Computer Vision and Image Understanding 151 (2016) 124–137

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

# Clinically inspired analysis of dermoscopy images using a generative model

Catarina Barata [a,*], M. Emre Celebi [b], Jorge S. Marques [a], Jorge Rozeira [c]

[a] Institute for Systems and Robotics, Instituto Superio Técnico, Lisboa, Portugal
[b] Louisiana State University, Shreveport LA, USA
[c] Hospital Pedro Hispano, Matosinhos, Portugal

ARTICLE INFO

*Article history:*
Received 14 February 2015
Accepted 22 September 2015

*Keywords:*
Melanoma
Dermoscopy
Computer-aided diagnosis
Color detection
ABCD rule
Correspondence-LDA
Von-Mises-Gaussian distribution

ABSTRACT

Dermatologists often prefer clinically oriented Computer Aided Diagnosis (CAD) Systems that provide medical justifications for the estimated diagnosis. The development of such systems is hampered by the lack of detailed image annotations (medical labels and segmentations of the associated regions). In most cases, we only have access to weakly annotated images (text labels) that are not sufficient to learn proper models. In this work we address this issue and propose a CAD System that uses medically inspired color information to diagnose skin lesions. We deal with the weakly annotated dermoscopy images using the Correspondence-LDA algorithm to learn a probabilistic model. The algorithm is applied with success to the identification of relevant colors in dermoscopy images, obtaining an average Precision of 83.8% and a Recall of 89.8%. The proposed color representation is then used to classify skin lesions, resulting in a Sensitivity of 77.6% and Specificity of 73.0% using Random Forests, and a Sensitivity of 75.1% and Specificity of 77.5% using SVM. These results comparable favorably with related works.



## 1. Introduction

The most dangerous characteristic of melanoma is its ability to grow and spread rapidly to other tissues and organs [1]. This makes melanoma the deadliest form of skin cancer, although it is by far one of the less common types of skin related neoplasms. According to the most recent data, the incidence rate of melanoma has been steadily increasing for the past three decades and it currently ranks in the ninth position among the most common types of cancer in Europe alone [2]. The advanced stage of melanoma is often incurable and leads to the death of the patient, but an early diagnosis of this disease (when the abnormal growing cells are still contained within skin tissue) can lead to a full recovery [1]. Thus, a great effort has been put on the development of skin lesion visualization and diagnosis techniques, that can help dermatologists improve their diagnostic accuracies.

Dermoscopy is among the most popular imaging methods used by dermatologists, because it combines magnification and special illumination techniques that render an improved image of the skin lesion [3]. With this method, dermatologists are able to observe and analyze surface and subsurface structures that are invisible to the naked eye [1,4]. The observed structures, called dermoscopic criteria, play an important role in the diagnosis of melanoma and are considered in different medical procedures, such as the ABCD rule [5] and 7-point checklist [6]. The main drawback of dermoscopy is that it can only be effectively applied by trained practitioners [7]. Other negative characteristics of this method are its subjectivity and lack of reproducibility [8]. These drawbacks fostered the development of Computer Aided Diagnosis (CAD) systems, such as the ones described in [9–12] (see [13] for a survey on this topic), that can act as a second opinion tool and be used by non-experienced dermatologists [14].

Despite the interesting experimental results achieved by some of the CAD systems, dermatologists have pointed out that several of them have not been designed to work as a support tool [15]. The practitioners see these systems more as parallel/second opinion tools that give an output of melanoma or benign, without providing comprehensive medical information to justify the diagnosis. This black box structure and lack of interaction are two of the main reasons why dermatologists avoid including CAD systems in their routine practices. These two issues can be addressed with the development of more clinically oriented systems that focus not only on the diagnosis but also on the identification of key dermoscopic criteria (e.g., clinically relevant colors).

The development of clinically oriented CAD systems is an active topic of research. Different research groups have proposed strategies to detect the presence of dermoscopic criteria, such as pigment network [16–20], streaks [21,22], dots [23], and colors [24–27] or color

* Corresponding author.
*E-mail address:* ana.c.fidalgo.barata@ist.utl.pt (C. Barata).

http://dx.doi.org/10.1016/j.cviu.2015.09.011
1077-3142/

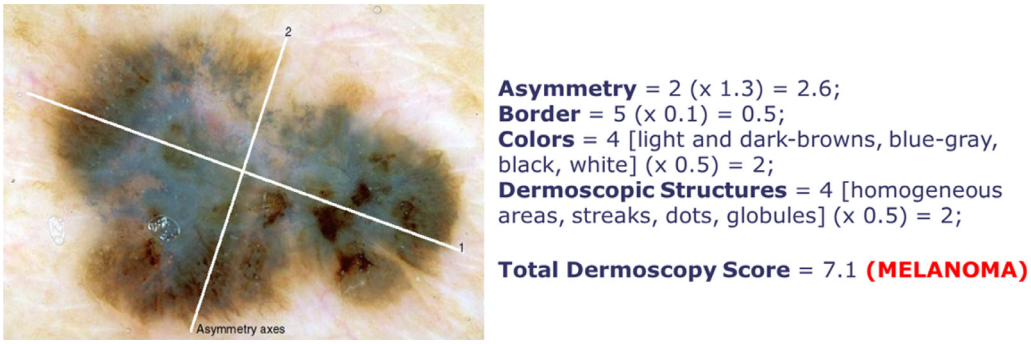


**Fig. 1.** Example of the application of the ABCD rule [1].

related structures [26–31]. However, only a few of these works use the detected criteria to obtain a diagnosis of the lesions as melanoma or benign [13], which demonstrates the difficulty of this problem.

One of the major challenges of developing a clinically oriented CAD system is that it might require a large number of detailed annotated images (i.e., medical text labels and segmentations of the relevant dermoscopic criteria). This detailed medical information must be obtained through consultation with experienced practitioners. Dermatologists usually provide text labels stating whether a dermoscopic criterion is present or absent, but do not perform their corresponding segmentations because it is a time consuming and subjective task. However, several of the methods described in the literature require detailed annotations (e.g., detection of colors [24–26], blue-whitish veil [28,30], and global patterns [32–34]), and can result in incomplete systems if the number of available segmentations is not sufficient. This limitation can be addressed through the design of systems that are capable of dealing with weakly annotated data (i.e., images for which there are text labels and it is not known which are the image regions that correspond to those labels). Such systems must be able not only to reproduce the medical labeling process in new images but also to identify the regions within the lesions that correspond to the text labels. Although one might argue that this last aspect is unnecessary as the system already provides text labels, it can be quite useful for dermatologists as it would allow them to associate the text outputs of the system with specific areas in the lesion and verify if the suggested output makes sense. It is also important for the designed systems to be able to diagnose the lesions as melanoma or benign using the detected medical criteria.

This work addresses the aforementioned problems and investigates the development of a clinically oriented CAD system, in which it is possible to learn a probabilistic model to represent the dermoscopic criteria using only medical text labels. The system is capable of i) reproducing the labeling process; ii) identifying the regions in the lesion associated with each of the labels, and iii) diagnosing the lesion as melanoma or benign. Various dermoscopic criteria could be used to study the labeling process. In this work we have selected the clinically relevant colors that are considered in the ABCD rule (Dark and Light Browns, Blue-Gray, Black, Red, and White) [5]. The selection of the color criterion is based both on the difficulty of the problem and on the fact that color detection systems usually require training examples of color segmentations. The probabilistic model used to learn the correlation between medical labels and image regions is Correspondence-LDA (corr-LDA) [35]. To the best of our knowledge this is the first time that such a model and approach are applied to the analysis and classification of dermoscopy images.

The paper is organized as follows. First we give an overview of the problem and the notation used (Section 2). Then, we discuss the state-of-the-art in annotation (Section 3), describe the probabilistic model (Section 4) and present the proposed modifications (Section 5). We discuss different possibilities to diagnose the skin lesions using the detected color information in Section 6. Finally we present the obtained results (Section 8) and conclude the paper (Section 9).

## 2. Problem formulation

### 2.1. Clinical analysis

A clinically oriented CAD system for the diagnosis of melanoma must have the following framework: i) identify relevant regions in the dermoscopy images and associate them with the dermoscopic criteria; ii) provide labels for the entire image stating whether the dermoscopic criteria are present or absent; and iii) use the identified medical information to estimate a diagnosis.

The first challenge that we must address is the selection of the dermoscopic criteria that must be identified by the developed CAD system. Medical procedures such as the ABCD rule [5] provide us with the necessary information regarding which are the criteria that dermatologists use to distinguish between benign lesions and melanomas. ABCD rule is a scoring approach that considers four different aspects of the lesion in order to obtain a diagnosis. The assessed criteria are: (A)symmetry regarding shape, color, and structures; irregular (B)orders; the number of (C)olors (up to six); and the existence of (D)ermoscopic structures, such as pigment network or streaks. During the diagnosis, dermatologists start by assigning an individual score to each of these criteria. Then, the scores are combined into a total lesion score using a weighted sum. The obtained score gives information about the level of suspiciousness/malignancy of the lesion. Fig. 1 shows an example of the ABCD rule [1].

In this work we address the detection of the clinically relevant colors considered by in the ABCD rule: Dark and Light Browns, Blue-Gray, Black, Red, and White (Fig. 2 shows some examples of lesions and the colors identified by experts). The detection of colors in dermoscopy images has already been addressed by some research groups [24–27,36]. Among these works, some require training examples of segmented color regions, which are not easy to obtain as was pointed out in the beginning of this paper. Other works do not use training examples and focus on the process of color quantification [27,36], usually using clustering methods, without actually identifying which are the colors that can be found in a given lesion. Our objective is to perform color identification and quantification first and then use this information to diagnose skin lesions. The main limitation is the lack of training examples of segmented color regions, since the segmentation of colors in dermoscopy images is a cumbersome and subjective task that is avoided by most dermatologists. Thus, we must investigate an alternative strategy that allows us to train a color model based on the available data.

