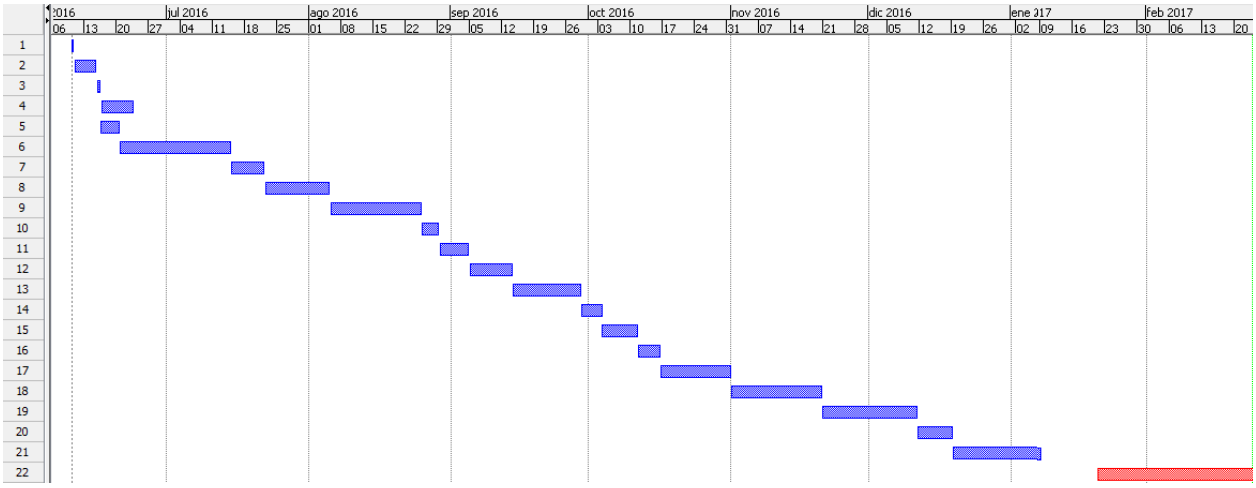


Table 2. Gantt chart.

4. ECONOMIC ENVIRONMENT

The current economic background tends to favor the growth of research and practice in Computer-Aided Systems for medical applications.

As presented in 5.3 COMMERCIAL CAD SYSTEMS, there is a direct translation of CAD developments to clinical use. Many commercial CAD systems, such as Molemax, Melafind and DB-Mips are commonly employed as aides in medical examinations.

Nowadays, there is a growing tendency towards user oriented appliances that bring CAD systems closer to patients, rather than being specialist exclusive. These conditions have led to the development of different tools that aim to offer a tentative lesion evaluation to refer individuals to experts, such as smartphone apps and attachments. MoleScope [8] falls within the second category: it is a magnifying lens attachment that scans lesions to supply the user with high-quality pictures in order to keep track of moles. Another product that is illustrative of the current environment is SkinVision. This is a smartphone app that processes lesion images taken on the phone itself to estimate risk of cancerous growth. It is the only instance of software so far to be able to detect different kinds of skin cancer besides melanoma: its algorithm was recently upgraded and now identifies Basal Cell Carcinoma (BCC) and Squamous Cell Carcinoma (SCC) [9].

4.1. FINANCES

This section details the total budget of 33.500€, derived from the cost of resources, both human and material.
Two people have been involved in the making of this project: a junior engineer and a senior engineer. Based on a set of phases devised from the Project Management Institute's PMBOK, an approximate estimation of the hours spent working on the different stages is shown below:

Phase	**Time**
Initiation	70 hours
Planning	80 hours
Execution	300 hours
Control	300 hours
Close	150 hours
Total	900 hours

Table 3. Project time breakdown.

Name	Position	Working time	Fee	Fee
Javier López Labraca	Senior engineer	100 hours	60 €/hour	6.000 €
Irene Hernández Serrano	Junior engineer	900 hours	20 €/hour	18.000 €
Total				24.000 €

Table 4. Human resources costs.

Resource	Price
Individual Matlab license	2.000 €
Personal computer	500 €
University computer	7.000 €
Total	9.500 €

Table 5. Material resources costs.

5. SOME NOTIONS OF MELANOMA DIAGNOSIS AND COMPUTER AIDED DIAGNOSIS

Melanoma is a type of skin cancer caused by cancerous growths due to damaged skin cells whose multiplication creates malignant tumors.
The aforementioned cells come from melanocytes. Melanocytes are found in the epidermis, which is the top layer of the skin, and are responsible for the production of melanin, a pigment meant to protect the skin from the negative impact of sun radiation.

5.1 DERMOSCOPY AND METHODS OF MELANOMA DIAGNOSIS

The most common procedure for melanoma diagnosis begins with a physical examination by a doctor, followed by a non-invasive diagnostic technique known as dermoscopy.
Dermoscopy entails the magnified visualization of the pigmented lesion with a device that allows digital imaging, such as a camera or a dermatoscope. Prior to inspection, the lesion is coated with fluid to make it translucent and avoid reflections. This enables the visualization of reflections from deeper layers of the epidermis and top dermis layers. The depth of melanin in the lesion can be determined through color: different hues are related to skin layers at different levels. Black is associated with the topmost layer of the epidermis. Brown shades correspond to layers below the external epidermis. Grey is linked to the dermoepidermal junction and blue-gray to the dermis.
Therefore, the higher the number of colors found in a dermoscopic lesion, the more chaotic the growth is, suggesting a higher risk of melanoma.

The diagnosis commonly follows, among others, one of these algorithms:

- The ABCD rule [10] streamlines the pattern criterion, thus simplifying the classification phase for the physicians. It takes into consideration more parameters present in the lesion: asymmetrical shape, irregular border, diameter and the amount and unevenness of its color, which is the central scope of this project. This rule has been further completed with the inclusion of the evolution factor (ABCDE rule).
- The 7-point checklist [11] assigns quantitative scores to a small set of identifiable features.
- The Menzies scale [12] offers a binary classification based on negative (axial asymmetry, color symmetry, presence of one color) and positive

features (discoloration, many colors, blue-white veil, certain patterns). The absence of negative features paired with any positive feature implies the presence of melanoma.

5.2 COMPUTER AIDED DIAGNOSIS (CAD) SYSTEMS FOR MELANOMA DETECTION

Computer-aided diagnosis systems can be used as support tools in the medical field. CAD systems are founded on data sets, usually derived from real cases. They help build statistical models through image processing, feature extraction and classification.

Provided the image data set, the first step is normalizing the source, to remove or at least reduce artifacts, which might include hairs, dermoscopic gel, air bubbles, irregular illumination, skin imperfections, etc. These flaws make the forthcoming courses of action more complex, hence the need for pre-processing.

A conventional approach to pre-processing is to apply smoothing filters; for instance, mean, median or Gaussian filters.

In order to diminish the appearance of lines and hairs, mathematical morphology may be used.

Other strategies are color quantification (fewer colors allow for easier management), contrast enhancement (borders become more noticeable) and conversion from RGB to gray (less noise and fewer artifacts), which is useful when the target feature is texture and color is then irrelevant and can be disregarded.

Source lightning normalization will be explored in greater detail in future sections.

The next logical step is to locate the lesion and single it out. This part of the process can be done manually, tracing the borders, or implementing segmentation algorithms, based on color, discontinuity, regions, soft computing or thresholding.

Next, feature extraction is performed. The concept of feature extraction refers to the characterization of the lesion by a series of measures on the pixels that belong to the segmented object.

Said features can be decided upon according to the previously explained diagnosis methodologies, or according to the attributes of the chosen feature extraction system (filtering or statistical).

Feature selection is a way to optimize the amount of parameters that have been extracted. The number of selected features, known as k, is essential to the fruition of the classifier and could lead to either overfitting or lack of discrimination between classes.

Subsequently, the classification of stage delivers a diagnostic which can be binary, melanoma or benign lesion, or, alternatively, the probability of belonging to each class.
In this document, the output of the classifier is a series of probabilities, indicating the most likely color for each region in the lesion.
Finally, it is crucial for a CAD system to assess its efficiency. There are many possible mechanisms to do so, some of which will be developed later on. For this thesis, test to train ratio was applied, adopting several different tools for performance evaluation: balanced accuracy, F-measure, confusion matrix and receiver operating curve.

5.3 COMMERCIAL CAD SYSTEMS

The only commercially available clinical decision support systems for skin cancer, also known as CCDSs, pertain exclusively to melanoma and no other type of skin cancer [13]. There are several programs that examine the lesion's statististics to provide a comparison with a database made up of catalogued cancerous and benign lesions.

- SIAscope [14]: This device carries out a spectrophotometric analysis, making use of infrared and visible light emissions, which prompt different responses from the skin's chromophobes. The reflected light is interpreted by the software in terms of melanin, collagen and blood distribution, according to the quantity of light absorbed at each wavelength.
- MoleTrac [15]: Also known as SolarScan, it produces a decision tree based on a comparison against lesion features of a given database. It employs image analysis software to pre-process the image prior to comparison.
- MelaFind: This non-invasive tool assists dermatologists using spectral imaging technology. The outcome is a recommendation for or against a biopsy (i.e., a medical test involving a sample of tissue to examine the cells more precisely). Just like the SIAscope, it is multiespectral, although it works with ten different wavelengths, as opposed to the twelve operated by the SIAscope. After determining the edge of the lesion, it generates a sequence of digital images and then evaluates several parameters, such as wavelet maxima, asymmetry, color variation, perimeter alteration and texture changes [16].
- MoleMax [17]: This appliance gives a score conforming to the obtained dermoscopy. MoleMax creates full-body risk maps to track changes, making it ideal for follow-up examinations. It is a polarized-light dermoscope that utilizes epiluminescence microscopy (ELM) without the need for any oil or fluid between skin and apparatus, which provides the means for standardized comparison with previous data.
- DB-Mips [18]: This diagnostic aid is a module that analyzes the patient's lesions in real-time, as it does not need to store the images prior to analysis.

6. STATE OF THE ART OF COMPUTER-AIDED MELANOMA DIAGNOSIS BASED ON COLOR

This section will dive into the current advancements in the field of color analysis for melanoma detection. Within this framework, several authors discuss the pre-processing stage and many works about CAD systems touch upon color in terms of feature extraction.

The approach to color detection in CAD for melanoma detection is diverse. Usually, statistical parameters, like mean color and color variance, are used to describe color. E.g., the feature descriptor for color variation can be obtained via four different channels: the red, the green, the blue and the intensity channel, which is obtained as a fusion of the RGB channels [19]. Then, the color variation for each channel is computed as the logarithmic ratio of the mean to the standard deviation.

A different solution for the characterization of the lesion is employing color histograms as descriptors, where each bin represents the amount of pixels in that color channel. In [20], for example, the lesion is split into cells, associating three descriptors with each cell of the grid: mean color vector, uni-dimensional color histogram and generalized color moments.

As for pre-processing, source standardization deals with artifacts present in the lesion, such as body hair. These algorithms perform image segmentation to characterize the image as an assortment of a fixed number of regions. Consequently, mostly isolated, small details become less relevant. The techniques that cause this are studied in ensuing sections.
A prevalent design for hair detection and eradication applies directional filters, exploiting the highly directional nature of the shape of hairs [21]. Commonly, a bank of Gabor filters is used. These filters' impulse response is a linear combination of Gaussian filters. They have a linear shape, like the artifacts of interest. They are combined, forming a bank, to evaluate every possible direction, since the orientation of the hair is not known.
Pre-processing researchers also engage in other aspects of source normalization. There are proposed alternatives that seek to improve the conventional color constancy algorithm. The image is initially transformed to achieve chromatic adaptation. This means that colors are modified to mimic the human visual system, so that a variation in the illumination does not change the perception of a color, which is independent of light conditions. Traditionally, this involves a single transformation matrix, like the von Kris diagonal model, for every color in the image. A new method [22] proposes dividing the source color gamut into regions and applying a unique

transformation matrix to each of those regions. It utilizes Delaunay triangulation (no point in the plane is inside the circumcircle of any of the triangles) and Macbeth color checker (calibrating color palette whose patches are designed to resemble the spectral reflectances of natural objects and be very stable over time and under varying lightning situations).

Studies focused on the role of color are notably scarce.

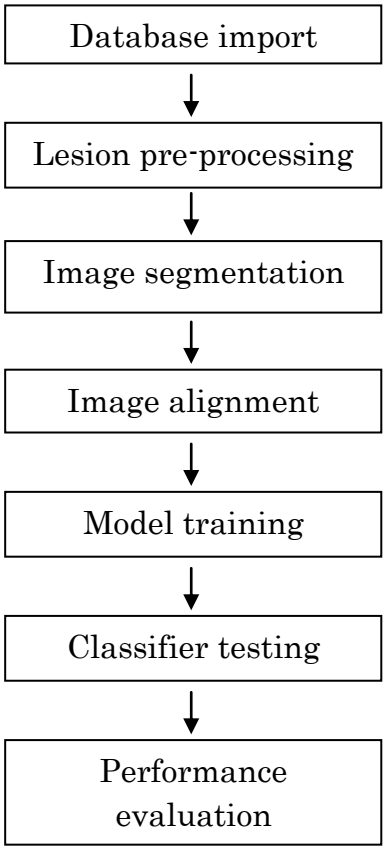
A research supported by NIH-SBIR (Small Business Innovation Research) [23] concluded that the most useful color discrimination information for lesion screening is found in the region closest to the skin lesion boundary. The color histogram is applied to evaluate two color features: percent melanoma color and color clustering ratio. Percent melanoma color refers to the proportion of pixels that are melanoma colors, whereas the color clustering ratio is the fraction of melanoma-colored eight-connected neighbors of melanoma-colored pixels to the number of eight-connected neighbors for all melanoma-colored pixels.

The same authors [24] pointed out the relevance of structured color information, given the superior results of color clustering ration versus percent melanoma color. Another conclusion was reached: color feature information is located at the center of the lesion, for the most part.

In [25], color detection in melanoma dermoscopy images was studied focusing exclusively on three shades of blue that were analyzed using fuzzy set techniques. Stoecker et al. [26] researched areas of granularity in skin pigmented lesions extracting texture and color features.