### 2.2. Preliminary information and goals

Our dataset comprises $D$ dermoscopy images in which the lesions were divided into small non-overlapping square patches, as shown

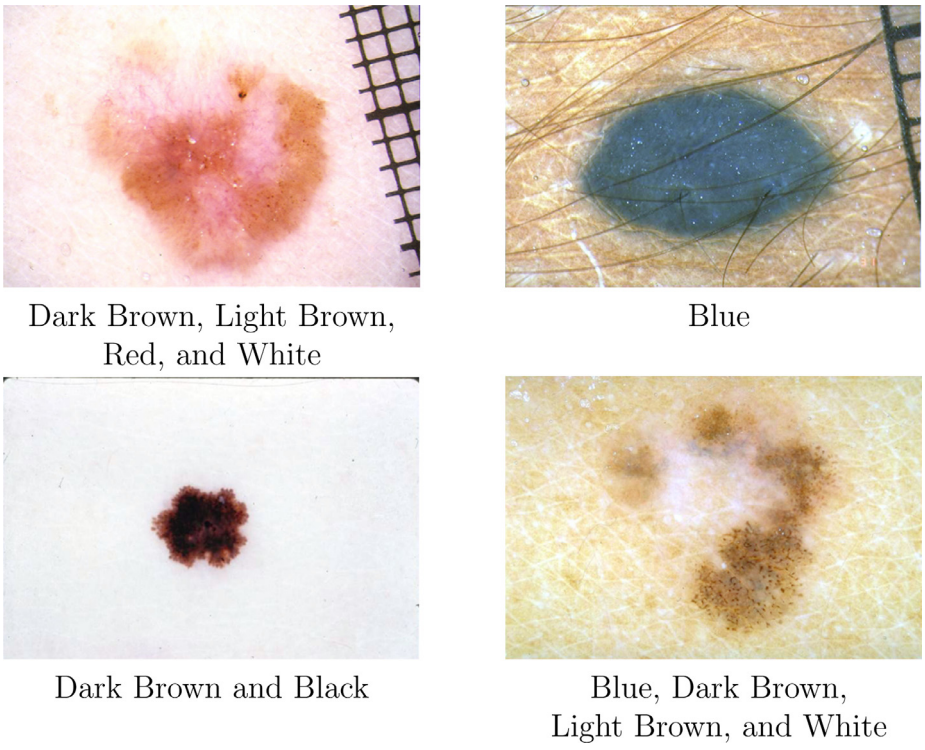


**Fig. 2.** Examples of the color identification/annotation performed by dermatologists [1]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

in Fig. 3. Each of the patches is characterized by a feature vector $r_n^d,\ n = 1, \ldots, N^d$, where $N^d$ is the number of patches of image $d$. A group of dermatologists provided a set of global text labels for each of the images $w_m^d,\ m = 1, \ldots, M^d$, stating which are the $M^d$ colors of the lesion $d$. The used notation is exemplified in Fig. 3 and see Fig. 2 for more examples of the color labels provided by experts.

Our work has two main goals. The first one is to find a correspondence between the patches and the medical text labels using a probabilistic model, so that we can automatically reproduce the color annotations for new images. Global colors labels allow an overall description of the lesion, which by itself could provide the dermatologists with sufficient medical information for a diagnosis. However, for visualization and medical validation purposes we are also strongly interested in being able to associate the global labels with specific regions of the lesions. Thus, the selected probabilistic model must represent the data in such a way that it can associate the color labels with specific image patches. In other words, the model must allow the computation of the following probabilities: i) the distribution of a label given a single patch $p(w_m|r_n)$, which is used to fulfill the task of patch labeling; and ii) the distribution of a label given the entire image/lesion $p(w_m|\mathbf{r})$ that can be used to estimate the global color labels.

After performing the labeling process, our second goal is to use the obtained color annotations to diagnose the lesion as melanoma or benign. The best strategy for using these annotations to diagnose the lesions is unknown. Thus it is necessary to define an approach that converts the color annotations into descriptors suitable to be used by a classification algorithm.

### 2.3. Proposed framework

Fig. 3 shows the training process of the proposed CAD system. The training phase can be divided into three steps, as shown in the figure. First, we estimate the parameters of a probabilistic model that relates the patch features $r_n^d$ with the text labels $w_m^d$ using the $D$ images of the training set. Then, we apply the estimated model to the training images and label their patches according to the most probable color (see Fig. 3). The global color labels are also obtained for each image during this stage. The probabilistic model and its application to dermoscopy images are discussed in Sections 4 and 5. Finally, we extract discriminative features from this color representation and use them to train a classification algorithm to distinguish between benign and lesions. The investigated features and classification approaches are discussed in Section 6.

Fig. 4 shows an example of the application of the CAD system to a new image. This is an example of the performance of the system in the real world, where the only information that we have access to are the image, its patches and collection of corresponding features $\mathbf{r}$. The analysis of a new image is performed in a sequential way. First, we apply the previously estimated probabilistic model to obtain the patch and global image labels $\mathbf{w}$, as exemplified in the figure. Finally, features are extracted and the previously trained classification algorithm is used to obtain a diagnosis. In the example, the lesion is diagnosed as a melanoma.

The main advantage of the proposed system is its ability to interact with the dermatologist, since it is performing a diagnosis relying on a color description that is medically inspired. By trying to identify relevant colors and showing the ones that are detected by the system, we are allowing the dermatologist to understand and validate the suggested lesion diagnosis. Furthermore, its sequential framework is similar to the analysis performed by an expert: first look for dermoscopic criteria (color) and then perform a diagnosis. These two characteristics of the proposed system make it valuable for the medical community and make it significantly different from other systems found in the literature [15].

## 3. Related work

The automatic reproduction of the image labeling process performed by humans is the goal of image annotation algorithms.

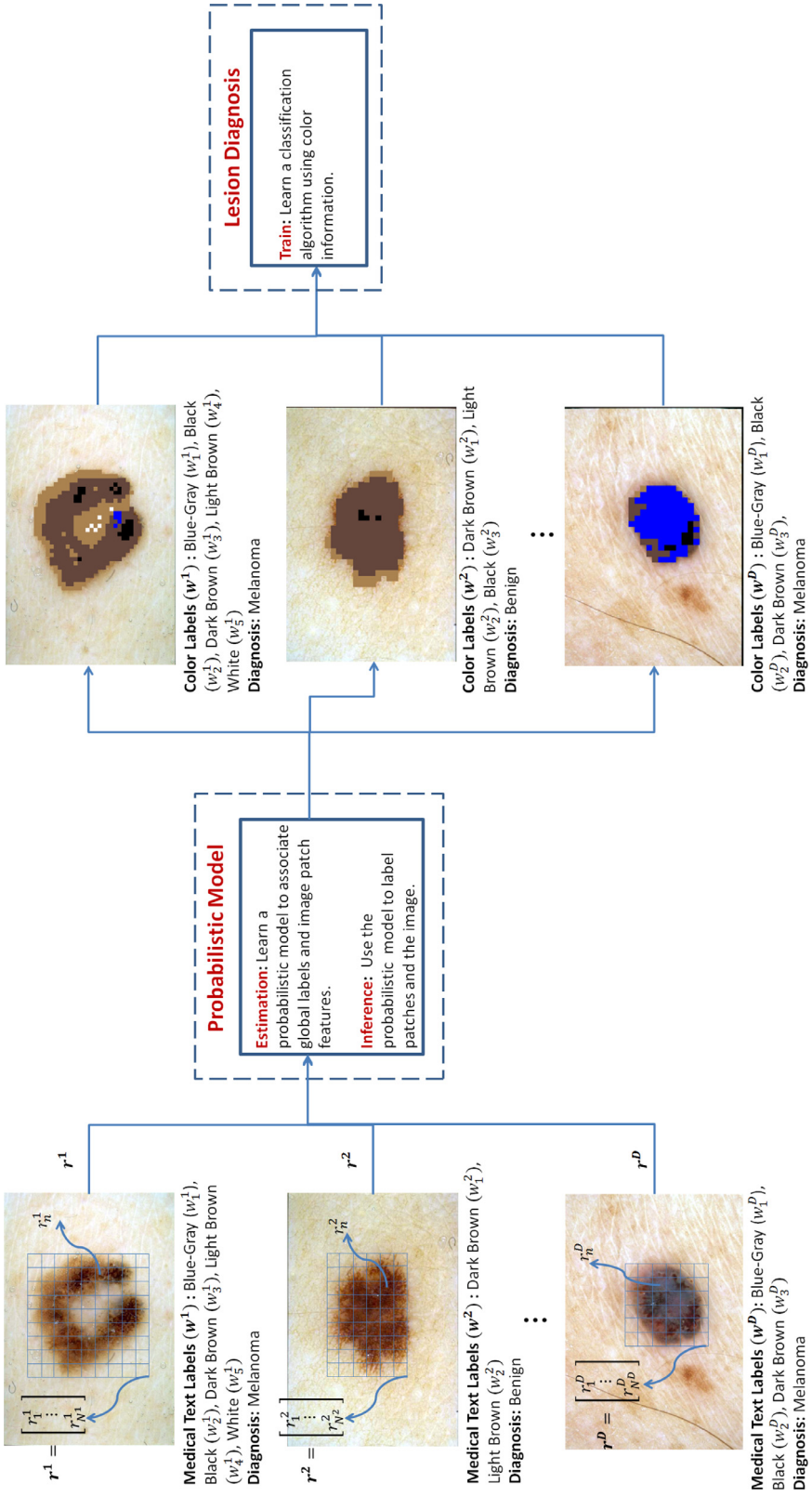


**Fig. 3.** Proposed framework: training step. Here $\mathbf{r}^d$ is the collection of local features $r_n^d$ for the $d$-th image and $\mathbf{w}^d$ represents the corresponding global text labels.

Nowadays, there are significantly large image and video repositories that require image annotation algorithms to speed up the labeling process. These annotations can then be used in tasks such as image retrieval or recognition. One of the major challenges of image annotation algorithms is that they have to be trained using weakly labeled data, i.e., they have image labels but no indication of the image regions that are connected to each of the labels [37]. Recalling the previous section, it is possible to notice that the we have the same difficulty in this paper: we have access to medical text labels but the corresponding segmentations are missing. Thus, it makes sense for

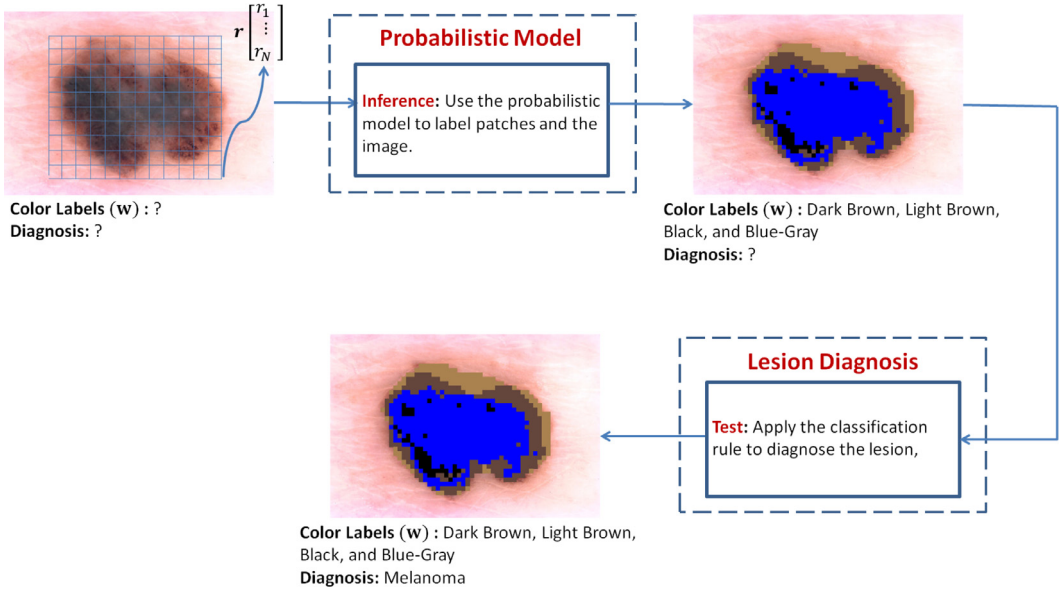


**Fig. 4.** Proposed framework: test step. Here **r** is the collection of local features $r_n$ and **w** represents the corresponding global text labels.

us to see our color detection problem as an image annotation one, and apply an annotation algorithm to solve it.

Image annotation algorithms can be divided into two different categories [38]. The first one can be seen as a supervised classification approach, where a classifier is trained separately for each of the possible labels. This leads to the application of several detection problems, whenever a new image is being annotated [37,38]. The supervised approach was used in the earliest works of image annotation since it guaranteed that the obtained labels were optimal to be subsequently used in an image retrieval or recognition method. However, learning a separate classifier for each of the labels might not be practical if we have a significant number of possible labels and/or training images. Furthermore, supervised annotation is less suitable for problems where multiple labels can be assigned to a single image, as is the case of this work.

An alternative is to use unsupervised learning [38]. The general idea behind this type of approach is to introduce hidden variables $z$ that capture the probabilistic relationship between image patches and text labels. Several algorithms have been proposed to perform unsupervised image annotation (e.g., [35,37,39]). Among these algorithms, we would like to select one that is capable of not only providing annotations, but also of associating the obtained labels with different image regions. This is important for medical validation purposes, since it will allow the dermatologists to associate the produced labels with regions of the lesions. An algorithm that fulfills this requirement is a generative probabilistic model called Correspondence Latent Derichlet Allocation (corr-LDA) [35]. This model assumes an image generation process based on hidden variables $z$ called topics, which are also used to model the joint density between the different image regions and the labels. Due to its probabilistic formulation, corr-LDA also allows the computation of the conditional probabilities $p(w_m|r_n)$ and $p(w_m|\mathbf{r})$, defined in the previous section, which can be used to perform image annotation and to obtain the desired medical representation. Furthermore, the output of corr-LDA can be used to characterize the lesions and obtain suitable feature vectors that can be used to classify the lesions as melanoma or benign.

Corr-LDA is an extension of another generative model called LDA [40], which was proposed for document retrieval and later used in image related tasks (e.g., scene recognition [41]). However, without modifications, LDA is unsuitable for image annotation. The purpose of LDA is to obtain a representation of the data based on the hidden variables that can be used for description and data classification. Its original formulation does not allow the computation of the desired conditional probabilities $p(w_m|r_n)$ and $p(w_m|\mathbf{r})$. Variants of corr-LDA can also be found in the literature (e.g., [42–44]). These new versions include additional information to improve the annotation process, namely new observed variables are considered in the joint probabilities. An example is the inclusion of a variable that identifies the class of the image, e.g., the type of scenario: landscape, seashore, mountain, etc [42]. In several annotation tasks it makes sense to define a relationship between the class of the image and the obtained labels. In our system we prefer not to enforce a relationship between colors and the class of the lesion (melanoma or benign), since the colors are related to different aspects, such as the skin layer where the lesion originated [1]. Furthermore, this allows us to separately address the two problems of this paper: i) obtain a medical representation of the lesion and ii) lesion classification.

The new lesion characterization can then be used to train a classifier in order to obtain the decision rule: melanoma or benign. Color detection in dermoscopy images has been addressed before. However, to the best of our knowledge this is the first work where an image annotation framework and corr-LDA are applied to this problem. The same can be said for the identification of any other dermoscopic criterion.

## 4. Correspondence Latent Dirichlet Allocation (corr-LDA)

corr-LDA is a generative model that first creates the patch features and then generates the annotation words conditioned on the image patches [35], as shown in Fig. 5(a). This figure depicts a simplified version of the generative process, which can be summarized as follows. First, $N^d$ feature vectors $r_n^d$ are generated to characterize each of the image patches. This allows the creation of an image described by $\mathbf{r}^d = \{r_1^d, \ldots, r_N^d\}$. Each of the descriptors is generated conditioned on a hidden variable (topic) $z_n^d$, where $\mathbf{z}^d = \{z_1^d, \ldots, z_N^d\}$ is the set of topics that was used to obtain the image $d$. Finally, the global image annotations are obtained as follows. For each of the $M^d$ annotations, one of the image patches is selected and a corresponding annotation $w_m^d$ is drawn conditioned on the topic that was used to generate the patch descriptor. The selection of the image patch is performed using a latent indexing variable $y_m^d$ that takes values between 1 and $N^d$.

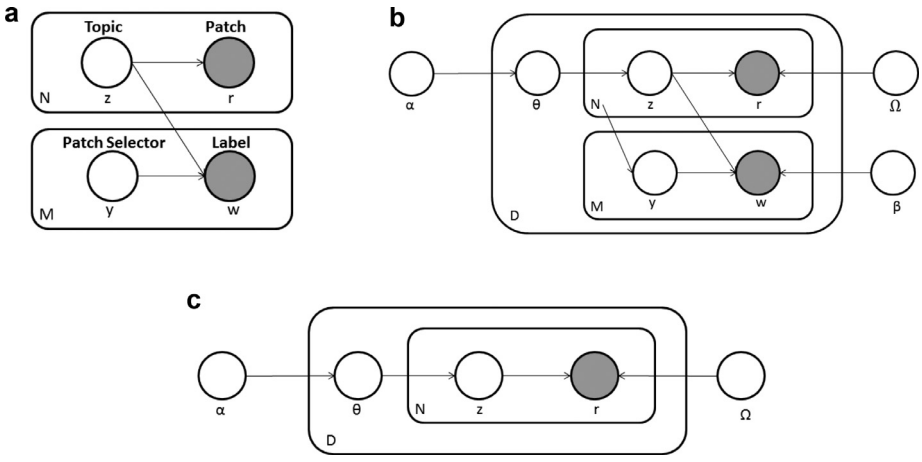


**Fig. 5.** Graphical representations of: (a) simplified Corr-LDA, (b) complete Corr-LDA, and (c) LDA. Each of the boxes represents an image, a patch or a label replication. The filled circles represent the variables observed in the training set.

Each of the previous variables is generated using a parametric distribution. The full generative process and the parameters involved are shown in Fig. 5(b). For a set of $D$ images the generative process can be summarized as follows [35]:

1. For each image $d$, from a set of $D$ images, sample a topic distribution $\theta \sim \text{Dirichlet}(\alpha)$.
2. For each of the $N$ image patches described by $r_n$
   (a) Sample a topic $z_n \sim \text{Multinomial}(\theta)$.
   (b) Sample a patch descriptor $r_n \sim p(r|z_n, \Omega)$ from a multivariate Gaussian distribution conditioned on $z_n$.
3. For each of the $M$ labels $w_m$
   (a) Sample an indexing variable $y_m \sim \text{Uniform}(1, \ldots, N)$.
   (b) Sample an annotation $w_m \sim p(w|y_m, \mathbf{z}, \beta)$ from a multinomial distribution conditioned on the $z_{y_m}$ topic.

Here, $\alpha$ is the Dirichlet parameter and equals the number of topics ($K$). $\Omega$ is the set of parameters of one of the $k = 1, \ldots, K$ multivariate Gaussian distributions that characterize the image patches, and $\beta$ is the distribution of the possible labels over each of the $k$ topics. These are model parameters, while $\theta$ is an image specific parameter that equals $K$ and is sampled once per image.

Fig. 5(c) shows the graphical representation of the traditional LDA model. A comparison of this model with the one of Corr-LDA (Fig. 5(b)) shows that the latter applies LDA to obtain the image patches. The main difference between the two methods is that Corr-LDA also includes a block that generates the annotations conditioned on the selected topics.

### 4.1. Inference

In order to use corr-LDA it is necessary to compute the posterior distribution of the latent variables $(\theta, \mathbf{z}, \mathbf{y})$ given the observations (patch features and annotations)

$$p(\theta, \mathbf{z}, \mathbf{y}|\mathbf{w}, \mathbf{r}, \alpha, \beta, \Omega) = \frac{p(\mathbf{r}, \mathbf{w}, \theta, \mathbf{z}, \mathbf{y}|\alpha, \beta, \Omega)}{p(\mathbf{r}, \mathbf{w}|\alpha, \beta, \Omega)}. \quad (1)$$

The joint distribution of image patches, annotations, and latent variables is obtained as follows

$$p(\mathbf{r}, \mathbf{w}, \theta, \mathbf{z}, \mathbf{y}|\alpha, \beta, \Omega) = p(\theta|\alpha)\left(\prod_{n=1}^{N} p(z_n|\theta)\, p(r_n|z_n, \Omega)\right) \cdot \left(\prod_{m=1}^{M} p(y_m|N)\, p(w_m|y_m, \mathbf{z}, \beta)\right), \quad (2)$$

where independence is assumed among the several image patches and the different annotations. The distribution $p(\mathbf{r}, \mathbf{w}|\alpha, \beta, \Omega)$ is obtained through the marginalization of $p(\mathbf{r}, \mathbf{w}, \theta, \mathbf{z}, \mathbf{y}|\alpha, \beta, \Omega)$ over the latent variables. This distribution is intractable, which means that it is not possible to obtain an exact computation of the posterior distribution of the latent variables. Fortunately, this issue can be addressed using variational inference to approximate the posterior.

Variational inference consists of applying Jensen's Inequality to obtain a family of lower bounds of the log-likelihood. A simple way to obtain the family of lower bounds is to define a factorized distribution on the latent variables [35]

$$q(\theta, \mathbf{z}, \mathbf{y}) = q(\theta|\gamma)\left(\prod_{n=1}^{N} q(z_n|\phi_n)\right) \cdot \left(\prod_{m=1}^{M} q(y_m|\lambda_m)\right), \quad (3)$$

using the variational parameters $(\gamma, \phi, \lambda)$. Each of these parameters is related to its respective latent variable, thus $\gamma$ is a $K$-dimensional Dirichlet parameter, $\phi_n$ are $N$ $K$-dimensional multinomial parameters and $\lambda_m$ are $M$ $N$-dimensional multinomial parameters.