7. SYSTEM DESCRIPTION

For more clarity, an schematic figure with blocks for each essential phase can be found below. It shows an overview of the chain of tasks that define the project's scope. The blocks in the graph, which are the components that illustrate the extent of this Bachelor's Thesis, are going to be studied further along.



7.1 SOURCE NORMALIZATION

In preceding sections, source normalization was deemed an essential step in CAD systems. Given the diverse circumstances under which each sample is taken, variations in the system's behavior are to be expected. In other words, fluctuating conditions when collecting data can alter the degree of efficiency: different cameras, hospitals, lightning situations and the like, may result in unpredictable feedback when subject to a system adjusted for other settings. There are other factors to be accounted for, such as hairs or oil bubbles. For the specific database used in this project, this kind of artifacts were not a problem, due to the effect of the algorithms applied later on.

The focus of the pre-processing phase, however, is on the normalization of colors. This task will be achieved applying the concept of color constancy or chromatic adaptation. The fundamental idea of color constancy is that every picture is perceived as though all of them were captured under the same light. The human visual system possesses this feature: the eye can generally adapt to changes in illumination, maintaining the appearance of the visualized colors. The human eye perceives color through a type of photoreceptor cells found in the retina, called cones. There are three types of cone cells (S,M,L) depending on their sensibility at short, medium and long wavelengths. The spectral response of the cones models the LMS color space, as seen in Figure 1.

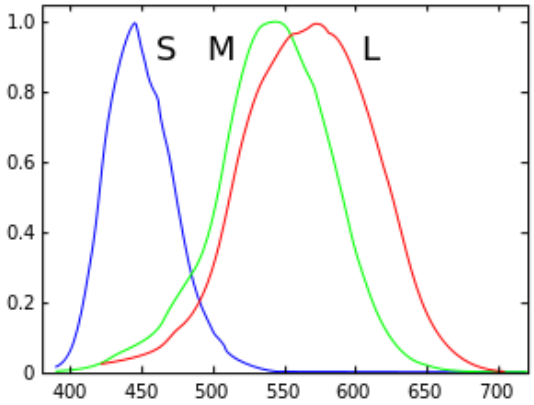


Figure 1. Normalized frequency response of S, M and L cones [35].

In order to mimic this adaptability, the initial illumination given by an unknown source illumination of the image is altered. Thus, the apparent illumination is modified to fit the LMS cone space. This way, the LMS response in the eye is equal to that of the desired illuminant (a perfect white light).

Chromatic adaptation starts with the estimation of the illuminant, which is then used to transform the image. Given the complexity of color perception in human vision, there is not just one transformation matrix that objectively describes the LMS response.

The approach assumed in this study follows techniques proposed in [27], which at the time of publication had not yet been applied to dermoscopic images.

First, a generalization of the Shades of Grey method, introduced by Finlayson and Trezzi [28], is utilized. It argues that better color constancy is achieved when assuming that the average of any scene is a shade of grey. The Shades of Grey algorithm merges the Max-RGB and Grey World color constancy methods.

Mathematically, the color is estimated as a vector or RGB coordinates (1), computing the normalized Minkowski *p*-norm (2).

$$v = \begin{pmatrix} v_R \\ v_G \\ v_B \end{pmatrix} \tag{1}$$

$$\left(\frac{\iint (I_c(x,y))^p \, dxdy}{\iint dxdy} \right)^{1/p} = k v_c \tag{2}$$

where c is the color component ($c \in \{R,G,B\}$), k is a constant such that v has unit length, and I is the value of the pixel in (x,y) for component c.

Optimum performance happens when p=6, according to Finlayson and Trezzi's experiments [28]. The Minkowski *p*-norm performs a weighted average over the intensity of the pixels, where the weights are directly proportional to the pixel value.

The second step involves a chromatic adaptation transform (CAT) of the colors. The chosen method is the von Kris diagonal model. The von Kris transformation matrix (3) applies a gain to each of the cone spectral responses in order to have a constant reference.

$$\begin{pmatrix} I_R^n \\ I_G^n \\ I_B^n \end{pmatrix} = \begin{pmatrix} \frac{1}{v_R} & 0 & 0 \\ 0 & \frac{1}{v_G} & 0 \\ 0 & 0 & \frac{1}{v_B} \end{pmatrix} \begin{pmatrix} I_R^u \\ I_G^u \\ I_B^u \end{pmatrix} \tag{3}$$

where I_c^n is the normalized pixel value for a perfect white light and I_c^u is the pixel value under unknown lightning conditions.

Some examples of the color constancy implementation on a few dermoscopic samples can be found below.

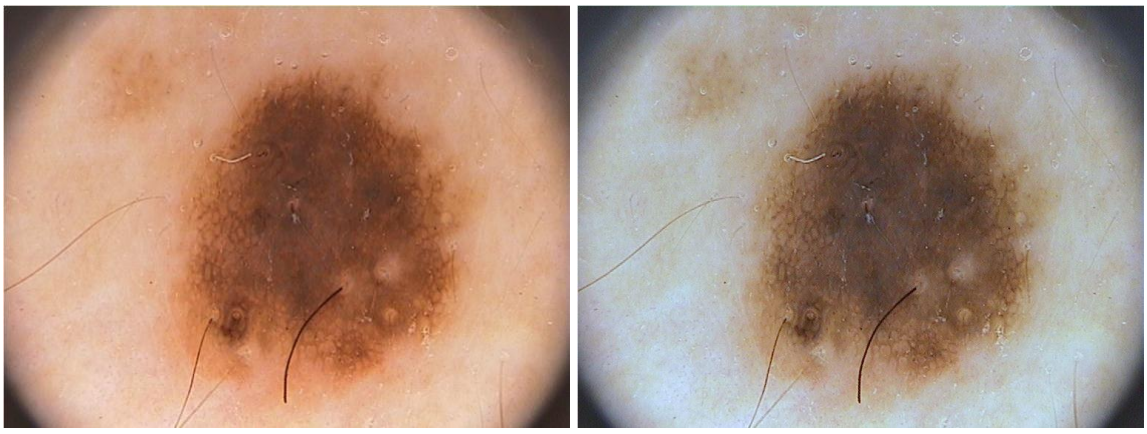


Figure 2. *IMD002* before (L) and after (R) color constancy.

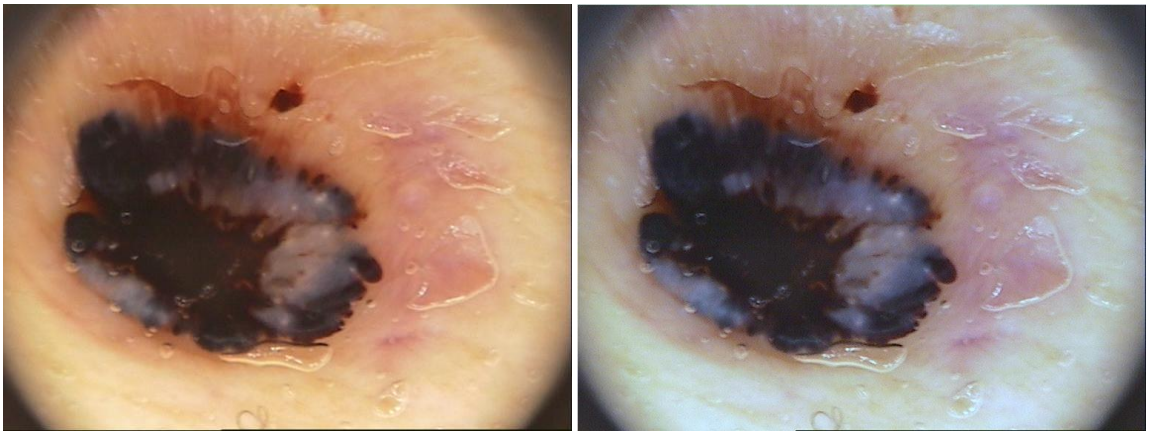


Figure 3. *IMD168* before (L) and after (R) color constancy.

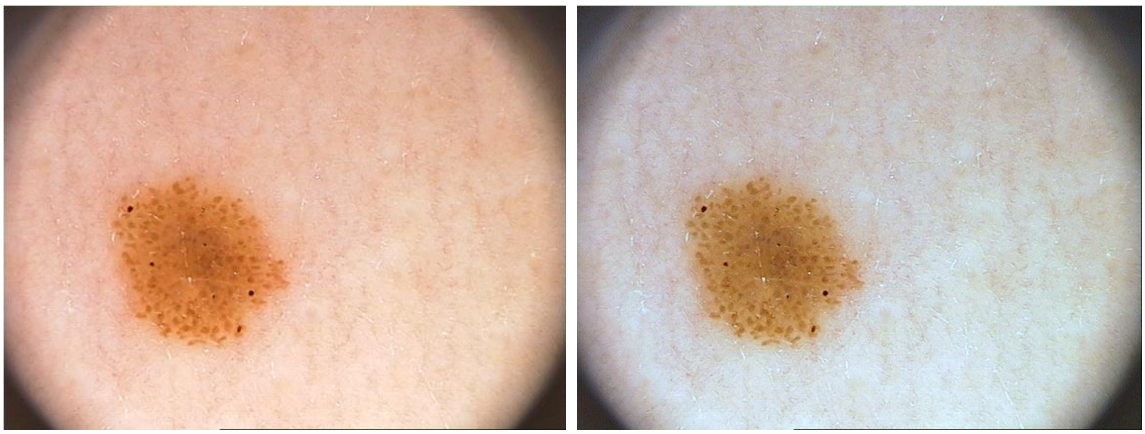


Figure 4. *IMD103* before (L) and after (R) color constancy.

7.2 SEGMENTATION

Once the lesion has been normalized with respect to the light source, it must be partitioned in order to handle its meaningful information efficiently. Every lesion will be segmented into a number of regions. The concept of region can be summed up as follows: a region is an area homogeneous in color, such that one label applies to every pixel of it.

The segmentation is performed with the available data: an image is made up of pixels, each of which is a point in a 3-dimensional space, comprising the intensities of the red, blue, and green channels for the RGB color space. The segmentation algorithms will treat each pixel in the image as a separate data point.

Given the crucial nature of segmentation, three algorithms were put into action: K-Means clustering, N-Cuts and Mean Shift.

K-Means

K-means is an unsupervised method that clusters data iteratively. Clustering data means creating groups of data objects that are similar in some respect. K-means color segmentation will assign a color membership to each color value. Hence, for any lesion, each pixel belongs to one of its k colors or clusters.

The principle of the K-means algorithm is that each observation is assigned to the cluster with the nearest mean. Therefore, the aim of each assignment is to minimize the expression (4).

$$argmin_j \, D(x_i, c_j) \tag{4}$$

where D is the Euclidean distance (straight-line metric) between the point x_i and the cluster center c_j, as per (5).

$$D(x_i, c_j) = \sqrt{\sum_{k=1}^{m} (x_{ik} - c_{jk})^2} \tag{5}$$

The means (centers) of the clusters are known as centroids. Centroids are randomly placed at the start, and then re-computed on every repetition according to the new mean of the cluster, until convergence. Convergence happens when cluster assignments stay constant.

The simplicity and low computational cost are its main advantages. Its drawbacks include the dependency on the random initial placement of centroids and on the aptness of the number of clusters.

The first analysis carried out involves is the data chosen as input to the K-means function. Initially, only color information is used, meaning that the input is a set of 3-dimensional points: the clustering is based on the color values for each pixel. With the intention of improving the algorithm results, more information was added to each sample. Specifically, given that usually colors appear in areas, not isolated, taking into account the location in the image of each pixel seems logical. Therefore, the second implementation describes each point using the color information as well as its x and y coordinates.

Basically, the clusters will be conditioned by an extra criterion: being assigned to a certain cluster implies similarity in color and proximity, since membership now depends on both features.

To better comprehend this matter, the dermoscopic image named *IMD058* (Figure 5) will serve as an example, by being clustered using K-means with different settings. For reference, according to the true values provided in the database information, the chosen lesion displays four colors: white, red, blue-grey and dark brown. These four labels are seen as RGB values in the known manual segmentation (Figure 6).

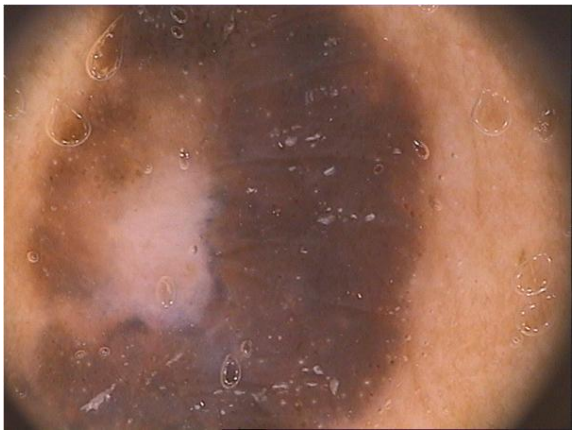


Figure 5. Original lesion (*IMD058*).

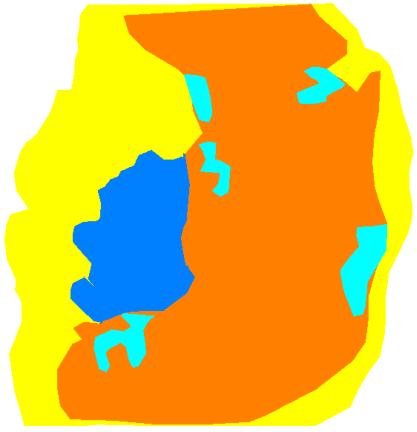


Figure 6. Ground truth image for *IMD058*.

Additional information about the true colors found in lesion is given in the form of color masks obtained by a dermatologist. The segmented color regions are shown below (Figure 7). The color red was not identified with a color mask. This is a common obstacle within this particular dataset, which will be explained when discussing database issues.

Figure 7. Segmentations for white, dark brown and blue-grey labels, respectively.

Now that the exemplifying image has been explained, K-means is performed, getting the results depicted next.