The optimal values of the variational parameters are found by minimizing the Kullback–Leibler (KL) divergence between the defined factorized distribution and the true posterior. This enforces a dependence on the data $(\mathbf{r}, \mathbf{w})$. Minimizing the KL divergence is equivalent to maximizing the lower bound obtained using Jensen's Inequality as follows (refer to [40] for more details):

$$\begin{aligned} \log p(\mathbf{r}, \mathbf{w}|\alpha, \beta, \Omega) &= \log \int_\theta \sum_{\mathbf{z}} \sum_{\mathbf{y}} p(\mathbf{r}, \mathbf{w}, \theta, \mathbf{z}, \mathbf{y}|\alpha, \beta, \Omega) \mathrm{d}\theta \\ &= \log \int_\theta \sum_{\mathbf{z}} \sum_{\mathbf{y}} \frac{p(\mathbf{r}, \mathbf{w}, \theta, \mathbf{z}, \mathbf{y}|\alpha, \beta, \Omega) q(\theta, \mathbf{z}, \mathbf{y})}{q(\theta, \mathbf{z}, \mathbf{y})} \mathrm{d}\theta \\ &\geq \mathbb{E}_q[\log p(\mathbf{r}, \mathbf{w}, \theta, \mathbf{z}, \mathbf{y}|\alpha, \beta, \Omega)] - \mathbb{E}_q[\log q(\theta, \mathbf{z}, \mathbf{y})]. \end{aligned} \quad (4)$$

The right side of (4) is the lower bound of the log-likelihood: $\mathcal{L}(\gamma, \phi, \lambda; \alpha, \beta, \Omega)$. The distributions $p(\mathbf{r}, \mathbf{w}, \theta, \mathbf{z}, \mathbf{y}|\alpha, \beta, \Omega)$ and $q(\theta, \mathbf{z}, \mathbf{y})$ can be factorized, leading to the following factorization of the lower bound

$$\begin{aligned} \mathcal{L}(\gamma, \phi, \lambda; \alpha, \beta, \Omega) = {} & \mathbb{E}_q[\log p(\theta|\alpha)] + \mathbb{E}_q[\log p(\mathbf{z}|\theta)] \\ & + \mathbb{E}_q[\log p(\mathbf{r}|\mathbf{z}, \Omega)] + \mathbb{E}_q[\log p(\mathbf{y}|N)] \\ & + \mathbb{E}_q[\log p(\mathbf{w}|\mathbf{y}, \mathbf{z}, \beta)] - \mathbb{E}_q[\log q(\theta|\gamma)] \\ & - \mathbb{E}_q[\log q(\mathbf{z}|\phi)] - \mathbb{E}_q[\log q(\mathbf{y}|\lambda)]. \end{aligned} \quad (5)$$

Each of the terms in (5) can be expanded into explicit functions of the model $(\alpha, \beta, \Omega)$ and variational $(\gamma, \phi, \lambda)$ parameters. For completeness, the expanded version of each of the terms of $\mathcal{L}(\gamma, \phi, \lambda; \alpha, \beta, \Omega)$ is included in Appendix A.

The variational parameters can be obtained by taking the derivatives of $\mathcal{L}(\gamma, \phi, \lambda; \alpha, \beta, \Omega)$ with respect to each of them and setting these derivatives to zero. This leads to an iterative process that is repeated until the change in the KL divergence is negligible [35]. In Section 5 we present the update equations for each of the variational parameters.

After obtaining an approximation for the posterior it is now possible to compute the conditional distributions of interest $p(w|r_n)$ and $p(w|\mathbf{r})$. The first probability can be used to perform the patch labeling [35]

$$p(w|r_n) \propto \sum_{z_n} q(z_n|\phi_n)\, p(w|z_n, \beta)\,, \tag{6}$$

while the second probability can be used to obtain the global labels [35]

$$p(w|\mathbf{r}) \propto \sum_{n=1}^{N} \sum_{z_n} q(z_n|\phi_n)\, p(w|z_n, \beta)\,. \tag{7}$$

### 4.2. Parameter estimation

Given a set of pairs images features/annotations $(\mathbf{r}^d, \mathbf{w}^d)$, $d = 1, \ldots, D$, our goal is to obtain the maximum likelihood estimates of the model parameters $(\alpha, \beta, \Omega)$. These estimates can be obtained using a variational Expectation-Maximization (EM) method that maximizes the aforementioned lower bound $\mathcal{L}(\gamma, \phi, \lambda; \alpha, \beta, \Omega)$. More specifically, this process consists of iteratively applying the following two steps until convergence

- *E-Step:* The variational parameters $(\gamma^d, \phi^d, \lambda^d)$ are estimated for each image in the dataset and the lower bound $\mathcal{L}(\gamma, \phi, \lambda; \alpha, \beta, \Omega)$ is computed, as described in Section 4.1.
- *M-Step:* The model parameters $\alpha$, $\beta$, and $\Omega$ are estimated by maximizing the lower bound $\mathcal{L}(\gamma, \phi, \lambda; \alpha, \beta, \Omega)$ obtained in the E-step.

The update equations of the model parameters are obtained by taking derivatives of $\mathcal{L}(\gamma, \phi, \lambda; \alpha, \beta, \Omega)$ with respect to each of them and then setting these derivatives to zero. In Section 5 we will show the update equations in detail.

## 5. Color detection using corr-LDA

This section describes the application of corr-LDA to dermoscopy images as well as some modifications introduced to the original algorithm.

### 5.1. Patch and feature extraction

Before the application of corr-LDA it is necessary to divide the dermoscopy images into several regions. Different approaches can be used to achieve this task. In this work we apply an uniform grid to divide the lesion into small non-overlapping square patches of size $12 \times 12$ pixels. This size was selected based on the average resolution of the dermoscopy images ($570 \times 760$) and on the results obtained in previous works [12,26]. Furthermore, these are the dimensions that allow us to identify small color regions in the lesions, without significantly increasing the computational running times.

We are only interested in patches that contain the lesion, thus patches containing less than 50% lesion pixels are discarded. The identification of the lesion's pixels is performed using a manual segmentation, which allow us to separate the lesion from the healthy skin. Ideally a CAD system should be fully automatic, meaning that it should not require any interaction, such as manual segmentation, from the user. Despite the large number of works on this topic, automatic border detection algorithms are far from perfection [45–47]. Incorrect segmentations can negatively influence the color detection process and hamper a proper estimation of the color detection model and lesion classification algorithms [48,49]. Thus, it is preferable to use manual segmentations in the learning stage and only incorporate an automatic segmentation method in a final version of the system, where its influence is mitigated, as has been shown in [49].

Each of the patches is characterized using the mean color vector in the HSV color space. Although RGB is the original color space of dermoscopy images, we prefer to use HSV because this color space performs a description of color similar to the human visual system. Furthermore, this color space has been shown to perform well in different dermoscopy image problems [12,26]. The uniform color space CIE L*a*b* was tested as well. However, it performed poorly for the dark brown, black, and white colors. This same limitation of L*a*b* had been noted before, in a previous color detection work [26]. The set of mean color vectors that characterize the patches corresponds to the set $\mathbf{r} = \{r_1, \ldots, r_N\}$ defined in the previous section.

### 5.2. Medical color annotations

The annotations provided by the dermatologists are strings. In order to simplify this information and make it usable for a computer, we have changed the format of the annotations. The annotation vector $\mathbf{w}$ is assumed to be a binary vector of length $M = 6$ (same as the number of colors [5]) where $w_m = 1$ if the $m$-th color is present and 0 otherwise.

### 5.3. Training and testing of corr-LDA

The application of corr-LDA to dermoscopy images can be divided into two phases: training and testing.

*Training:* We use a set of $D$ annotated images to estimate the model parameters $(\alpha, \beta, \Omega)$ as described in Section 4.2.

*Testing:* The annotation of new images is performed as follows. First we apply the E-step to each of the images in order to determine their corresponding variational parameters. Then we use (6) to label each image patch according to each of the six possible colors (blue-gray, dark brown, light brown, white, red, or black).

Our strategy to obtain the global labels is different from the one describe in [35]. Corr-LDA and other annotation algorithms are usually trained using a large dictionary of possible text labels (these dictionaries can contain more than 10k words). However, it is assumed that each training image is only associated with a very small set of all the possible text labels. During the annotation of a new image (test phase), the possible labels of the dictionary are sorted according to their conditional probabilities $p(w|\mathbf{r})$ (computed using (7)). This means that labels with higher probabilities will come first and the labels with lower probabilities will be the last ones. Then, a fixed number of these sorted annotations is selected per image (e.g., in [35] they select 4 words per image) and are set as the global labels. This restricts the number of possible annotations that can be associated with an image. Finally, these global labels are compared with the ones provided by human users and the corresponding performance metrics are computed.

The problem addressed in this work is slightly different because we have a dictionary of only six words and there is no restriction regarding the number of colors that can be found in the lesions. We can find just one color but we can also find all of the six colors in a single lesion. Thus, the approach described in the previous paragraph could not be applied to our problem and it was necessary to adopt an alternative strategy based on the number of patches per color that could be found in the lesion. In our simplified method, an image is labeled with a certain color if the same label was assigned to at least three patches of image during the patch labeling step.

### 5.4. Inclusion of the von-Mises distribution

The HSV color space represents the colors in terms of their (H)ue, (S)atu-ration, and (V)alue. This is a mixture of angular (H) and linear (S and V) information. The original formulation of corr-LDA considers that the patch features are modeled using multivariate Gaussian distributions. This representation is not appropriate in the case of the H channel, since this is a periodic angular measure. Therefore, it is necessary to find an alternative distribution that is more suitable for our kind of data. In our approach, we use a von-Mises distribution to model the content of the H channel, while S and V channels are modeled using a multivariate Gaussian, as before. We have selected a von-Mises distribution to represent the H channel because this is a periodic distribution that has been used before to represent this angular information [50].

Assuming independence between H and the other channels, it is possible to obtain the following distribution

$$p(r_n|z_n,\Omega) = \nu(\mathrm{H}_n|z_n,\tau,\varepsilon)\ .\ G(\mathrm{S}_n,\mathrm{V}_n|z_n,\mu,\Sigma), \tag{8}$$

where $G$ is the 2-dimensional Gaussian and $\nu$ is a von-Mises distribution

$$\nu(\mathrm{H}_n|z_n,\tau,\varepsilon) = \frac{1}{2\pi I_0(\varepsilon)} e^{\varepsilon\cos(\mathrm{H}_n-\tau)}, \tag{9}$$

where the normalization factor $I_0$ is the modified zero-order Bessel function of the first kind and $\varepsilon \geq 0$ denotes the concentration of the distribution around the mean $\tau$.