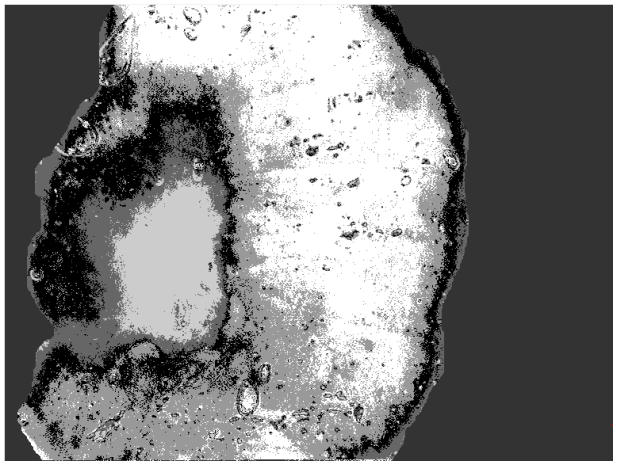


Figure 8. K-means (K=6) segmented lesion (*IMD058*) using color information.

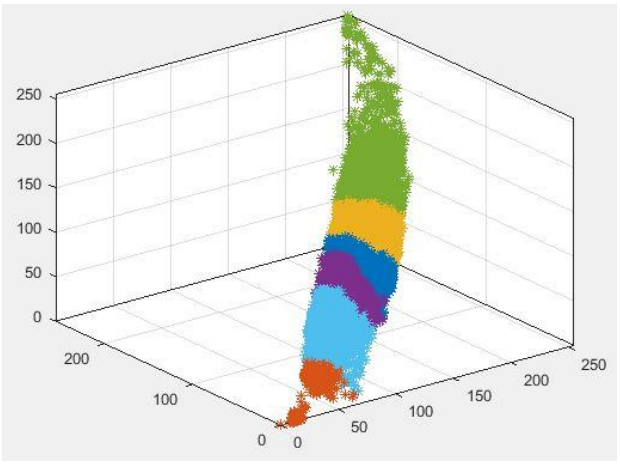


Figure 9. K-means (K=6) clusters in RGB space using color information from lesion *IMD058.*

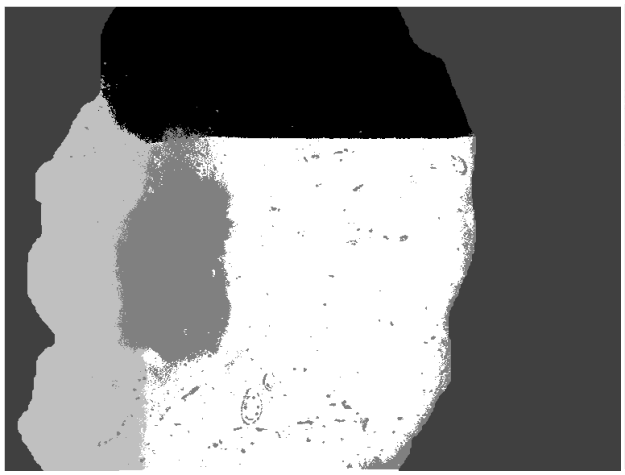


Figure 10. K-means (K=5) segmented lesion (*IMD058*) using color and position information.

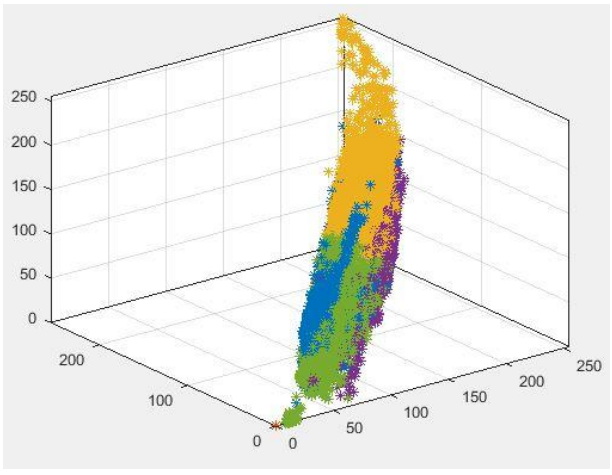


Figure 11. K-means (K=5) clusters in RGB space using color and position information from lesion *IMD058.*

The graphs shown in Figure 9 and (11Figure 11 display the grouping of pixels distributed along the axis that represent the color components: red, blue and green, given that the current color space is RGB.

There is a clear improvement on the method when including the position of the pixels. The visualization of the clusters in three dimensions is more uniform in Figure 9, which makes sense, because color resemblance is all that matters in the first K-means implementation. Connected regions are produced in the second case (Figure 10) when spatial information is introduced, which is a desirable, given the nature of color distribution in dermoscopic lesions. What's more, parameter *k* (number of clusters; i.e., number of colors) for image *IMD058* is correctly estimated as being equal to five (including background color as another cluster).

From now on, position information, along with color data, shall be used for every experimental output.

Another key aspect that is part of this evaluation is the spatial organization of color: the color space. Two color spaces are considered: RGB and CIE 1976 L*a*b*. The RGB (red-green-blue) color space is based on the RGB color model. A color model is a mathematical structure that represents colors as a small set of numbers. The RGB color produces an array of hues from three additive primaries: red, green and blue. The CIE 1976 L*a*b*, or simply Lab, color space describes each color in three dimensions: lightness (*L*) and two color channels (*a* and *b*), were *a* characterizes red/green opponent colors and *b* describes the yellow/blue opponent colors. Compared to RGB, Lab comprises a larger subset of colors: its gamut or range is bigger because it contains all perceivable colors (Figure 13).

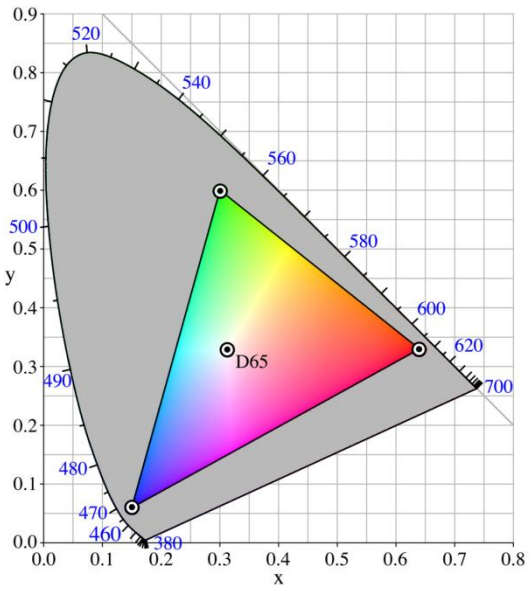


Figure 12. Chromacity diagram for RGB color space [36].

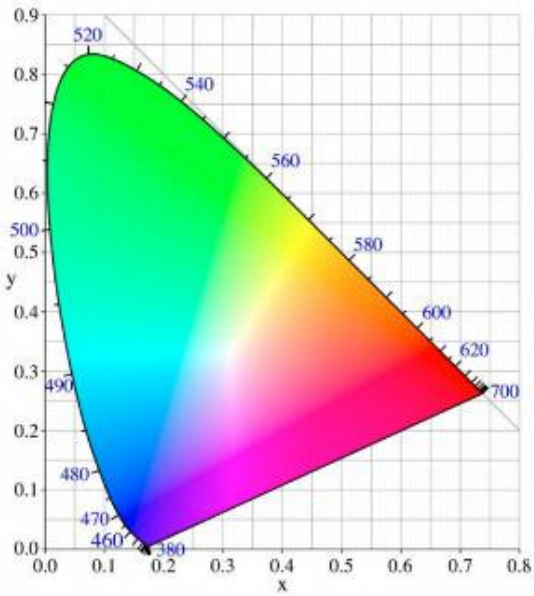


Figure 13. Chromacity diagram for Lab color space [37].

For comparison purposes, the lesion picture was converted to Lab color space. The outcome for the K-means segmentation is shown in Figure 14 and Figure 15.

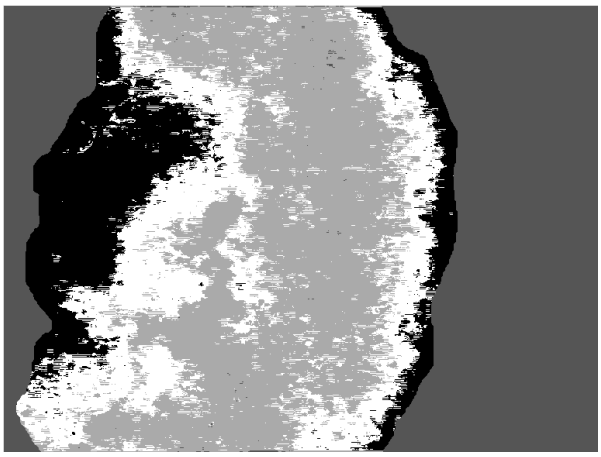


Figure 14. K-means (K=4) segmented lesion (*IMD058*) in Lab color space.

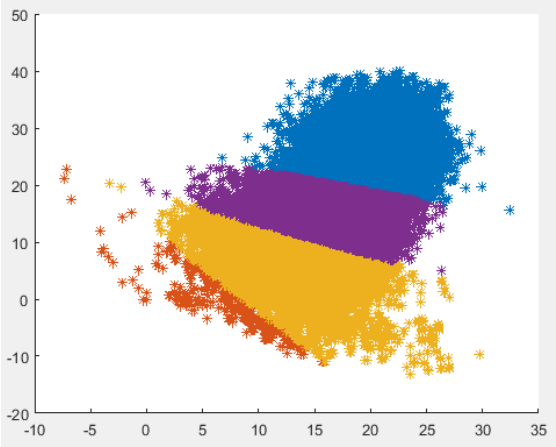


Figure 15. K-means (K=4) clusters in Lab color space from lesion (*IMD058*).

The clusters for the Lab color space are displayed in two dimensions because only the *a* and *b* components are used in the clustering process. One of the advantages to working in Lab space is that it decorrelates luminance (*L*) and chrominance (*a*, *b*), so any color is completely defined in two dimensions. This implies less computational cost. While this could be a great asset, the clustering of the color brown (light and dark) models would be hindered, given their dependence on luminance to be differentiated. All three components are relevant under these circumstances. The main benefit of this color space is its precision: the distance between colors correspond to the perceived color differences; the Euclidean distance allows accurate distinction of hues. Consequently, Lab is the most recommended color space in literature for colorimetric tasks.

Finally, the *K* parameter becomes the research matter. As previously stated, *K* stands for the number of clusters or colors into which pixels are categorized. So far, *K* has been assumed to be automatically computed. This is tackled next.

K could be an input parameter, but considering the purpose of this project, a certain number of colors cannot be assumed beforehand; the system should provide the user with the amount of colors found, which is an important trait in melanoma diagnosis.

To achieve the optimum number of clusters for each lesion, the Calinski-Harabasz index [29] was found to be the most effective option. Said

clustering validation method decides on the best number of clusters depending on the average between-cluster sum of squares (6) and within-cluster sum of squares (7).

$$SS(B) = \sum_{i} (x_i - c_j)^2 \tag{6}$$

$$SS(W) = \sum_{j} \sum_{i} (x_{ij} - c_j)^2 \tag{7}$$

The sum of squares is a metric that describes the dispersion of data points. From the perspective of "between" clusters, it measures the variation among group means (centroids). "Within" clusters refers to variation of individual points with respect to each centroid. It is an ANOVA-based criterion: it performs an analysis of the variance, understood as the difference between groups.

The graph for the Calinski-Harabasz index (Figure 16) offers a visual insight into the result: it has a maximum at the optimum value for *K*. For the one shown in Figure 16, there would be 3 colors.

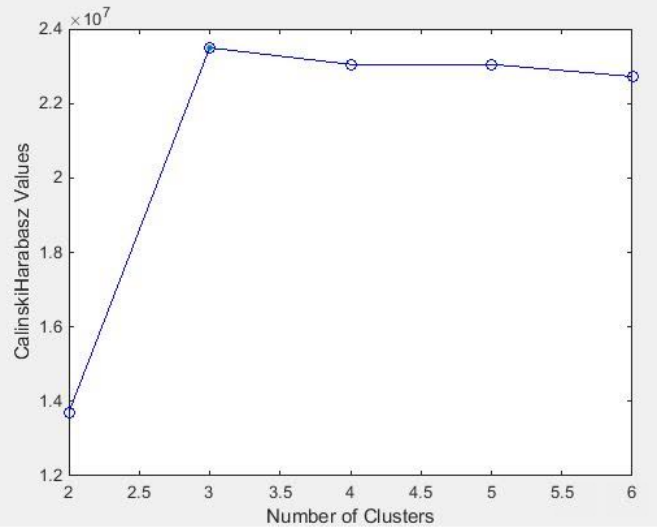


Figure 16. Calinski-Harabasz graph example.

N-Cuts

The algorithm for Normalized Cuts, as presented in [30], offers a clustering method based on global perception, rather than on local properties.

The theoretical ground of it starts with graph theory: the method is developed on the basis of structures that model object to object relationships; i.e., graphs. To be more specific, weighted graphs. A weighted graph is one whose edges have weights associated to them.

Given the purpose of this section, it is relevant to study the division of such a graph. A preliminary study proposed the minimization of the degree of dissimilarity when cutting a graph into two disjoint parts. This measurement, known as *cut* (8), is computed as the sum of the weights (costs) of the edges that were removed to split the graph.

$$cut(A,B) = \sum_{u \epsilon A, v \epsilon B} w(u,v) \tag{8}$$

However, the computation of a minimum cut tends to create small regions of isolated nodes. This happens because outlying nodes have fewer edges and, as a consequence, fewer weights to sum over, giving a smaller result for *cut*. Therefore, this minimization mechanism is not always correct.

N-cuts improves the partitions by normalizing the cut criteria:

$$Ncut(A,B) = \frac{cut(A,B)}{\sum_{u \epsilon A, t \epsilon V} w(u,t)} + \frac{cut(A,B)}{\sum_{u \epsilon B, t \epsilon V} w(u,t)} \tag{9}$$

The denominators of expression (9) stand for the sum of the costs from each point (A or B) to all the nodes in the graph. This is known as association (10).

$$Assoc(A,V) = \sum_{u \epsilon A, t \epsilon V} w(u,t) \tag{10}$$

The concept of association refers to how well connected that point is to the rest of the nodes.