The new distribution of $p(r_n|z_n,\Omega)$, $\Omega = (\mu,\Sigma,\tau,\varepsilon)$ is used to define the update equations of the variational and model parameters. The update equations of the parameters were obtained by taking derivatives of $\mathcal{L}(\gamma,\phi,\lambda;\alpha,\beta,\Omega)$ (see Appendix A) and setting them to zero (refer to [35,40] for details on the derivatives). Below are the equations obtained. The equations are divided according to their corresponding step of the variational EM-algorithm and are sorted by order of computation.

- *E-step* (performed for any lesion $d$ during train or test)

$$\phi^d_{nk} \propto p(r^d_n|z_n=k,\tau,\varepsilon,\theta,\Sigma)\exp\{E_q[\log q(\theta_k|\gamma^d)]\}.$$
$$.\exp\left\{\sum_{m=1}^{M=6}\lambda^d_{mn}\log p(w^d_m|y_m=n,z_m=i,\beta)\right\}, \tag{10}$$

$$\lambda^d_{mn} \propto \exp\left\{\sum_{k=1}^{K}\phi^d_{nk}\log p(w^d_m|y_m=n,z_m=i,\beta)\right\}, \tag{11}$$

$$\gamma^d_k = \alpha_k + \sum_{n=1}^{N_d}\phi^d_{nk}, \tag{12}$$

  It is necessary to perform an initialization of the variational parameters. This is performed as proposed in [35,40].
- *M-step* (performed using the training set of size $D$)

$$\mu_k = \frac{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi^d_{nk}[\mathrm{S,V}]^d_n}{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi^d_{nk}}, \tag{13}$$

$$\tau_k = \tan^{-1}\left(\frac{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi^d_{nk}\sin\mathrm{H}^d_n}{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi^d_{nk}\cos\mathrm{H}^d_n}\right), \tag{14}$$

$$\Sigma_k = \frac{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi^d_{nk}([\mathrm{S,V}]^d_n-\mu_k)([\mathrm{S,V}]^d_n-\mu_k)^T}{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi^d_{nk}}. \tag{15}$$

  An analytical computation of the parameter $\varepsilon_k$ is not possible. Different approximations have been proposed to tackle this issue. In this work we use the approach described in [51], which makes use of Newton–Raphson's method to obtain an approximation. This method requires a few iterations $t$ of the following equation:

$$\varepsilon^t_k = \varepsilon^{t-1}_k - \frac{A(\varepsilon^{t-1}_k)-\bar{R}}{1-A(\varepsilon^{t-1}_k)}, \tag{16}$$

  where

$$A(\varepsilon^{t-1}_k) = \frac{I_1(\varepsilon^{t-1}_k)}{I_0(\varepsilon^{t-1}_k)}, \tag{17}$$

  and the variable $\bar{R}$ is computed as follows

$$\bar{R} = \frac{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi^d_{nk}\cos([\mathrm{H}]^d_n-\tau_k)}{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi^d_{nk}}. \tag{18}$$

  In the first iteration $\varepsilon^0_k$ is set as follows [51]

$$\varepsilon^0_k = \frac{\bar{R}-\bar{R}^3}{1-\bar{R}^2}. \tag{19}$$

  We perform the update equation until convergence is reached.

  The next parameter to be estimated is $\beta$ that relates the color labels with the topics $k$

$$\beta_{km} \propto \sum_{d=1}^{D} w^d_m \sum_{n=1}^{N_d}\phi^d_{nk}\lambda^d_{mn}. \tag{20}$$

  The Dirichlet parameter $\alpha$ is updated. An exact computation of this parameter not possible, thus Blei and Jordan make use of Newton–Raphson's method [40] to obtain an update equation. Please refer to [40] for details.

Fig. 6 shows two examples of the performance of the proposed von-Mises Gaussian formulation against the traditional formulation that uses a multivariate Gaussian. These examples clearly demonstrate that the proposed formulation outperforms the traditional one, rendering better results and more consistent color regions.

## 6. Lesion diagnosis

corr-LDA allows us to obtain local (patch) and global (image) color labels for the lesions. Our second goal is to use this medical color information to diagnose the lesions as melanoma or benign. To achieve this goal it is necessary to convert the color annotations to an appropriate description that can be used by machine learning algorithms. Since we do not know the optimal way to describe the lesions, we investigate four different strategies:

- *Number of Colors (i):* This is the simplest and most clinically oriented description. We simply count the number of global labels (colors) that are obtained for a given lesion and use this number to characterize the lesion.
- *Present/Absent Colors (ii):* Instead of counting the number of colors, we can describe the lesion stating which are the colors that are present or absent. We represent the lesion by a feature vector $\mathbf{c}^d$ of length 6, where $c^d_m$ is equal to 1 if the $m$th color is present and 0 otherwise. The reader might identify this description as the same one that we use to represent the medical color annotations during the train of corr-LDA.
- *Distribution of Color Annotations (iii):* Another possibility is to describe the images using the conditional distribution $p(w|\mathbf{r})$, which provides an approximation of the distribution of each color in a given lesion. We represent each lesion by a feature vector $\mathbf{c}^d$ of size 6, where $c^d_m = p(w_m|\mathbf{r}^d)$ and $m$ identifies one of the six colors.
- *Number of Patches per Topics (iv):* Recalling (12) it is possible to see that each of the variational parameters $\gamma^d_k$ corresponds to approximately to the $k$th model parameter $\alpha_k$ plus the expected number of patch features that were generated by the $k$th topic. This means

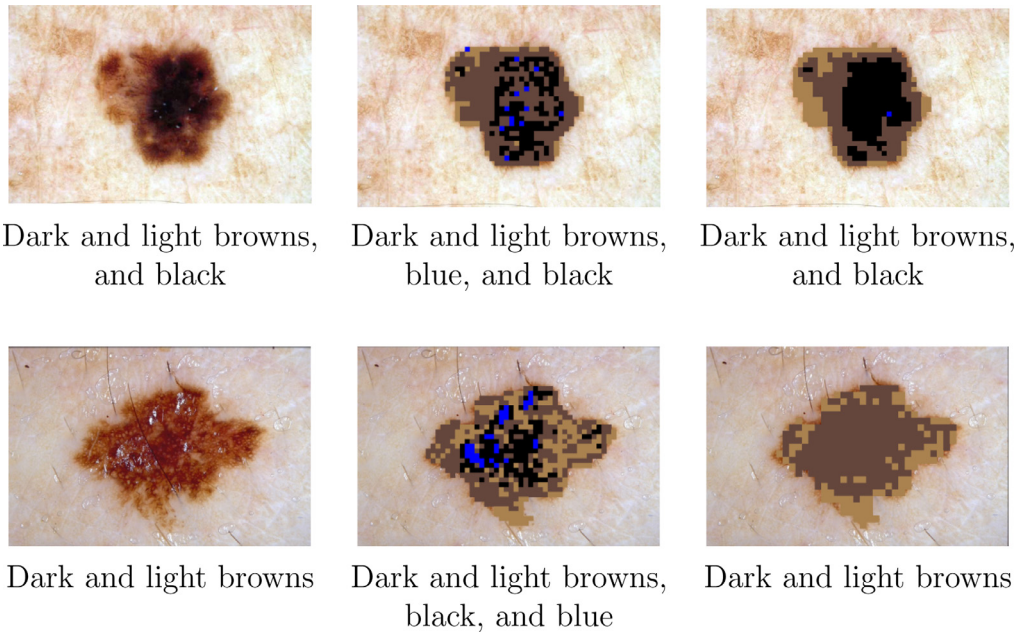


**Fig. 6.** Original image and medical labels (left), output of corr-LDA using a multivariate Gaussian (mid), and output of corr-LDA using a von Mises-Gaussian (right).

that if we subtract the model parameter $\alpha$ from a variational parameter $\gamma^d$ of a given image, we will obtain the expected number of patches that was generated by each of the topics. The number of patches per topic can also be used to characterize the lesions, as proposed in [40]. The feature vector $\mathbf{c}^d$ obtained in this case has the same length as the number of topics and $c_k^d = \alpha_k - \gamma_k^d$.

Each of the aforementioned descriptors is used to classify the lesions as melanoma or benign. The classification method based on feature (i) is the simplest one. We classify the lesion as melanoma if the number of annotations/colors is higher than 3. This threshold is defined based on the findings of MacKie et al. [52] and has been previously used in the color-based system proposed by Seidenari et al. [24]. The diagnosis based on the remaining descriptors requires the use of a classification algorithm. This means that we have to train a classifier using a training set of images previously diagnosed by an expert. Then, the obtained classification rule is used to classify new lesions as melanoma or benign. Four classification algorithms are considered in this work: AdaBoost, Support Vector Machines (SVM), k-Nearest Neighbor (kNN), and Random Forests.

## 7. Experimental setup

This section describes the experimental setup used to train and evaluate the different parts of the CAD system, namely color detection using corr-LDA and lesion diagnosis (recall Fig. 3).

### 7.1. Dataset and performance metrics

The experiments were performed using a dataset of 482 dermoscopy images (50% melanomas) randomly selected from the commercial database EDRA [1]. This is a multi-source database that contains dermoscopy images from three different universities hospitals: University Federico II of Naples (Italy), University of Graz (Austria), and University of Florence (Italy). Each of the lesions has been analyzed by a group of experienced dermatologists, who provided several annotations regarding the presence or absence of clinically relevant structures as well as a diagnosis. The same dataset of 482 images has been used in previous works of lesion diagnosis, namely [53], which allows us to have a fair comparison between approaches.

Color labels were provided according to the ABCD rule [1], which means that we had information about the presence/absence of six clinically relevant colors (dark and light browns, blue-gray, black, red, and white). The ABCD rule can only be applied when the lesion is fully contained within the image. Unfortunately not all the images in our dataset complied with this constraint, thus we had color annotations for 344 out of the 482 images. In order to tackle this issue, we used the reduced set to train and evaluate the color detection method based on corr-LDA while the full set was used to train and test the lesion diagnosis block.

All of the images were pre-processed in order to remove acquisition artifacts and skin hair as described in [19] and their colors were normalized as proposed in [53]. Color normalization has been shown to improve the task of color detection in previous works [54]. In order to separate the lesions from healthy skin we have performed manual segmentations.

To evaluate the performance of the color detection strategy we computed two metrics for each color: Precision and Recall, defined as follows

$$\text{Precision} = \frac{\#I_{rel} \cap \#I_{ret}}{\#I_{ret}}, \tag{21}$$

$$\text{Recall} = \frac{\#I_{rel} \cap \#I_{ret}}{\#I_{rel}}. \tag{22}$$

where $\#I_{rel}$ is the number of images that was annotated with a certain color by the experts and $\#I_{ret}$ is the number of images that was annotated with the same color by the probabilistic model.