This way, *Ncut* is not small for sets of isolated nodes anymore, because their association measure will be smaller, so the previous issue is solved. The objective of the algorithm is to maximize association within groups, which is the same as minimizing the disassociation between groups.

The minimization of *Ncut* is computationally prohibitive, which is why the solution takes an approximate form through the generalized eigenvalue

system. This allows the N-Cuts algorithm to be summed up in a few steps. From the image, a weighted graph is created, using the similarity between two nodes as the weight of the edge connecting them. Then, the generalized eigenvalue system $((D - W)x = \lambda Dx)$ is solved for eigenvectors with the smallest eigenvalues. The eigenvector with the second smallest eigenvalue is used to divide the graph in two. These actions are repeated recursively if needed.

Although this algorithm is well extended in low level segmentation tasks, the available implementation only considers grey scale. This limits its ability to outline proper regions, and the other two methods give better results.

Figure 17. Lesion *IMD016* with binary mask.

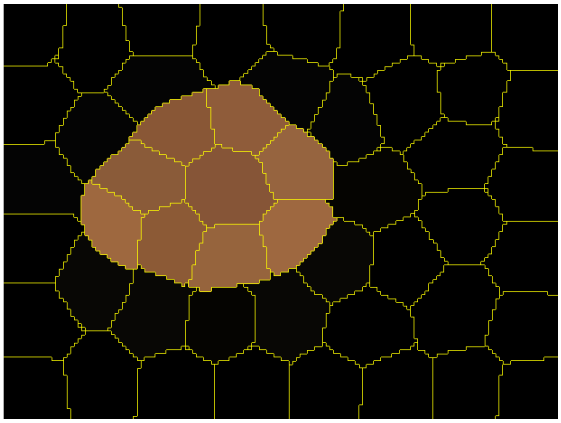


Figure 18. N-Cuts segmentation (visualized with borders) for masked lesion *IMD016*.

Mean shift

Mean shift is a data mining algorithm with many applications, from finding local maxima to data clustering in computer vision.

This analysis method does not assume the probability distribution of the variables of interest. The number of parameters is not constant, but proportional to the amount of training data. Hence, mean shift is a non-parametric algorithm.

Besides, this technique is feature-space based. This means that it works with a vector space composed of feature vectors, which are representations of objects in numerical features.

When mean shift is used, the feature space is seen as an empirical probability density function (PDF). The input (set of data points) are samples from that PDF. The mode or local maxima of the probability density function constitute clusters (dense areas).

The mean shift algorithm associates a point to the closest maximum in the feature space; a window is placed around it and the mean shift vector is computed (12). This vector is the difference between the weighted mean and the center of the window (kernel).

The window is then shifted according to $\boldsymbol{m}_{h,G}$. Local maxima are found through gradient ascent; regions of high-density value are therefore of interest. The mean shift vector always points towards the maximum increase of density, because at point x, it is proportional to the density gradient estimate $\hat{\nabla} f_{h,K}$, obtained for kernel K (multivariate normal), which is normalized by the density estimator $\hat{f}_{h,G}$, computed for kernel G ($G = c_{g,d}\, g(||x||^2)$) (11).

$$\hat{f}_{h,G}(x) = \frac{c_{g,d}}{nh^d} \sum_{i=1}^{n} g(|| \frac{x - x_i}{h} ||^2) \tag{11}$$

$$\boldsymbol{m}_{h,G}(x) = \frac{1}{2} h^2 c \frac{\hat{\nabla} f_{h,K}(x)}{\hat{f}_{h,G}(x)} \tag{12}$$

where c is a normalization constant, h is the bandwidth and n is the number of data points.

The computation of the mean shift vector and the relocation of the kernel are iterative processes. Consequently, $\boldsymbol{m}_{h,G}$ creates a path that leads to a stationary point of the estimated density $\hat{f}_{h,G}$. These steps are performed

until convergence, which happens when the gradient estimate obtained for the multivariate normal kernel is zero.

A key aspect that makes this method preferable to N-Cuts is the fact that color is taken into account for the segmentation, whereas the normalized cuts implementation does not.

Some advantages of mean shift compared to K-Means are the lower sensitivity to initializations and outliers, plus the lack of assumptions regarding the number of clusters, which depends on the number of modes. Mean shift offers more flexibility because it does not take for granted the shape of the clusters (they are elliptical for K-Means clustering); they can be arbitrarily shaped.

Figure 19. Lesion *IMD009* with binary mask.

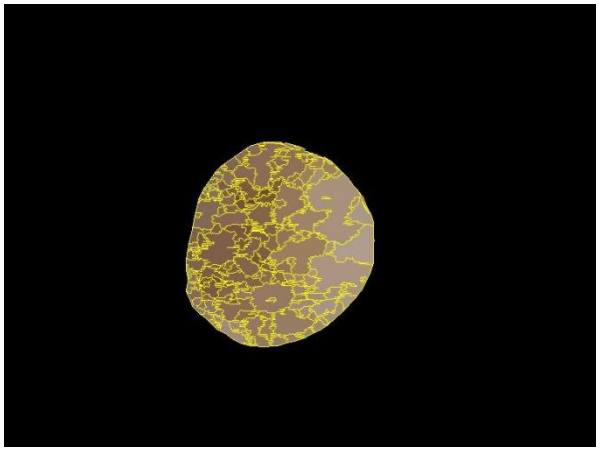


Figure 20. Mean shift segmentation (visualized with borders) for masked lesion *IMD009*.

7.3. FEATURE EXTRACTION

Through feature extraction, relevant characteristics are plucked from the original data.
The concept of ground truth should be made clear beforehand. In machine learning, ground truth stands for the true representation of the data. It is the standard the results will be compared to. As far as this project is concerned, the ground truth images are labeled lesions, divided into areas that correspond to a color class, as observed in the following visualizations of ground truth data.

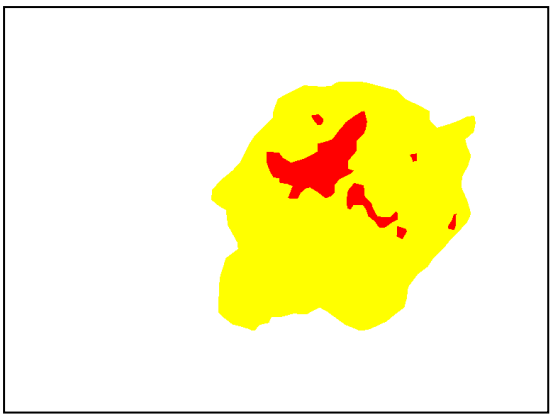


Figure 21. Ground truth (for *IMD208*) with 2 colors (dark brown and light brown).

Figure 22. Ground truth (for *IMD372*) with 1 color (light brown).

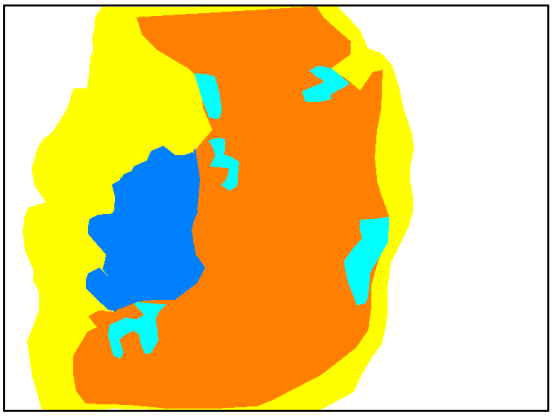


Figure 23. Ground truth (for *IMD435*) with 3 colors (dark brown, light brown, blue gray and black).

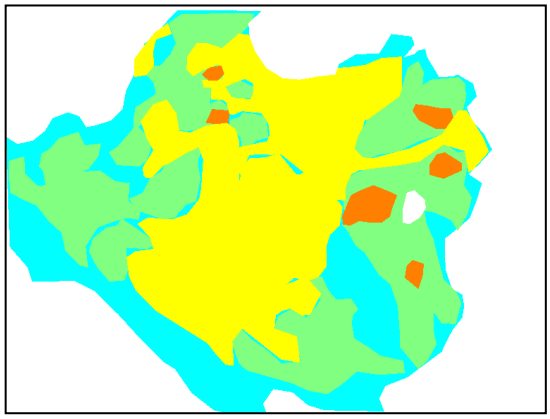


Figure 24. Ground truth (for *IMD435*) with 4 colors (white, red, dark brown and blue gray).

With the purpose of displaying the pictures shown above, labels were converted to RGB values.

Once the meaning of ground truth has been established, feature extraction can be explained as the need to summarize the information contained in a region into a single feature vector.

The information in this study, at this stage, is presented in the form of images segmented into regions. Every region has a value assigned to it.
However, this region label does not match the values found in the ground truth images, which range from 1 to 6. Those values depict each of the possible colors commonly present in a skin lesion:

- 1: White.
- 2: Red.
- 3: Light brown.
- 4: Dark brown.
- 5: Blue gray.
- 6: Black.

Image alignment enables the mapping of the reference data (segmented lesions) to the target data (manually segmented ground truth images). That is, the values of the regions in the lesions, which were segmented via N-Cuts, Mean Shift or K-Means algorithm, will be transformed to fit the range (1-6) of the ground truth segmentations.
To do so, the reference or source image is overlapped with the target or subject image. Each region of the segmented lesion acquires a single representative label: the most repeated value in the ground truth for that same region.

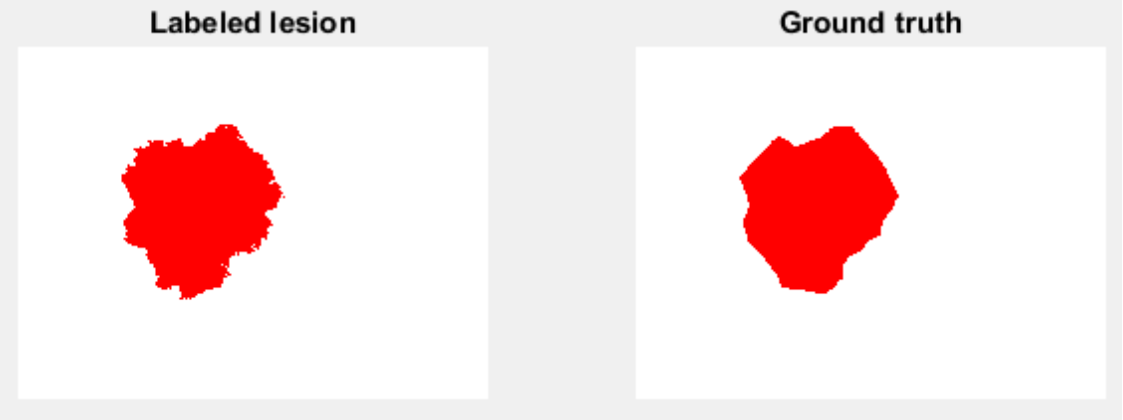


Figure 25. Labeled lesion *IMD003*: after segmentation and alignment (L) and ground truth (R).

Figure 26. Labeled lesion *IMD042*: after segmentation and alignment (L) and ground truth (R).

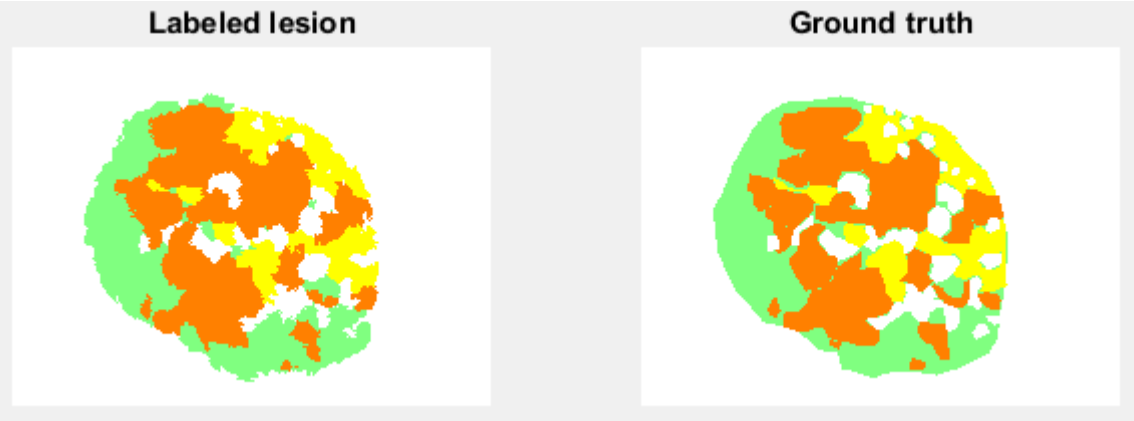


Figure 27. Labeled lesion *IMD150:* after segmentation and alignment (L) and ground truth (R).

At the end of the execution of this algorithm, several data structures are returned:

- A label vector containing the color label (as a number between 1 and 6) for each region. This is the outcome of the alignment procedure.
- A data matrix with the color information (i.e., one value per color component) for each region. A region is represented by a single color whose components are obtained from all the region pixels, taking the mean of each color channel.
- A vector that links the regions to its lesion. The point of this vector is to keep track of the lesions for each region descriptor, since they have been broken up into regions. This is useful for future tasks.

To sum up, the extracted features are characterized as follows:

	Label Vector	Image Vector	Data Matrix: R/L	Data Matrix: G/a	Data Matrix: B/b
Region 1 →	Color label ∈ (1,...,6)	Name of lesion it belongs to ('IMDxxx')	Value for red or lightness channel	Value for green or a^* channel	Value for blue or b^* channel
	. . .	. . .	. . .	. . .	. . .
Region N →	Color label ∈ (1,...,6)	Name of lesion it belongs to ('IMDxxx')	Value for red or lightness channel	Value for green or a^* channel	Value for blue or b^* channel

N regions

7.4. CLASSIFICATION

Overview

Classification in CAD systems refers to accomplishing a diagnosis by inference for the input lesion, based on information that was previously gathered.
In machine learning tasks, there are two general approaches: unsupervised and supervised learning. Unsupervised learning clusters or associates data based exclusively on a set of unlabeled inputs. Supervised learning trains a set of data whose correct labels are known and used to infer a mapping function, such that any new input can be predicted according to it. This work focuses on supervised learning, because the training process aims to develop a model from a feature matrix that is representative of labeled training data. Given that the targets are classes (colors), the problem becomes a classification task.
From the perspective of the result, there are mainly two means of tackling classification: hard and soft. Hard classification assigns to each region the most likely class. In melanoma diagnosis, this can be a binary decision: label 0 or 1 depending on whether it is melanoma or benign.
Instead, soft classification consists in recognizing that each region can belong to more than one class and giving membership grades (probabilities) to each. The chosen option is a soft classifier, meaning a region can be each of the six colors with a certain probability. Within the context of medical diagnosis, which deals with delicate matters, it makes sense to pick a soft or fuzzy classifier that measures the confidence of the prediction.

The classification stage was divided into two phases. First, the training data is selected, constituting 70% of the available database. The color features from the training data, collected in the structures explained before (7.3. FEATURE EXTRACTION), are used to estimate several statistical parameters. These will define the function that shapes the distribution of each color.
Then, using the remaining 30% (test data), the obtained models are used to predict which colors each region might belong to.
Hence, the approach taken is model-oriented. Specifically, training of the data was done for two different statistical models to base decisions on: multivariate Gaussian model and Gaussian mixture model.

Multivariate Gaussian Model

In the first case, each color was modeled as a Multivariate Gaussian: a generalization of the univariate Gaussian distribution to k dimensions.

Its probability density function is, by definition:

$$p(x;\mu,\sigma) = \frac{1}{2\pi^{k/2}|\sum|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T\sum^{-1}(x-\mu)} \tag{13}$$

where $|\cdot|$ denotes the determinant, $\sum$ is the covariance matrix (symmetric, positive and semi-definite) and μ is the mean.

Any Gaussian distribution is fully characterized by its mean and covariance matrix. This fact allows the depiction of each color as a three variable Gaussian model. Therefore, each model is defined by the mean value of each color channel in the color space, as well as the covariance matrix that contains the variance of each component and the variances between each pair of variables.

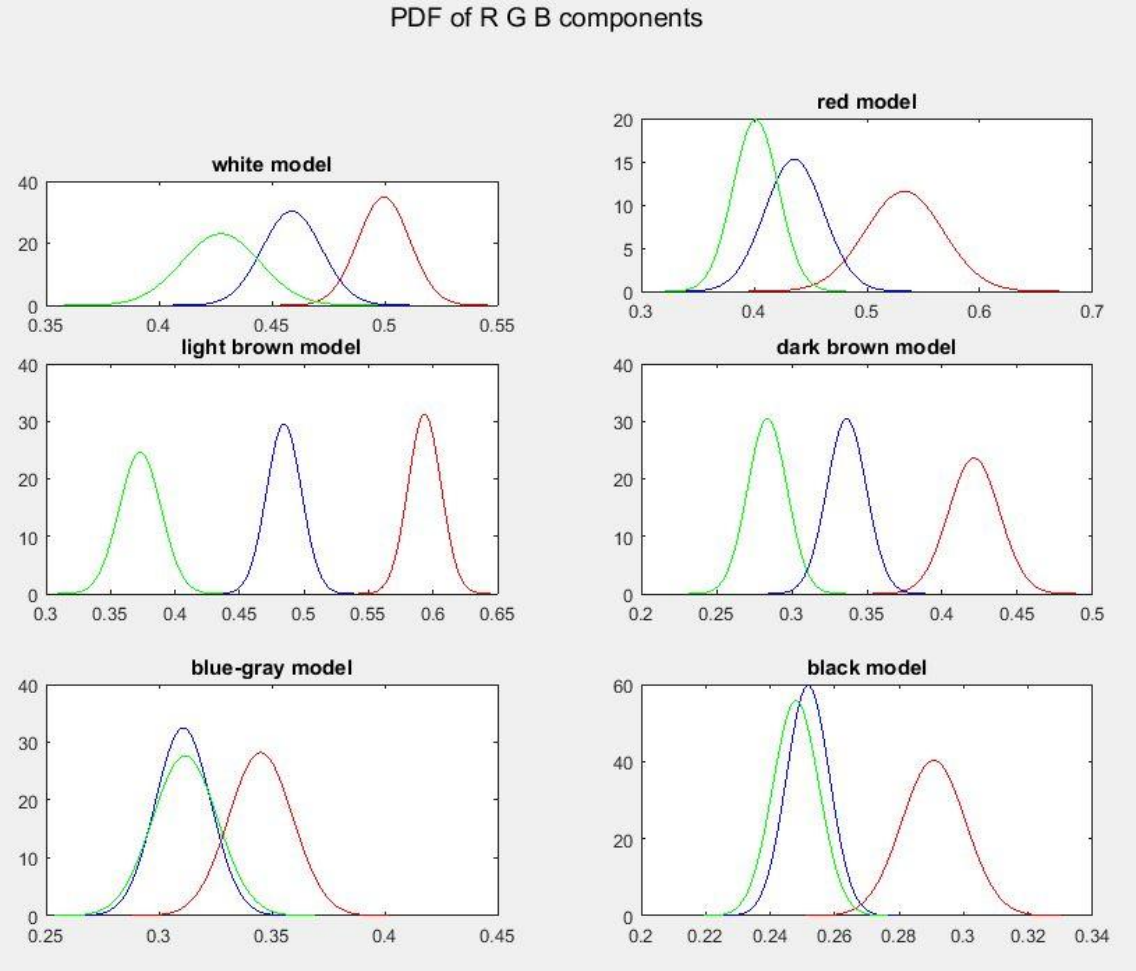


Figure 28. Probability density distributions of Multivariate Gaussian Models in RGB color space.

When it comes to classifying test samples, for each new sample or region, the decision criterion relies on the value for the probability density function (PDF). The PDF is obtained for each of the six color models. For completion, in order to provide a tangible decision, the highest probability is selected. The color associated to that PDF is assigned as the inferred value for the region. The classifier maximizes the posterior probability of the class; it is governed by the maximum a posteriori decision rule (14).

$$MAP = argmax\ P_i(c|x) \tag{14}$$

where c is the color class and x is the observation.

The decisions made for a randomly selected subset (30%) of the database, after training the remaining lesions to build multivariate Gaussian models, can be observed in Figure 29. Each point appears in the color of the color class it was assigned to.

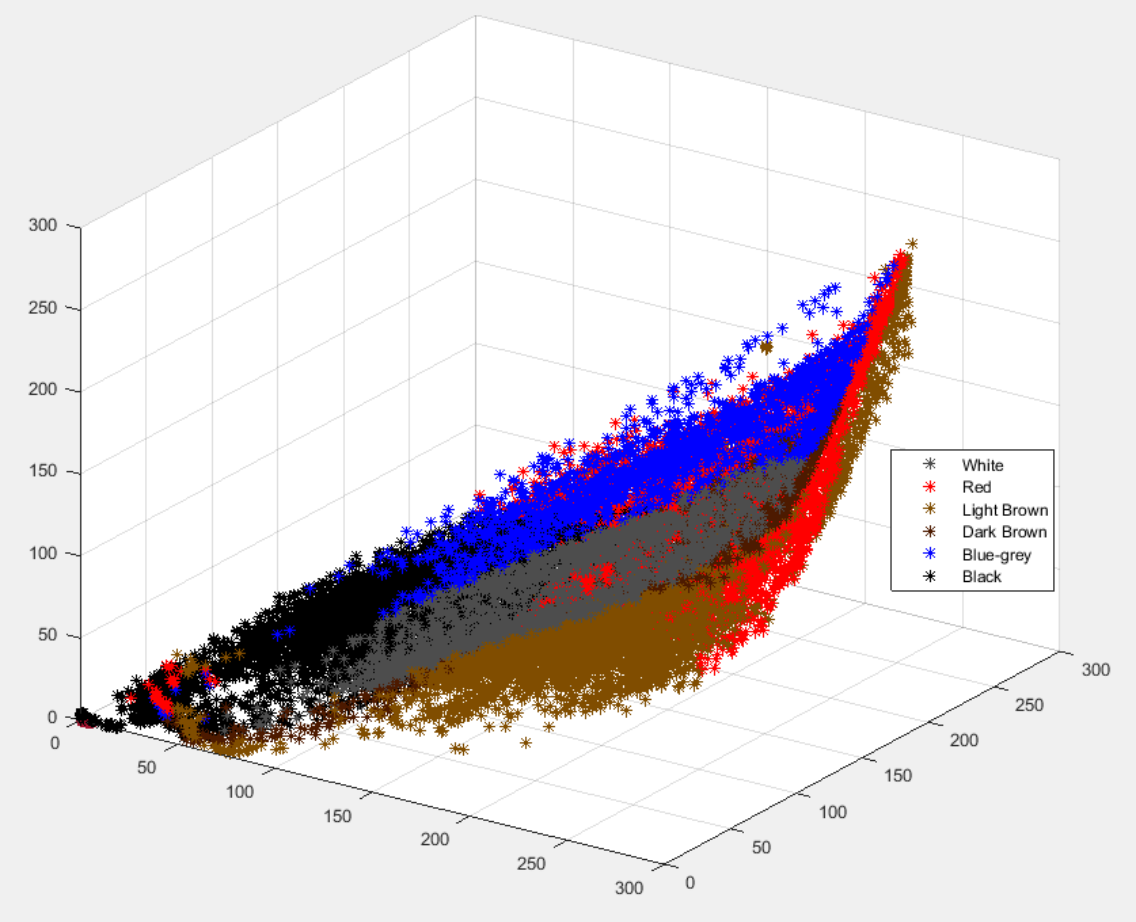


Figure 29. Plot of the class membership obtained for a test dataset, based on the Multivariate Gaussian classification.

Gaussian Mixture Model

A mixture of Gaussians is a probability distribution whose probability density function is a combination of a finite number of multivariate Gaussian probability density functions with specific probabilities or mixing coefficients. It can be expressed as the sum of weighted Gaussian components (15).

$$p(\vec{x}) = \sum_{c=1}^{C} w_c * N(\vec{x}|\vec{\mu}_c, \sum_c) \tag{15}$$

where C is the number of components, w_c is the weight for component c, and $\vec{\mu}_c, \sum_c$ are the mean vector and covariance matrix for component c.

Particularizing for this project, the meaning of using GMM is that each color will be defined by a mixture of Gaussians, each of which can be seen as a shade or *subcolor*.
GMM provide more complex classes of density models, which render them more capable of modeling realistic problems when compared to regular Gaussian distributions.
Since the parameters that describe the Gaussian components are unknown, the Maximum Likelihood estimation is employed to fit the data. The ML criterion attempts to estimate the model's parameters by maximizing the log-likelihood (16).

$$\ln p(\boldsymbol{X}|\vec{w}, \vec{\mu}, \sum) = \sum_{n=1}^{N} \ln \left\{ \sum_{c=1}^{C} w_c * N(\vec{x}|\vec{\mu}_c, \sum_c) \right\} \tag{16}$$

where $\boldsymbol{X}$ is a data matrix with N mixtures of Gaussians of C components.

However, when solving for Maximum Likelihood, Gaussian mixture models do not offer a closed solution. This lack of solution is explained due to singularities and the summation over C. The first problem occurs when the mean of a Gaussian component is equal to a data point (the Gaussian would collapse into a point and if the likelihood function goes to infinity, so does the log likelihood function). The second issue is only relevant when trying to interpret the obtained parameters.

The EM algorithm is an iterative technique that obtains the ML estimate in the presence of latent variables (they are not directly observed and must be inferred from the available data) so it is often applied to problems with missing or hidden data.

The first point in the process, given observed variables $\mathbf{X}$, latent variables $\mathbf{Z}$ and parameters $\boldsymbol{\theta}$, is to choose initial values for the parameters $\boldsymbol{\theta}$.
Then, in the E (expectation) step, the parameters are evaluated for the posterior distribution of the latent variables (17).

$$p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}) \tag{17}$$

This posterior is used to compute the expectation (18).

$$Q(\boldsymbol{\theta},\boldsymbol{\theta}') = \sum_{Z} p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}) ln(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}) \tag{18}$$

Next, the M (maximization) step maximizes the expectation.

$$\boldsymbol{\theta}' = argmax_{\boldsymbol{\theta}}\, Q(\boldsymbol{\theta},\boldsymbol{\theta}') \tag{19}$$

The E phase evaluates the log-likelihood with the estimated parameters and the M phase maximizes the log-likelihood to estimate new parameters.
Starting with the E step, the EM algorithm is performed iteratively until a stop condition is met. Given that each step increases the log-likelihood of the model, such a condition is usually established to be a log-likehood increment that is smaller than an input threshold.

In the context of GMM, the parameters $\boldsymbol{\theta}$ stand for the means $\vec{\mu}_c$, covariances $\sum_c$, and mixing coefficients w_c. The posterior evaluated in step E is the responsibility γ_c that a certain component C has over observation x, which is simply the normalized weighted likelihood for observation $\boldsymbol{x}_n$ (20).

$$\gamma_{nk} = \frac{w_k N(\vec{x}|\vec{\mu}_k,\sum_k)}{\sum_{c=1}^{C} w_c N(\vec{x}|\vec{\mu}_c,\sum_c)} \tag{20}$$

The responsibilities are used to update the parameters in the M step.

$$\vec{\mu}_k{}^{new} = \frac{1}{N_k}\sum_{n=1}^{N} \gamma_{nk}(\vec{x}_n) \tag{21}$$

$$\sum_k{}^{new} = \frac{1}{N_k}\sum_{n=1}^{N} \gamma_{nk}(\vec{x}_n - \vec{\mu}_c{}^{new})(\vec{x}_n - \vec{\mu}_c{}^{new})^T \tag{22}$$

$$w_c{}^{new} = \frac{N_k}{N} \tag{23}$$

The EM algorithm is highly sensitive to initialization. In this instance, it is initialized applying K-means++. First, it assumes a number k of clusters, an initial mean or centroid selected uniformly from all data points, and diagonal covariance matrices whose elements are the variances of each data component. The subsequent centroids are found through a recurrent process. It computes the Mahalanobis distances from data points to each centroid and then chooses one with a probability proportional to the distance to the current centroid. This step is repeated until k centroids are selected.

The method that was just described provides models such as the following.