The performance of lesion diagnosis is evaluated using the metrics Sensitivity and Specificity. Sensitivity is the percentage of correctly diagnosed melanomas and Specificity stands for the percentage of correctly classified benign lesions. The aforementioned metrics were computed using a 10-fold cross validation approach in which the dataset is divided into ten subsets, each with approximately the same number of melanomas and benign lesions. Nine folds were used for training the classifier and the remaining one was used for testing. The results correspond to the average performance on the ten test folds. We used the same folds to train and test the color detection and lesion diagnosis blocks. Therefore, ensured that the reduced set of 344 images was fairly split among the 10 folds, such that we had enough images for train and test each time.

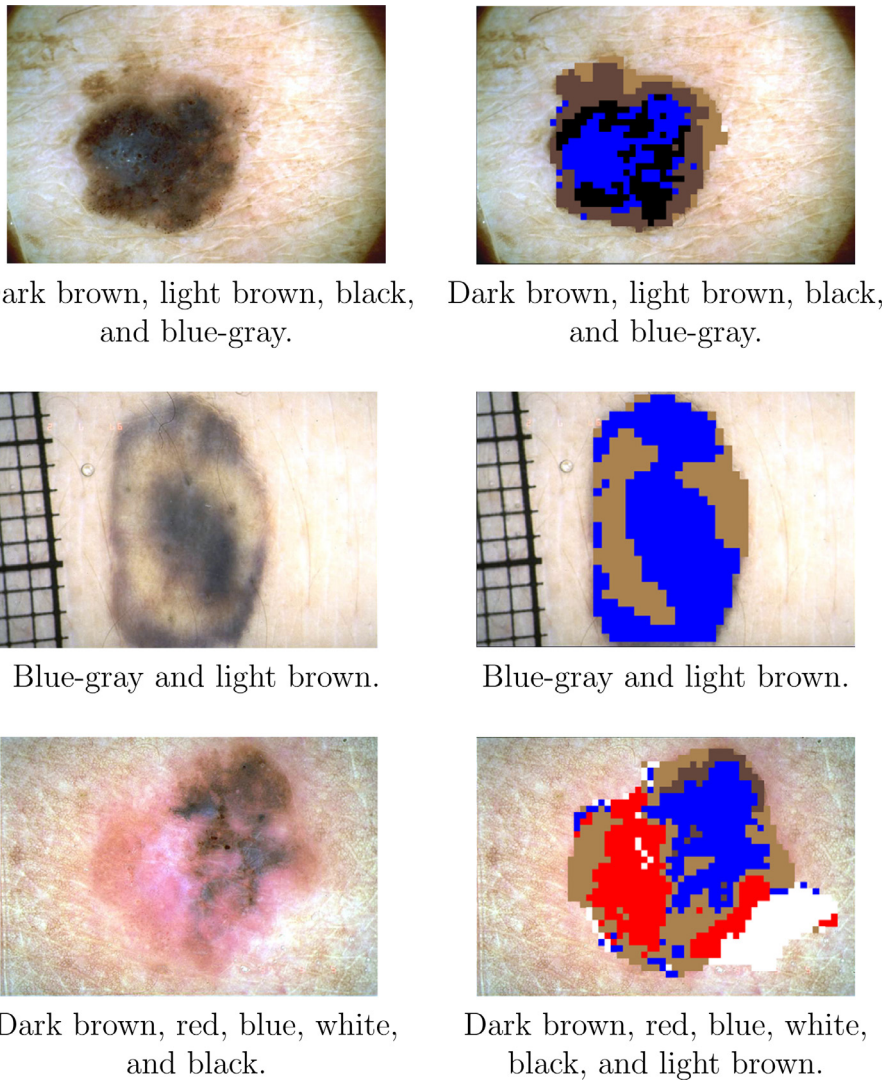


**Fig. 7.** Original image medical labels (left) and output of corr-LDA (right) .

### 7.2. Training of the color detection block

The training of corr-LDA corresponds to finding the optimal set of model parameters $(\alpha, \beta, \Omega)$ using annotated images. This process is described in Section 5. The only variable that can be tuned by the user is the number of topics $K$ considered in the generative process. In this work we have searched for the best number of topics in the set $K = \{50, 75, \ldots, 300\}$.

### 7.3. Training of the lesion classification block

In Section 6 we described different features that can be extracted from the medical color representation obtained with corr-LDA. Since we do not know which is the best feature, we evaluated each of them separately. This means that we tested four diagnosis systems, each developed using a different feature.

Features (ii), (iii), and (iv) require the learning of a classification rule. We investigated three classifiers in our work: AdaBoost, SVM, and kNN. Each of these classifiers requires the tuning of at least one parameter. In the case of AdaBoost we have set the number of weak classifiers $W \in \{1, 2, \ldots, 150\}$, for kNN we set the number of neighbors $p \in \{1, 3, \ldots, 25\}$, and in the case of Random Forests we defined the number of trees $T \in \{1, 2, \ldots, 50\}$. SVM requires the tuning of a larger number of parameters. We studied two kernels in this work: Radial Basis Function (RBF) and polynomial. According to this choice of kernels, we tuned the following parameters: the width of the RGB kernel $\rho \in \{2^{-6}, 2^{-5}, \ldots, 2^{6}\}$, the degree of the polynomial kernel $d \in \{1, 2, \ldots, 5\}$ and the cost $C \in \{2^{-5}, 2^{-4}, \ldots, 2^{5}\}$ given to the soft margin (tuned for both kernels).

We have also combined the four types of features into a single feature vector in order to determine if the results could be improved by combining the different features. A feature selection method was applied to check if there was a subset of the different features that was more informative than using the entire feature vector. We tuned the number of selected features to range from 1 to the length of the original feature vector. The used feature selection strategy was Mutual Information with the max-dependency criterion [55].

## 8. Results and discussion

### 8.1. Color detection

Fig. 7 shows some examples of the output of the color detection block as well as the ground truth labels provided by the experts. The performance of color detection is shown in Table 1. This table shows the scores obtained for each color as well as the average performance of the probabilistic model. The best results were obtained with a configuration of $K = 150$ topics.

**Table 1**
Color detection results.

| | #Images | Precision | Recall |
|---|---|---|---|
| **Blue-gray** | 226 | 87.6% | 94.2% |
| **Dark-brown** | 303 | 95.7% | 95.7% |
| **Light-brown** | 247 | 89.1% | 92.7% |
| **Black** | 179 | 81.5% | 88.8% |
| **Red** | 31 | 79.3% | 74.2% |
| **White** | 15 | 63.6% | 93.3% |
| **Average** | – | 83.8% | 89.8% |

**Table 2**
Comparison of color detection methods. In **bold** we highlight the best results.

| | Proposed method | | Gaussian mixtures [26,54] | |
|---|---|---|---|---|
| | Precision | Recall | Precision | Recall |
| **Blue-gray** | **87.6%** | **94.2%** | 86.5% | 92.2% |
| **Dark-brown** | 95.7% | **95.7%** | **98.3%** | 76.4% |
| **Light-brown** | 89.1% | **92.7%** | **97.0%** | 81.0% |
| **Black** | 81.5% | **88.8%** | **90.9%** | 67.0% |
| **Red** | **79.3%** | **74.2%** | - | - |
| **White** | **63.6%** | **93.3%** | 42.1% | 85.7% |

corr-LDA performs well, achieving good average detection scores. It is possible to obtain a good probabilistic model for the Dark and Light-Brown colors, Blue-Gray, and Black. However, the model does not perform so well for the red and white colors. The different performances can be justified by the number of examples that we have for each color. There is a large number of images where the Dark and Light Brown and the Blue-Gray colors are represented, and for each of this colors corr-LDA shows a good performance. In the case of the Black color, the results are still good, but slightly worse than for the three previous colors. The number of images that contains this color is smaller than in the previous cases. Finally, let us inspect the red and white colors. For each of these colors, the number of examples is very small. Red and white are the colors that are more difficult to find in skin lesions, mainly because they are associated with malignant lesions [1] and images from melanomas are more difficult to obtain.

Despite the different performances for each color, it is important to keep in mind that the detection of colors in skin lesions is a challenging task, especially if one is only using text labels and does not have segmentation examples of each of the colors.

Table 2 compares the performance of the method described in this paper with the one proposed in [26]. This method uses a set of Gaussian mixtures to model the colors. The Gaussian mixtures that represent each color were estimated using color segmentations from a set of 27 images, obtained from the publicly available $PH^2$ dataset [56]. We applied each of the mixtures to our images, in order to identify the colors that were present. Although the mixtures were trained using the $PH^2$ dataset, they can still be applied to images from other datasets as shown in [54]. This work reports the importance of color normalization and how it can be used to successfully improve the analysis of dermoscopy images in different tasks for single and multi-source datasets, as is the case of EDRA. Color detection using the method described in [26] was one of the investigated tasks. Here we report the results obtained in [54], since we are using exactly the same subset of EDRA images for color detection. Unfortunately, as reported in [26], the number of color segmentations was small and it was not possible to model the red color due to lack of training examples. Nonetheless, it is still possible to compare the performance of the two methods for the remaining colors.

Both color detection methods perform well for most of the colors. It is easy to notice that the proposed method outperforms [26] for the white color. Moreover, it seems to achieve a better recall score for all of the colors. The method described in [26] achieves a better precision score for three of the colors. Nonetheless, the precision scores obtained with the approach described in this paper are also high and promising, especially if one considers that in this work we are trying to identify colors using only text labels as training data.

We believe that Corr-LDA can also be applied in the detection of other relevant dermoscopic features such as blue-whitish veil and regression areas [1]. This can help us improve the developed clinically oriented CAD system, since color is not the only criteria considered by dermatologists in their diagnosis. Hence one of our future goals is to extend this methodology to the detection of other dermoscopic structures.

**Table 4**
Best configuration – feature and classifier.

| | Classifier/Parameters | | | |
|---|---|---|---|---|
| Feature | AdaBoost | SVM | kNN | Random forests |
| (iv) | W = 145 | RBF kernel, $C = 2^2$, $\rho = 2^2$ | $k = 13$ | $T = 48$ |
| All | W = 60 | RBF kernel, $C = 2^5$, $\rho = 2^5$ | $k = 61$ | $T = 29$ |

### 8.2. Lesion classification

Table 3 shows the classification scores obtained using each of the features described in Section 6 and their combination (labeled as "All", last row). These results show that each feature performs differently and that some of them are more appropriate to identify melanomas than others. The investigated features were computed over the output of the best corr-LDA model (number of topics $K = 150$). Table 4 shows the best configuration: feature and classifier (refer to Section 7 for a definition of each of the classifier's parameters).