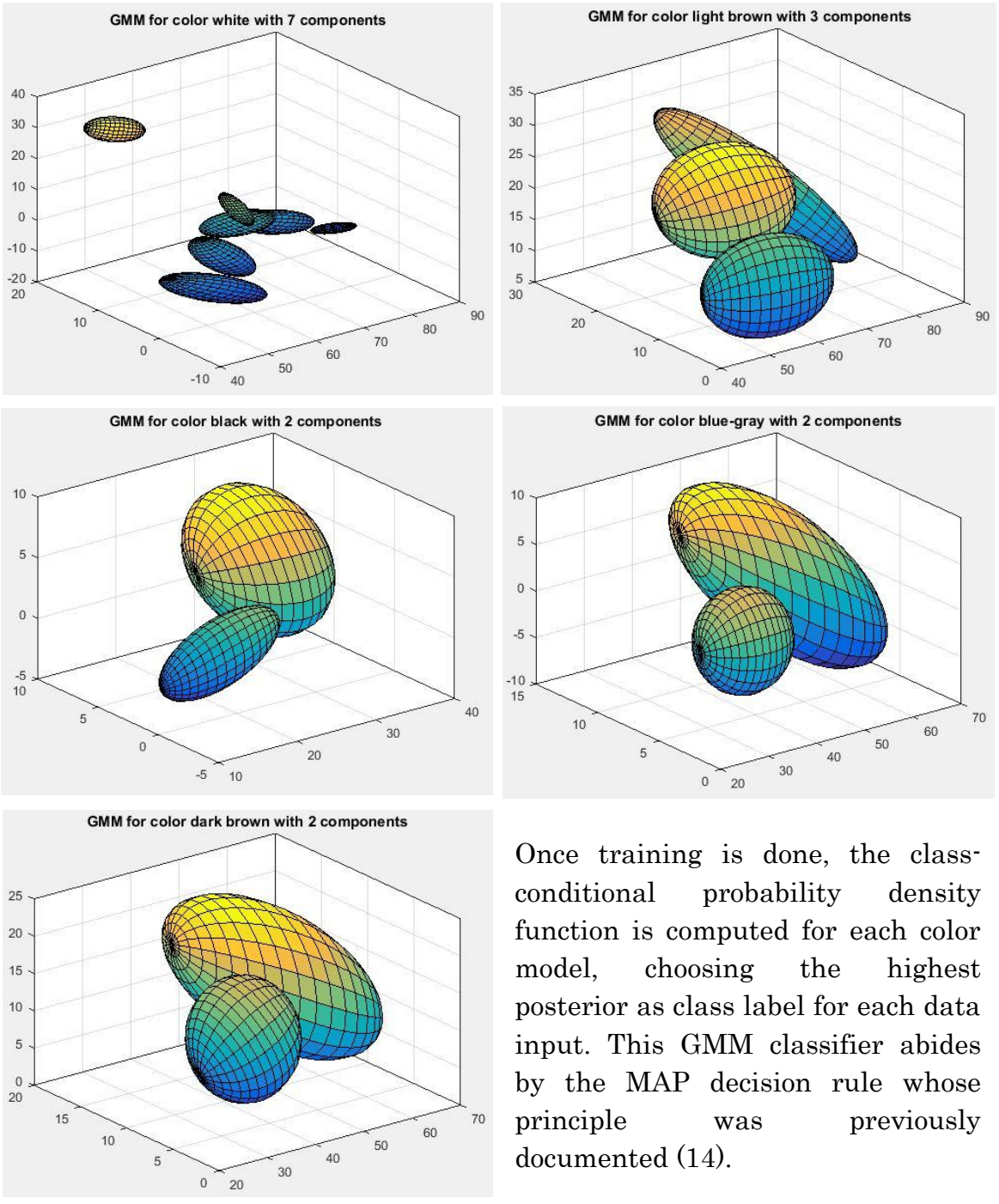


Figure 30. Gaussian Mixture Models for each color.

Once training is done, the class-conditional probability density function is computed for each color model, choosing the highest posterior as class label for each data input. This GMM classifier abides by the MAP decision rule whose principle was previously documented (14).

7.5. SUGGESTION OF COLORS PRESENT ON THE LESION

Color inspection in pigmented lesions implies examining the number of colors and how they are distributed. The presence of two or more colors as well as an irregular chromatic distribution are tell-tale signs of potentially cancerous lesions.

The work illustrated in this document aims to provide a suggestion regarding which colors are found in a lesion. It is based on the six most relevant colors in dermoscopy (white, red, light brown, dark brown, blue gray and black). The result for each dermoscopic image is a number of vectors equal to its number of regions. Each of those vectors has six values that correspond to the probability of existence of each color.

Region $i \rightarrow$	○	●	●	●	●	●
	P(white | i)	P(red | i)	P(light-brown | i)	P(dark-brown | i)	P(blue-grey | i)	P(black | i)

Furthermore, hard decisions were included as an additional result based on the highest class posterior probability for each region. Thus, a region is assumed to be the most likely color. Applying this concept, lesions can be reconstructed according to the predicted color labels.

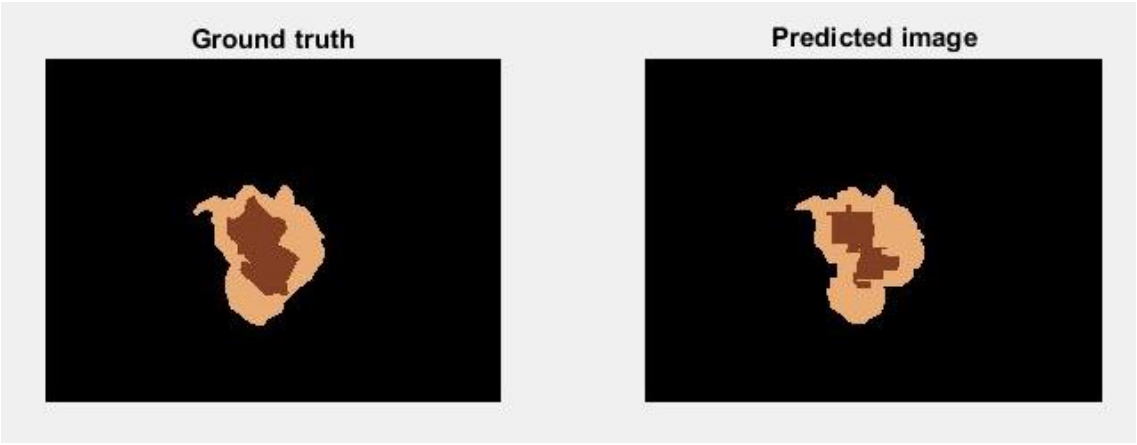


Figure 31. Labeled lesion IMD374: ground truth (L) and classifier result (R).

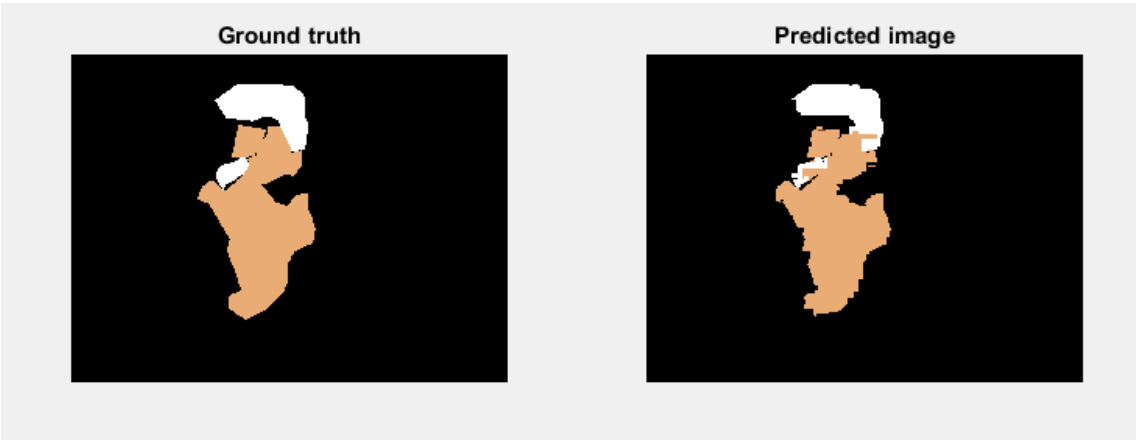


Figure 32. Labeled lesion *IMD035*: ground truth (L) and classifier result (R).

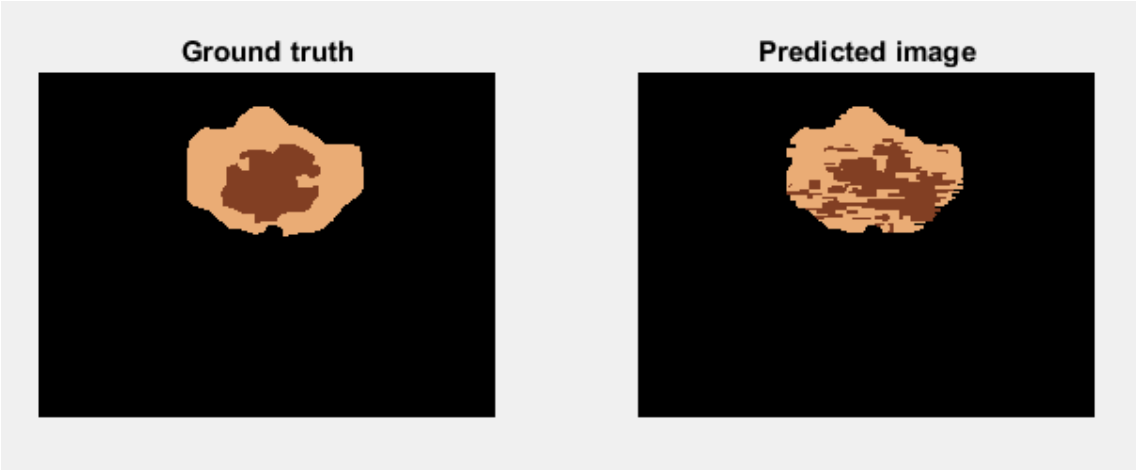


Figure 33. Labeled lesion *IMD146*: ground truth (L) and classifier result (R).

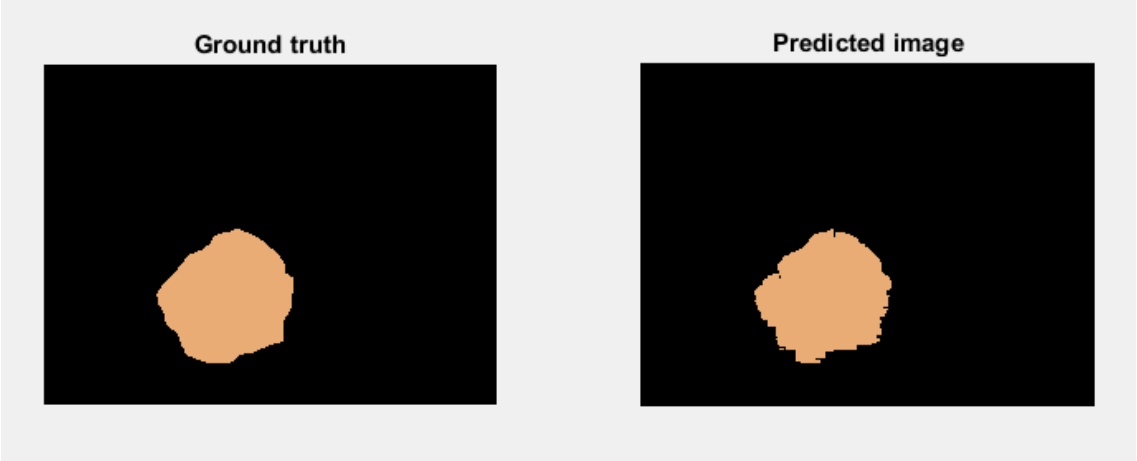


Figure 34. Labeled lesion *IMD383*: ground truth (L) and classifier result (R).

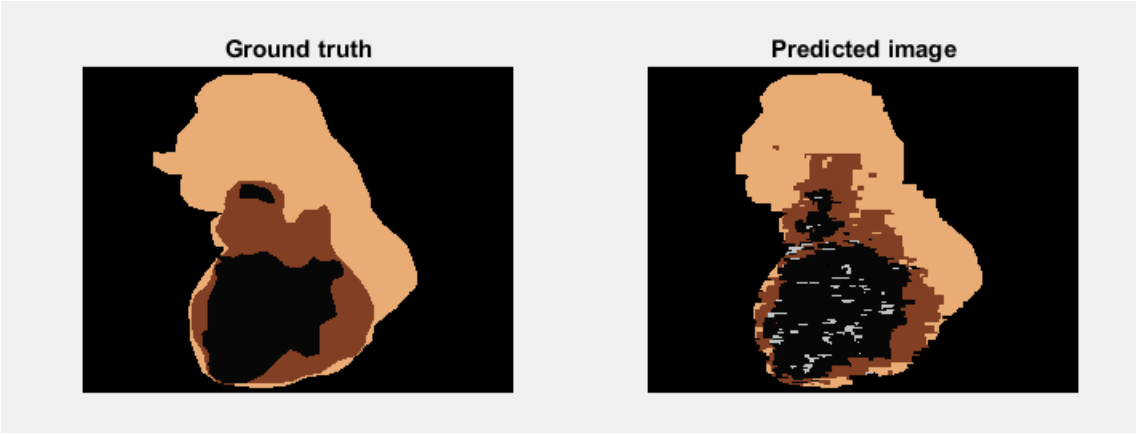


Figure 35. Labeled lesion *IMD407*: ground truth (L) and classifier result (R).

8. EXPERIMENTAL SETUP

The experiments were the means through which the algorithms would be put to the test in order to appraise the performance of the color analysis executed by the system.
The state of the art of medically-oriented image processing software has shaped the software-related decisions made for this thesis. Matlab was the obvious choice for the experiments. First of all, there is a large community of users and researchers, which is testament for its capacity and establishes a broad map of shared knowledge. Moreover, one of Matlab's best assets is its clear and extensive documentation and its capability to read a wide variety of image formats (over fifteen types). Additionally, the work environment streamlines data management, which is easy to store and keep track of, making the debugging process more straightforward, because testing changes is uncomplicated. Also, Matlab possesses a vast library of bi-annually updated algorithms. Another advantage is the numerical accuracy for uttermost precision that allows images to have double precision pixels throughout the whole process.

The experimental arrangement incorporates three different segmentation algorithms:

- K-Means
- N-Cuts
- Mean Shift

For every segmented dataset, two different statistical models will be constructed:

- Multivariate Gaussian Model
- Gaussian Mixture Model

Each of these combinations will be analyzed using four different evaluation measurements:

- Balanced accuracy
- F-Measure
- Confusion matrix
- Receiver Operating Curve

8.1. DATABASE DESCRIPTION AND PREPARATION

The image dataset used for this study came from the ADDI (Automatic computer-based Diagnosis system for Dermoscopy Images) project [7] [31]. This program gathers experts from the medical and academic field (Pedro Hispano Hospital, University of Porto, University of Aveiro and Instituto Superior Técnico) who seek to develop an automatic dermoscopic image analysis system.

The PH^2 database intends to facilitate comparative studies regarding the topic of dermoscopic images. It contains 200 (8-bit RGB, with 768x560 px resolution) pictures of melanocytic lesions taken at the Dermatology Service of Hospital Pedro Hispano in Matosinhos, Portugal. The tool employed for the task was the Tuebinger Mole Analyzer with a 20x magnification.
The database includes benign (melanocytic nevi) and malignant (melanoma) lesions: 80 common nevi, 80 atypical nevi and 40 melanomas.
Besides, the dataset contains the binary mask of the segmented lesions and the binary masks of the color classes found in them.
All images are medically documented by an expert dermatologist, according to well-known parameters: colors, pigment network, presence of dots, globules and streaks, regression areas and blue-whitish veil. This information is collected in an attached file, later used for system evaluation.

The preparation of the data entailed creating an structure array whose fields mirrored the given information, as text or as paths to files. As for the training set retrieval, 70% of the images were selected at random, using the remaining 30% for test purposes. This random partition is preserved for all the experiments, so as to guarantee legitimate comparisons. Even though classification is performed on the basis of regions, each percentage was ensured to contain every region of an image.

8.2 SMALL DATASETS PROBLEM INSIGHTS

The main issue that stemmed from the dataset's small size was the lack of color representation. Specifically, the color red was found in only 5% of the total set of lesions. Its limited presence made the characterization of its corresponding model highly inaccurate, and consequently gave an unreliable training result. As expected, performance assessment after testing showed poor efficiency for that class. This led to the decision of neglecting said color. A similar circumstance was found for the white model, present in only 10% of the lesions, although parametric modeling was possible.

8.3. EVALUATION MEASURES FOR THE PERFORMANCE OF THE SYSTEM

In order to assess the functionality of the classifier, several performance measures are employed.

Balanced Accuracy

Accuracy refers to the closeness of the estimated result to the true value. For this specific task, accuracy provides information about the proximity of the predictions, which are the color decisions inferred for each region, to the known colors, found in the ground truth images. This concept can be expressed as follows:

$$Accuracy = \frac{Number\ of\ regions\ correctly\ predicted}{Number\ of\ predictions} = \frac{\sum TP}{N} \tag{24}$$

where TP is the number of true positives (predictions match ground truth) and N stands for the total number of test regions.

Due to the multi-label nature of the classifier, the accuracy metric is not good enough. Balanced accuracy is an alternative measure that intends to solve the issues that arise from accuracy evaluation for multi-class and unbalanced cases [32]. Balanced accuracy is defined as follows:

$$Balanced\ Accuracy\ = \frac{1}{n}\sum_{i}\frac{TP_i}{N_i} \tag{25}$$

where n is the number of classes, TP_i is the number of correct predictions for class i and N_i is the number of test samples that belong to class i.

Due to the disparity in the number of class members for each color, the accuracy value might be affected, giving an unrealistic result. Balanced accuracy accounts for this possibility by dealing with each class independently.

F-Measure

The F-measure gives further information on the classifier's accuracy by combining two scores related to both type I (false positive) and type II (false negative) errors. False positive errors for a given class i refer to incorrect assignments of label i. False negative errors for class i indicate a wrong decision that does not classify a point as a member of i.

$$F - measure = \frac{2 * Precision * Recall}{Recall + Precision} \tag{26}$$

$$Precision = \frac{TP_i}{TP_i + FP_i} \tag{27}$$

$$Recall = \frac{TP_i}{TP_i + FN_i} \tag{28}$$

where TP_i is defined like in (25); FP_i and FN_i are, respectively, the number of false positives (wrong color is predicted) and false negatives (color is wrongly not predicted) for class *i*.

Confusion Matrix

A confusion or error matrix is a set of measures that relate the predictions for each class to the actual class memberships.
This evaluation measure has the advantage of establishing relationships between each class and the rest, which can be of help to expose if the system is mixing up certain categories that might be similar.
The rows are attributed to the actual classes, whilst the columns are designated for the predictions. In consequence, the matrix diagonal contains the True Positives for each color.

		Predicted				
	Color 1	**Color 2**	**Color 3**	**Color 4**	**Color 5**	**Color 6**
Color 1	TP1	FP2,1	FP3,1	FP4,1	FP5,1	FP6,1
Color 2	FP2,1	TP2	FP3,2	FP4,2	FP5,2	FP6,2
Actual **Color 3**	FP3,1	FP2,3	TP3	FP4,3	FP5,3	FP6,3
Color 4	FP4,1	FP2,4	FP3,4	TP4	FP5,4	FP6,4
Color 5	FP5,1	FP2,5	FP3,5	FP4,5	TP5	FP6,5
Color 6	FP6,1	FP2,6	FP3,6	FP4,6	FP5,6	TP6

Table 6. Confusion matrix.

where TP_i was defined in (25); $FP_{i,j}$ represents an erroneous prediction of color class *i* instead true label *j*.

Receiver Operating Curve

The Receiver Operating Curve (ROC) graphically conveys the ratio of correctly classified test samples versus data erroneously assigned to that same class; i.e., it plots true positive rate or recall (28) against false positive rate or inverse recall (29).

$$Inverse\ recall = \frac{TN_i}{FP_i + TN_i} \tag{29}$$

where FP_i was defined in (27) and TN_i stands for the number of true negatives (color is appropriately not predicted) for class i.

ROC curves, unlike previous measures, acknowledge the class probability estimates. At optimum performance, the results are maximized on both axis.

9. EXPERIMENTAL RESULTS AND DISCUSSION

In order to evaluate the classifier, several configurations of the system were executed. All three segmentation algorithms along with both parametric models are assessed according to the metrics previously explained.

The **k-means** algorithm produced the results seen below.

Confusion Matrix				
	Light brown	Dark brown	Blue-grey	Black
Light brown	6517	1697	312	89
Dark brown	1895	4392	696	1834
Blue-grey	94	298	1287	1342
Black	1	27	64	536
F-Measure	0.76	0.58	0.48	0.24
Balanced Accuracy		0.63		

Table 7. Performance evaluation measures for K-means segmentation with Gaussian Mixture Model classification.

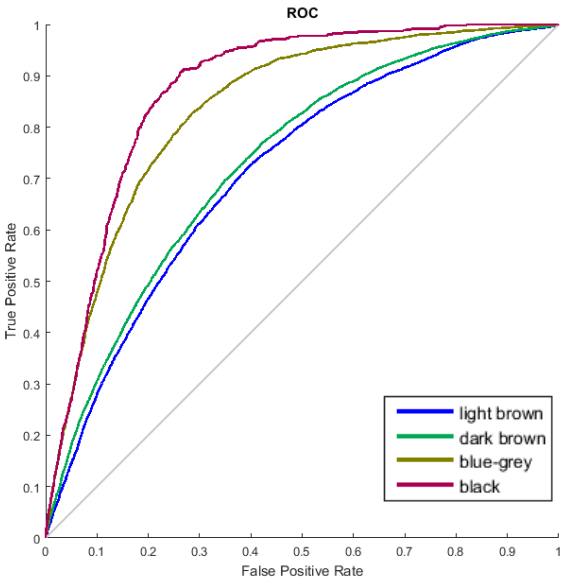


Figure 36. ROC for K-means segmentation with Gaussian Mixture Model classification.

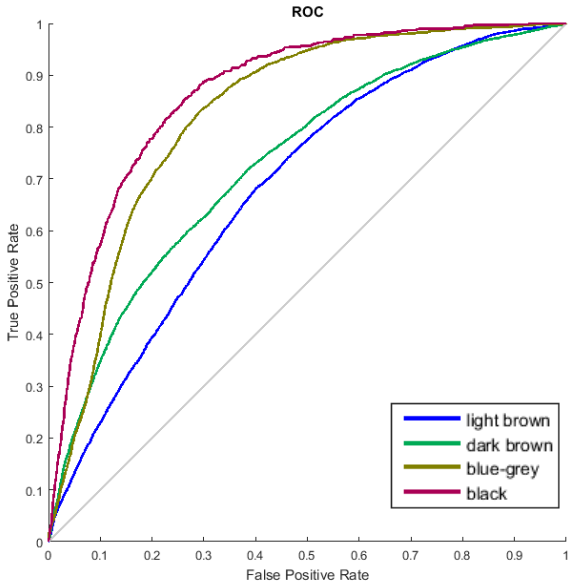


Figure 37. ROC for K-means segmentation with Multivariate Gaussian Model classification.

Confusion Matrix				
	Light brown	Dark brown	Blue-grey	Black
Light brown	6561	1709	279	66
Dark brown	2323	4135	1173	1186
Blue-grey	77	329	1449	1166
Black	1	65	141	421
F-Measure	0.75	0.55	0.48	0.24
Balanced Accuracy		0.60		

Table 8. Performance evaluation measures for K-means segmentation with Multivariate Gaussian Model classification.

Next, the **n-cuts** algorithm was the segmentation method employed.

Confusion Matrix				
	Light brown	Dark brown	Blue-grey	Black
Light brown	284	73	14	6
Dark brown	81	178	29	57
Blue-grey	2	35	133	27
Black	0	3	2	10
F-Measure	0.76	0.56	0.71	0.17
Balanced Accuracy		0.65		

Table 9. Performance evaluation measures for N-cuts segmentation with Gaussian Mixture Model classification.

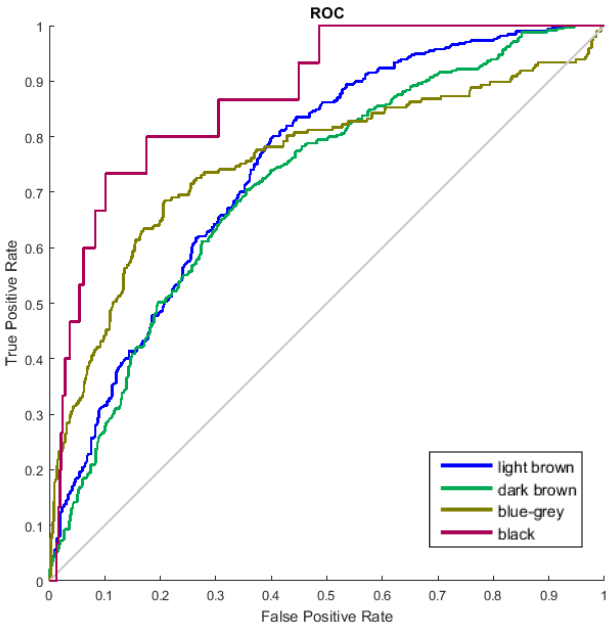


Figure 38. ROC for N-cuts segmentation with Gaussian Mixture Model classification.

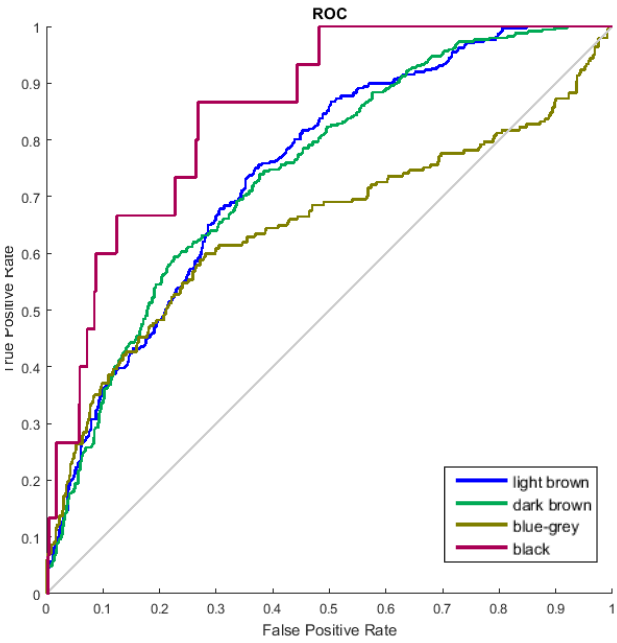


Figure 39. ROC for N-cuts segmentation with Multivariate Gaussian Model classification.

Confusion Matrix				
	Light brown	Dark brown	Blue-grey	Black
Light brown	303	52	18	4
Dark brown	109	115	43	78
Blue-grey	1	10	144	45
Black	0	0	3	12
F-Measure	0.77	0.44	0.70	0.16
Balanced Accuracy		0.54		

Table 10. Performance evaluation measures for N-cuts segmentation with Multivariate Gaussian Model classification.

Finally, the **mean shift** algorithm was applied to the lesions, which were then classified.

Confusion Matrix				
	Light brown	Dark brown	Blue-grey	Black
Light brown	9397	1370	701	30
Dark brown	4600	6712	1533	4418
Blue-grey	366	725	11267	5588
Black	8	188	567	1820
F-Measure	0.73	0.51	0.70	0.25
Balanced Accuracy		0.65		

Table 11. Performance evaluation measures for Mean Shift segmentation with Gaussian Mixture Model classification.