The number of colors (feature (i)) was based on the findings of MacKie et al. [52]. They found that the presence of more than 3 colors was a sign of malignancy, obtaining an SE = 92% and an SP = 51% on their experiments. Applying the same strategy to our database lead to different results, with a significantly lower SE (54.8%) an higher SP (83.8%). Similar observations were made by Seidenari et al. [24], during the development of their color-based CAD system. In their work, they also used the number of colors to classify the skin lesions and the scores were an SE = 69.9% and SP = 85.8%. The differences between our scores and the ones reported in the literature can be related to the dataset used or with different performances of the color detection method (in the case of [24]).

**Table 3**
Lesion diagnosis results. In **bold** we highlight the best results.

| | Threshold | | AdaBoost | | SVM | | kNN | | Random forests | |
|---|---|---|---|---|---|---|---|---|---|---|
| Feature | SE | SP | SE | SP | SE | SP | SE | SP | SE | SP |
| (i) | 54.8% | 83.8% | – | – | – | – | – | – | – | – |
| (ii) | – | – | 67.7% | 73.0% | 65.6% | 73.4% | 44.8% | 86.7% | 64.8% | 75.9% |
| (iii) | – | – | 60.2% | 63.4% | 76.3% | 50.1% | 63.5% | 57.6% | 61.5% | 57.2% |
| (iv) | – | – | 70.9% | 70.1% | 77.6% | 57.2% | 55.6% | 79.7% | **76.3%** | **70.5%** |
| All | – | – | **74.3%** | **73.8%** | **75.1%** | **77.5%** | **70.5%** | **78.0%** | **77.6%** | **73.0%** |

**Table 5**
Comparison of the proposed approach with related works.

| Method | SE | SP |
|---|---|---|
| [24] | 69.9% | 85.8% |
| [27] | 61.6% | 75.8% |
| Proposed – Feature (iv) | 76.3% | 70.5% |
| Proposed – All | 74.3% / 75.1% / 77.6% | 73.8% / 77.5% / 73.0% |

Assessing the colors that can be found in the lesion (feature (ii)) leads to better SP scores than SE. Feature (iii), which corresponds to the distribution $p(w|\mathbf{r})$, do not allow a good discrimination between the melanoma and benign classes.

The best classification results are obtained using feature (iv), i.e., the number of regions per topic. Both AdaBoost and Random Forests obtain interesting classification results, more specifically we achieve the best classification results using Random Forests (SE = 76.3% and SP = 70.5%). Despite being the less clinically inspired feature, it still incorporates medical knowledge, since it is possible to associate each of the topics with the colors through the model parameter $\beta$ (recall Section 4).

The last row of Table 3 shows the performance of the classifiers after combining all the features and performing feature selection. The feature vector had a length of 163 features (recall Section 6 for information on the dimensions of each type of feature), from which we selected a subset of the $S$ most informative features. This combination lead to a small improvement in the performance of Random Forests and AdaBoost. This was expected because these two algorithms belong to the family of boosting algorithms, which means that they are already capable of selecting the most relevant features to obtain the best classification results. On the other hand, the performance of SVM and kNN is degraded for high dimensional spaces (curse of dimensionality) and these classifiers are not capable of selecting the best features. Thus, as expected, feature fusion and selection significantly improves the performance of these classifiers.

The achieved results are comparable to those obtained in a previous work using exactly the same dataset [53]: SE = 73.9%, SP = 80.1% and a Bag-of-Features (BoF) framework with HSV color histograms. This shows us that it is possible to develop clinically oriented approaches, where medical representations of the lesions are used, and still obtain performances similar to those of traditional pattern recognition strategies.

A direct comparison with other related works is not possible due to the different datasets used. Nonetheless, we can still check if our results lead to similar conclusions. Table 5 shows the comparison between our method and two related works that can be found in the literature. Seidenari et al. [24] report scores of SE = 69.9% and SP = 85.8% on calibrated image data, while Celebi and Zornberg [27] report SE = 61.6% SP = 75.8% on uncalibrated image data. The SE scores obtained in our work are higher than the ones reported in these works. On the other hand, our SP scores are lower than the one reported in [24] and similar to the one reported in [27]. Overall, our method obtains a better trade-off between SE and SP.

Although the achieved results are promising, the developed system cannot be used in clinical practice yet. A diagnosis is performed based on more criteria besides color and basing the decision solely on this criterion could lead to an incorrect diagnosis. In future work we would like to extend our corr-LDA model to other relevant dermoscopic structures and use that information to obtain a more reliable CAD system and improve the classification results.

## 9. Conclusions

This work describes a clinically oriented CAD system where the diagnosis of the skin lesions is performed based on a medical color description. The proposed system comprises two main tasks: i) detection of relevant colors using a probabilistic model and ii) diagnosis of lesions using the obtained color information. The system was trained and tested on a dataset of 482 dermoscopy images.

Our main challenge was to learn a probabilistic model using weakly annotated dermoscopy images, i.e., we knew the color labels of each image but did not know the location of each of the colors in the lesion. We addressed this issue using the Correspondence Latent Dirichlet Allocation (corr-LDA) algorithm to obtain a probabilistic model that relates text labels and image features. This allowed us to simultaneously obtain global image labels and individually annotate image patches. Due to the type of image features used (mean color in the HSV space), it was necessary to modify the original formulation of corr-LDA in order to incorporate a von-Misses-Gaussian distribution. This distribution was more suitable to describe the data. The results were promising, with the following average scores for color detection: Precision = 83.8% and Recall = 89.8%.

We have also addressed the problem of lesion classification using the extracted medical information. Four different strategies were studied in order to determine which is the best way to use the color information to classify the lesions. Our results showed that one of the strategies outperformed the remaining, leading to a diagnostic Sensitivity of 76.3% and a Specificity of 70.5% using Random Forests. Combining the four types of features allowed us to achieve the best classification results, with a Sensitivity of 77.6% and a Specificity of 73.0% using Random Forests and a Sensitivity of 75.1% and Specificity of 77.5% using SVM.

In the future we would like to extend our model to other clinically relevant dermoscopic criteria in order to obtain a more robust and reliable system.

## Acknowledgments

This work was partially funded with grant SFRH/BD/84658/2012 830 and by the FCT projects FCT [UID/EEA/50009/2013] and PTDC/EEIPRO/0426/2014.

## Appendix A. Expanded lower bound $\mathcal{L}(\gamma, \phi, \lambda; \alpha, \beta, \Omega)$

Here we show the expanded version of every term in (5), such that we obtain a lower bound $\mathcal{L}(\gamma, \phi, \lambda; \alpha, \beta, \Omega)$ as a function of the model $(\alpha, \beta, \Omega)$ and variational $(\gamma, \phi, \lambda)$ parameters. We do not show how to expand each of the terms, so for details on these derivations please refer to [35,40].

$$\mathbb{E}_q[\log p(\theta|\alpha)] = \log\Gamma\left(\sum_{k=1}^{K}\alpha_k\right) - \sum_{k=1}^{K}\log\Gamma(\alpha_k) + \sum_{k=1}^{K}(\alpha_k - 1)\left(\Psi(\gamma_k) - \Psi\left(\sum_{j=1}^{K}\gamma_j\right)\right), \quad \text{(A.1)}$$

where $\Gamma(.)$ is the gamma function and $\Psi(.)$ is the digamma function, i.e., the first derivative of the gamma function.

$$\mathbb{E}_q[\log p(\mathbf{z}|\theta)] = \sum_{n=1}^{N}\sum_{k=1}^{K}\left(\Psi(\gamma_k) - \Psi\left(\sum_{j=1}^{K}\gamma_j\right)\right)\phi_{nk}. \quad \text{(A.2)}$$

$$\mathbb{E}_q[\log p(\mathbf{r}|\mathbf{z}, \Omega)] = \sum_{n=1}^{N}\sum_{k=1}^{K}\phi_{nk}\log p(r_n|z_n = k, \Omega). \quad \text{(A.3)}$$

$$\mathbb{E}_q[\log p(\mathbf{y}|N)] = C, \quad \text{(A.4)}$$

where $C$ is a constant.

$$\mathbb{E}_q[\log p(\mathbf{w}|\mathbf{y}, \mathbf{z}, \beta)] = \sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{k=1}^{K}\lambda_{mn}\phi_{nk}\log p(w_m|y_m = n, z_n = k, \beta). \quad \text{(A.5)}$$

$$\mathbb{E}_q[\log q(\theta|\gamma)] = \log\Gamma\left(\sum_{k=1}^{K}\gamma_k\right) - \sum_{k=1}^{K}\log\Gamma(\gamma_k) + \sum_{k=1}^{K}(\alpha_k - 1)\left(\Psi(\gamma_k) - \Psi\left(\sum_{j=1}^{K}\gamma_j\right)\right). \quad \text{(A.6)}$$

$$\mathbb{E}_q[\log q(\mathbf{z}|\phi)] = \sum_{n=1}^{N}\sum_{k=1}^{K}\phi_{nk}\log\phi_{nk}. \quad \text{(A.7)}$$

$$\mathbb{E}_q[\log q(\mathbf{y}|\lambda)] = \sum_{m=1}^{M}\sum_{n=1}^{N}\lambda_{mn}\log\lambda_{mn}. \quad \text{(A.8)}$$

## References


[1] G. Argenziano, H.P. Soyer, V. De Giorgi, D. Piccolo, P. Carli, M. Delfino, A. Ferrari, V. Hofmann-Wellenhog, D. Massi, G. Mazzocchetti, M. Scalvenzi, I.H. Wolf, Interactive Atlas of Dermoscopy, EDRA Medical Publishing & New Media, 2000.

[2] J. Ferlay, E. Steliarova-Foucher, J. Lortet-Tieulent, S. Rosso, J. Coebergh, H. Comber, D. Forman, F. Bray, Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012, Eur. J. Cancer 49 (6) (2013) 1374–1403.

[3] A.W. Kopf, M. Elbaum, N. Provost, The use of dermoscopy and digital imaging in the diagnosis of cutaneous malignant melanoma, Skin Res. Technol. 3 (1) (1997) 1–7.

[4] J. Mayer, Systematic review of the diagnostic accuracy of dermatoscopy in detecting malignant melanoma., Med. J. Aust. 167 (4) (1997) 206–210.

[5] W. Stolz, A. Riemann, A.B. Cognetta, ABCD rule of dermatoscopy: a new practical method for early recognition of malignant melanoma, Eur. J. Dermatol. 4 (1994) 521–527.

[6] G. Argenziano, G. Fabbrocini, P. Carli, V. De Giorgi, E. Sammarco, E. Delfino, 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 (1998) 1563–1570.

[7] M. Binder, M. Puespoeck-Schwarz, A. Steiner, H. Kittler, M. Muellner, K. Wolff, H. Pehamberger, Epiluminescence microscopy of small pigmented skin lesions: short-term formal training improves the diagnostic performance of dermatologists, J. Am. Acad. Dermatol. 36 (2) (1997) 197–202.