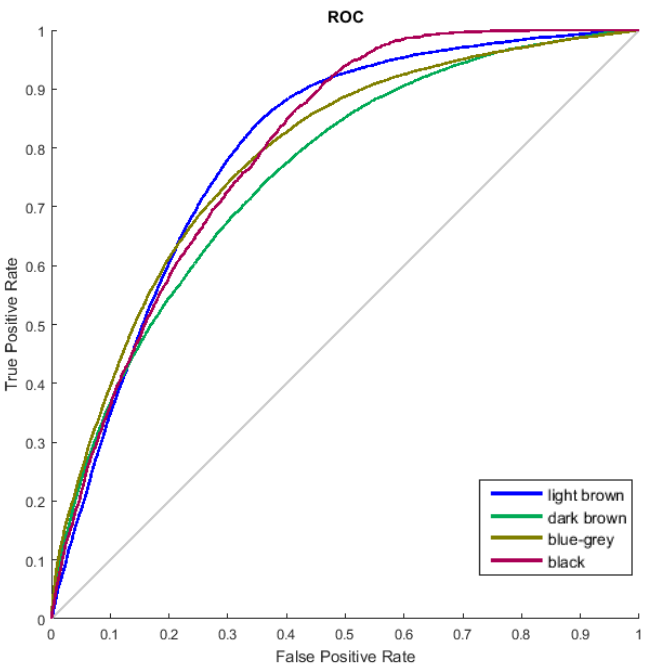


Figure 40. ROC for Mean Shift segmentation with Gaussian Mixture Model classification.

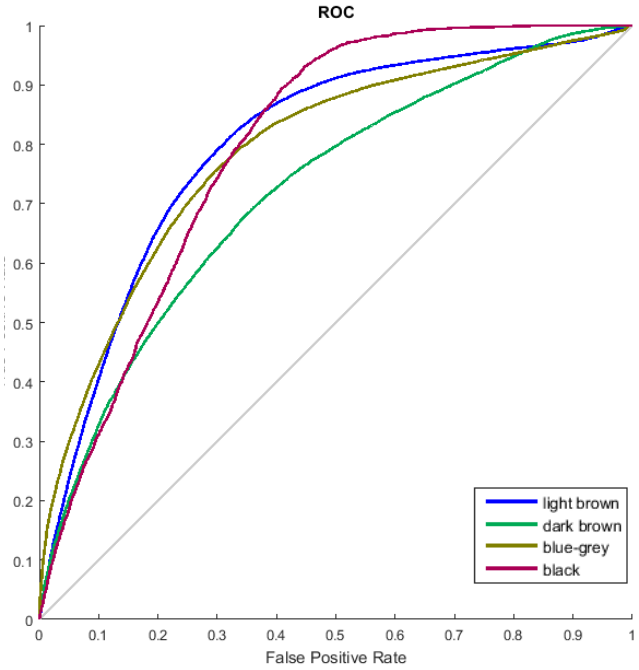


Figure 41. ROC for Mean Shift segmentation with Multivariate Gaussian Model classification.

Confusion Matrix				
	Light brown	Dark brown	Blue-grey	Black
Light brown	9210	1232	1021	35
Dark brown	4215	6185	1215	5648
Blue-grey	266	649	9939	7092
Black	5	78	508	1992
F-Measure	0.73	0.49	0.65	0.23
Balanced Accuracy		0.60		

Table 12. Performance evaluation measures for Mean Shift segmentation with Multivariate Gaussian Model classification.

Multiple conclusions can be drawn from the classifier outputs. The first thing that should be addressed is the drawback imposed by the lack of samples for color white. While color red (also deficient in the available images) was excluded from the classification process, white regions did allow for model training with fair test results (Figure 42), yielding a balanced accuracy equal to 0.62 at a 0.2 baseline.

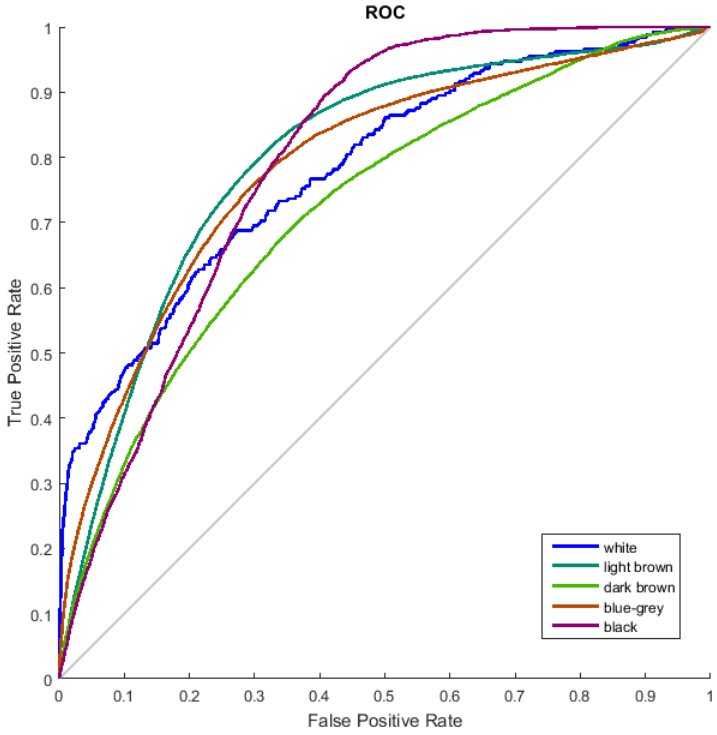


Figure 42. ROC for Mean Shift segmentation with Gaussian Mixture Model (includes white class).

Nevertheless, it is more reasonable to provide representative results to extract unbiased conclusions.

Overall, in relation to color segmentation, the mean shift algorithm grants superior performance levels related to a couple of factors: it benefits from low-dimensional data, such as pixel intensity, and it clusters points attending to both location and color but, unlike K-means, it is not sensitive to outliers. Predictably, the normalized cuts method underperforms due to its inherent nature, as it only uses spatial data.

The various results attained for each implementation hint at the strength of Gaussian mixture models to describe complex features. A mixture of Gaussians offers flexibility to model any distribution shape. This is relevant to this system because it prevents erroneous class assignations attributed to color similarity. As manifested by the confusion matrices, this is very likely to happen for the brown shades. GMM are favorable because every class defined as such is more intricate, as a combination of latent subcolors.

10. CONCLUSIONS AND FUTURE WORK

The presented solution offers a solid base for color-based feature classification in dermoscopic lesions. The optimum setup requires a segmentation technique that focuses on color and gathers plenty of information to build a complex enough parametric color model. This renders K-means (with spatial information) and mean shift algorithms, as well as Gaussian mixture models, perfectly suited for the task. By virtue of the fulfillment of the segmentation objectives, diagnostic features based on color can be generated from it.
The system provided satisfying balanced accuracy results for a multiclass classifier, being able to visibly disclose the color of regions.

With regards to future lines of work, the system introduced by means of the present document could be upgraded to interpret at a larger scale the results obtained for each dermoscopic image, regarding its malignancy. Higher levels of performance could be achieved with a larger, more balanced database, as the lack thereof could be considered one of the main hindrances encountered throughout the development of the project.

Of course this tool could be paired with feature classifiers that explore other melanoma characteristics, such as texture, to provide a more wholesome diagnosis.

The work accomplished in this project offers a powerful, stand-alone assisting tool for color analysis. Color inspection is regarded as a key marker for malignant lesions in early melanoma detection, which is clinically crucial [33]. In fact, a more direct application is found in clinical principles like "ugly duckling", a sign that indicates risk of melanoma when a patient displays a particular lesion that looks different to the rest. E.g., a blue-grey lesion among predominantly brown ones. This indicator is remarkably sensitive for detecting melanoma [34].

Skin cancer is a pressing matter and there is still a lot of progress to be done in this area, especially considering the intersection of the vast world of computer vision with the interest and efforts of researchers. Experts in the field of melanoma detection can use all the help they can get, and this project constitutes one step towards a collaboration that ensures trustworthy diagnosis, as early as possible.

REFERENCES

[1] D. S. Gareau et al., "Digital imaging biomarkers feed machine learning for melanoma screening," *Experimental Dermatology*, no. 1600-0625, 2016.

[2] Catarina Barata, Mario A. T. Figueiredo, and M. Emre C, "Color identification in dermoscopy images using Gaussian mixture," 2014.

[3] American Cancer Society. Key Statistics for Melanoma Skin Cancer. [Online]. https://www.cancer.org/cancer/melanoma-skin-cancer/about/key-statistics.html

[4] AECC. Incidencia melanoma. [Online]. https://www.aecc.es/SobreElCancer/CancerPorLocalizacion/melanoma/Paginas/incidencia.aspx

[5] Agencia Española de Protección de Datos. Cesión. [Online]. https://www.agpd.es/portalwebAGPD/canalresponsable/obligaciones/tratamiento_cesion/cesion/index-ides-idphp.php

[6] European Commission. Protection of personal data. [Online]. http://ec.europa.eu/justice/data-protection/

[7] Faculdade de Ciências da Universidade do Porto. ADDI Project. [Online]. http://www.fc.up.pt/addi/

[8] MetaOptima. MoleScope. [Online]. https://molescope.com/product/

[9] Oliver Griffin. (2017, January) SkinVision unveils breakthrough algorithm that takes the fight to skin cancer. [Online]. http://startupbeat.com/2017/01/cancer-app-skinvision-breakthrough-algorithm/

[10] Dermoscopy: ABCD Rule. [Online]. http://www.dermoscopy.org/atlas/base.htm

[11] Dermoscopy: 7-point Checklist. [Online]. http://www.dermoscopy.org/atlas/base.htm

[12] Kajsa Møllersen et al., "Computer-Aided Decision Support for Melanoma Detection Applied on Melanocytic and Nonmelanocytic Skin Lesions: A Comparison of Two Systems Based on Automatic Analysis of Dermoscopic Images," *BioMed Research International*, no. 579282,

2015.

[13] Al-Jumaily, Ammara Masood, and Adel Ali, "Computer Aided Diagnostic Support System for Skin Cancer: A Review of Techniques and Algorithms," *International Journal of Biomedical Imaging*, vol. 2013, no. 323268, 2013.

[14] Vanessa Ngan. Spectrophotometric analysis of skin lesions. [Online]. http://www.dermnetnz.org/topics/spectrophotometric-analysis-of-skin-lesions/

[15] Neeraj Jasrotia, Er. Satnam Singh, and Er.Vikas Saini, "The Review on the Color Constancy based Dermoscopy," *International Journal of Advance Research in Education, Technology & Management*, vol. 5, no. 1.

[16] Maryam Sadeghi, Towards prevention and early dianosis of skin cancer: computer-aided analysis of dermoscopic images, 2012.

[17] Derma Medical Systems. About MoleMax. [Online]. https://www.dermamedicalsystems.com/index.php?menu_id=104&locale=en

[18] ddax3. DBMips System. [Online]. http://www.ddax3.com/eng/dbmips.html

[19] Jacob Scharcanski and M. Emre Celebi, *Computer Vision Techniques for the Diagnosis of Skin Cancer*.

[20] Margarida Ruelaa, Catalina Barata, Jorge S. Marques, and Jorge Rozeira, "A System for the Detection of Melanomas in Dermoscopy Images Using," *Computer Methods in Biomechanics and Biomedical Engineering*.

[21] Catarina Barata, Jorge S. Marques, and Jorge Rozeira, "A System for the Detection of Pigment Network in Dermoscopy Images Using Directional Filters," *IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING*, vol. 59, no. 10, October 2012.

[22] Rok Kreslin, Pilar M. Calvo, Luis G. Corzo, and Peter Peer, "Linear Chromatic Adaptation Transform Based on," *Mathematical Problems in Engineering*, vol. 2014, no. 760123, p. 9.

[23] Jixiang Chen, R. Joe Stanley Stanley, Randy H. Moss, and William Van Stoecker, "Color analysis of skin lesion regions for melanoma discrimination in clinical images," *Skin Res Technol.*

[24] R. Joe Stanley, William V. Stoecker, and Randy H. Moss, "A relative color approach to color discrimination for malignant melanoma detection in dermoscopy images," *Skin Res Technol.*

[25] Mounika Lingala et al., "Fuzzy logic color detection: Blue areas in melanoma dermoscopy images," *Comput Med Imaging Graph*, 2014.

[26] William V. Stoecker et al., "Detection of Granularity in Dermoscopy Images of Malignant Melanoma Using Color and Texture Features," *Comput Med Imaging Graph*, vol. 35, no. 2, pp. 144-147, March 2011.

[27] Catarina Barata, Jorge S. Marques, and M. Emre Celebi, "IMPROVING DERMOSCOPY IMAGE ANALYSIS USING COLOR CONSTANCY," in *ICIP*, 2014.

[28] G. Finlayson and Elisabetta Trezzi, "Shades of gray and colour constancy," in *IS&T/SID Twelfth Color Imaging Conference*, 2004, pp. 37-41.

[29] T. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Comm. in Statistics*, vol. 3, no. 1, pp. 1-27, 1974.

[30] Jianbo Shi and Jitendra Malik, "Normalized Cuts and Image Segmentation," *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, vol. 22, no. 8, October 2000.

[31] Teresa Mendonça, Pedro M. Ferreira, Jorge Marques, Andre R. S. Marcal, and Jorge Rozeira, "35th International Conference of the IEEE Engineering in Medicine and Biology Society," in *PH² - A dermoscopic image database for research and benchmarking*, Osaka, 2013.

[32] Kay H. Broderse, Cheng Soon Ong, Klaas E. Stephan, and Joachim M. Buhmann, "The balanced accuracy and its posterior distribution," in *2010 International Conference on Pattern Recognition*, 2010.

[33] M. C. Mihm et al., "Early Detection of Primary Cutaneous Malignant Melanoma — A Color Atlas," *The New England Journal of Medicine*.

[34] A. Scope, S.W. Dusza, A.C. Halpern, H., Braun, R.P. Rabinovitz, and I. Zalaudek, "The “ugly duckling” sign: agreement between observers,"

Arch Dermatol, vol. 144, pp. 58–64, 2008.

[35] Fco. Javier Martínez. El Color (I). [Online]. http://www.qvision.es/blogs/javier-martinez/2011/02/13/el-color-i/

[36] RGB color model. [Online]. http://www.wikiwand.com/en/RGB_color_model

[37] Color spaces. [Online]. http://www.color-management-guide.com/color-spaces.html

[38] NationsMaster. Kruskal's algorithm. [Online]. http://www.statemaster.com/encyclopedia/Kruskal's-algorithm