[8] I. Wolf, J. Smolle, H. Soyer, H. Kerl, Sensitivity in the clinical diagnosis of malignant melanoma, Melanoma Res. 8 (5) (1998) 425–429.

[9] H. Ganster, A. Pinz, R. Rohrer, E. Wildling, M. Binder, H. Kittler, Automated melanoma recognition, IEEE Trans. Med. Imag., 20 (3) (2001) 233–239.

[10] M.E. Celebi, H. Kingravi, B. Uddin, H. Iyatomi, Y.A. Aslandogan, W. Stoecker, R. Moss, A methodological approach to the classification of dermoscopy images, Comput. Med. Imag. Gr. 31 (6) (2007) 362–373.

[11] H. Iyatomi, H. Oka, M.E. Celebi, M. Hashimoto, M. Hagiwara, M. Tanaka, K. Ogawa, An improved internet-based melanoma screening system with dermatologist-like tumor area extraction algorithm, Comput. Med. Imag. Gr. 32 (7) (2008) 566–579.

[12] C. Barata, M. Ruela, M. Francisco, T. Mendona, J.S. Marques, Two systems for the detection of melanomas in dermoscopy images using texture and color features, IEEE Syst. J. 8 (3) (2014) 965–979.

[13] K. Korotkov, R. Garcia, Computerized analysis of pigmented skin lesions: a review, Artif. Intell. Med. 56 (2) (2012) 69–90.

[14] K. Hoffmann, T. Gambichler, A. Rick, M. Kreutz, M. Anschuetz, T. Grünendick, A. Orlikov, S. Gehlen, R. Perotti, L. Andreassi, Diagnostic and neural analysis of skin cancer (danaos). a multicentre study for collection and computer-aided analysis of data from pigmented skin lesions using digital dermoscopy, Brit. J. Dermatol. 149 (4) (2003) 801–809.

[15] S. Dreiseitl, M. Binder, Do physicians value decision support? a look at the effect of decision support systems on physician opinion, Artif. Intell. Med. 33 (1) (2005) 25–30.

[16] G. Betta, G. Di Leo, G. Fabbrocini, A. Paolillo, P. Sommella, Dermoscopic image-analysis system: estimation of atypical pigment network and atypical vascular pattern, in: Proceedings of the 2006 IEEE International Workshop on Medical Measurement and Applications (MeMea), IEEE, 2006, pp. 63–67.

[17] M. Anantha, R. Moss, W. Stoecker, Detection of pigment network in dermatoscopy images using texture analysis, Comput. Med. Imag. Gr. 28 (5) (2004) 225–234.

[18] M. Sadeghi, M. Razmara, T.K. Lee, M.S. Atkins, A novel method for detection of pigment network in dermoscopic images using graphs, Comput. Med. Imag. Gr. 35 (2) (2011) 137–143.

[19] C. Barata, J.S. Marques, J. Rozeira, A system for the detection of pigment network in dermoscopy images using directional filters, IEEE Trans. Biomed. Eng. 59 (10) (2012) 2744–2754.

[20] J.L. García, B.Z. García, Detection of pigment network in dermoscopy images using supervised machine learning and structural analysis, Comput. Biol. Med. 44 (2014) 144–157.

[21] H. Mirzaalian, T. Lee, G. Hamarneh, Learning features for streak detection in dermoscopic color images using localized radial flux of principal intensity curvature, in: Proceedings of the 2012 IEEE Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA), IEEE, 2012, pp. 97–101.

[22] M. Sadeghi, T. Lee, H. Lui, D. McLean, S. Atkins, Detection and analysis of irregular streaks in dermoscopic images of skin lesions, IEEE Trans. Med. Imag. 32 (2013) 849–861.

[23] S. Yoshino, T. Tanaka, M. Tanaka, H. Oka, Application of morphology for detection of dots in tumor, in: Proceedings of the SICE 2004 Annual Conference, vol. 1, IEEE, 2004, pp. 591–594.

[24] S. Seidenari, G. Pellacani, C. Grana, Computer description of colours in dermoscopic melanocytic lesion images reproducing clinical assessment, Brit. J. Dermatol. 149 (3) (2003) 523–529.

[25] A.R.S. Marcal, T. Mendonca, C.S.P. Silva, M.A. Pereira, R. J., Evaluation of the menzies method potential for automatic dermoscopic image analysis, in: Proceedings of the Computational Modelling of Objects Represented in Images - CompImage 2012, 2012, pp. 103–108.

[26] C. Barata, M.A.T. Figueiredo, M.E. Celebi, J.S. Marques, Color identification in dermoscopy images using gaussian mixture models, in: Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 3611–3615.

[27] M.E. Celebi, A. Zornberg, Automated quantification of clinically significant colors in dermoscopy images and its application to skin lesion classification, IEEE Syst. J. 8 (3) (2014) 980–984.

[28] M.E. Celebi, H. Iyatomi, W.V. Stoecker, R.H. Moss, H.S. Rabinovitz, G. Argenziano, H.P. Soyer, Automatic detection of blue-white veil and related structures in dermoscopy images, Comput. Med. Imag. Gr. 32 (2008) 670–677.

[29] G. Di Leo, G. Fabbrocini, A. Paolillo, O. Rescigno, P. Sommella, Towards an automatic diagnosis system for skin lesions: estimation of blue-whitish veil and regression structures, in: Proceedings of the 6th International Multi-Conference on Systems, Signals and Devices, 2009. SSD'09., IEEE, 2009, pp. 1–6.

[30] A. Madooei, M.S. Drew, M. Sadeghi, M.S. Atkins, Automatic detection of blue-white veil by discrete colour matching in dermoscopy images, in: Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013, Springer, 2013, pp. 453–460.

[31] M. Lingala, R. Stanley, R. Rader, J. Hagerty, H. Rabinovitz, M. Oliviero, I. Choudhry, W. Stoecker, Fuzzy logic color detection: Blue areas in melanoma dermoscopy images, Comput. Med. Imag. Gr. 38 (5) (2014) 403–410.

[32] C. Serrano, B. Acha, Pattern analysis of dermoscopic images based on markov random fields, Pattern Recog. 42 (6) (2009) 1052–1057.

[33] Q. Abbas, M.E. Celebi, C. Serrano, I.G. Fondón, G. Ma, Pattern classification of dermoscopy images: a perceptually uniform model, Pattern Recog. 46 (1) (2013) 86–97.

[34] A. Sáez, C. Serrano, B. Acha, Model-based classification methods of global patterns in dermoscopic images, IEEE Trans. Med. Imag. 33 (5) (2014) 1137–1147.

[35] D. Blei, M. Jordan, Modeling annotated data, in: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, ACM, 2003, pp. 127–134.

[36] M.E. Celebi, Q. Wen, S. Hwang, G. Schaefer, Color quantization of dermoscopy images using the k-means clustering algorithm, in: Color Medical Image Analysis, Springer, 2013, pp. 87–107.

[37] G. Carneiro, A. Chan, P. Moreno, N. Vasconcelos, Supervised learning of semantic classes for image annotation and retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 29 (3) (2007) 394–410.

[38] D. Zhang, M.M. Islam, G. Lu, A review on automatic image annotation techniques, Pattern Recog. 45 (1) (2012) 346–362.

[39] K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. Blei, M. Jordan, Matching words and pictures, J. Mach. Learning Res. 3 (2003) 1107–1135.

[40] D. Blei, A. Ng, M. Jordan, Latent dirichlet allocation, J. Mach. Learning Res. 3 (2003) 993–1022.

[41] L. Fei-Fei, P. Perona, A bayesian hierarchical model for learning natural scene categories, in: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, IEEE, 2005, pp. 524–531.

[42] C. Wang, D. Blei, L. Fei-Fei, Simultaneous image classification and annotation, in: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2009, pp. 1903–1910.

[43] L. Li, R. Socher, L. Fei-Fei, Towards total scene understanding: Classification, annotation and segmentation in an automatic framework, in: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)., IEEE, 2009, pp. 2036–2043.

[44] R. Socher, L. Fei-Fei, Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora, in: Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2010, pp. 966–973.

[45] M. Silveira, J.C. Nascimento, J.S. Marques, A. Marçal, T. Mendonça, S. Yamauchi, J. Maeda, J. Rozeira, Comparison of segmentation methods for melanoma diagnosis in dermoscopy images, IEEE J. Selected Topics Signal Process. 3 (1) (2009) 35–45.

[46] M.E. Celebi, H. Iyatomi, G. Schaefer, W.V. Stoecker, Lesion border detection in dermoscopy images, Computer. Med. Imag. Gr. 33 (2) (2009) 148–153.

[47] M.E. Celebi, Q. Wen, S. Hwang, H. Iyatomi, G. Schaefer, Lesion border detection in dermoscopy images using ensembles of thresholding methods, Skin Res. Technol. 19 (1) (2013) 252–258.

[48] G. Capdehourat, A. Corez, A. Bazzano, R. Alonso, P. Musé, Toward a combined tool to assist dermatologists in melanoma detection from dermoscopic images of pigmented skin lesions, Pattern Recog. Lett. 32 (16) (2011) 2187–2196.

[49] C. Barata, J.S. Marques, M.E. Celebi, Towards an automatic bag-of-features model for the classification of dermoscopy images: the influence of segmentation, in: Proceedings of the 2013 8th International Symposium on Image and Signal Processing and Analysis (ISPA), IEEE, 2013, pp. 274–279.

[50] S. Calderara, A. Prati, R. Cucchiara, Mixtures of von Mises distributions for people trajectory shape analysis, IEEE Trans. Circuits Syst. Video Technol. 21 (4) (2011) 457–471.
[51] S. Sra, A short note on parameter approximation for von mises-fisher distributions: and a fast implementation of $I_s(x)$, Comput. Stat. 27 (1) (2012) 177–190.
[52] R. MacKie, C. Fleming, A. McMahon, P. Jarrett, The use of the dermatoscope to identify early melanoma using the three-colour test, Brit. J. Dermatol. 146 (3) (2002) 481–484.
[53] C. Barata, M. Celebi, J. Marques, Improving dermoscopy image classification using color constancy, IEEE J. Biomed. Health Informat. 19 (3) (2015) 1146–1152.
[54] C. Barata, J. Marques, M. Celebi, Improving dermoscopy image analysis using color constancy, in: Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP), IEEE, 2014, pp. 3527–3531.
[55] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (8) (2005) 1226–1238.
[56] T. Mendonça, P.M. Ferreira, J.S. Marques, A.R.S. Marcal, J. Rozeira, Ph 2-a dermoscopic image database for research and benchmarking, in: Proceedings of the 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2013, pp. 5437–5440.