Pattern Recognition 69 (2017) 270–285

Contents lists available at ScienceDirect

Pattern Recognition

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

Development of a clinically oriented system for melanoma diagnosis

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

[a] Institute for Systems and Robotics, Instituto Superio Técnico, Lisboa, Portugal
[b] Department of Computer Science, University of Central Arkansas, USA

ARTICLE INFO

Article history:
Received 10 January 2017
Revised 24 March 2017
Accepted 17 April 2017
Available online 25 April 2017

Keywords:
Melanoma diagnosis
Computer aided diagnosis system
Detection of dermoscopic structures
Image annotation
Corr-LDA
Feature fusion

ABSTRACT

Dermatologists have stated their preference for computer aided diagnosis (CAD) systems that provide medical justifications for the estimated diagnosis of a skin lesion. Such systems are considered to be clinically oriented in the sense that they try to detect clinical criteria and then perform a diagnosis based on that information. Unfortunately, the development of clinically inspired systems is hampered by several challenges: (i) the lack of datasets with detailed information regarding the presence and location of clinical criteria; (ii) the subtlety of some diagnostic criteria, which makes them difficult to detect; and (iii) the difficulty of using the detected criteria to predict a diagnosis. In this work, we propose a machine learning framework to address these issues. First, an image annotation approach is used to detect various medical criteria (color, texture and color structures). Information is, then, extracted from the detected criteria and a late fusion method is used to obtain a lesion diagnosis. A sensitivity of 84.6% and a specificity of 74.2% are obtained on a multi-source dataset of 804 images.

1. Introduction

1.1. Motivation

Cancer is one of the leading causes of death worldwide, being the second major cause of death and morbidity in Europe alone [1]. Melanoma is the deadliest form of skin cancer, ranking in the ninth position among the most common types of cancer in Europe, while the American Cancer Society estimates that more than 73,000 new cases are diagnosed each year [2]. The mortality rates of melanoma are associated with its high potential to metastasize in later stages, propagating to other sites in the body such as lungs, bones, or brain. A common technique used by dermatologists to diagnose this disease is dermoscopy, which allows the observation of structures and colors, otherwise invisible to the naked eye. Although there are established medical procedures to analyze dermoscopy images (*e.g.*, ABCD rule [3] and 7-point checklist [4]), this is still a subjective process that heavily relies both on the experience and visual acuity of the practitioner [5]. These limitations have fostered the proposal of several computer aided diagnosis (CAD) systems, for the analysis of dermoscopy images [6].

This paper describes a clinically oriented CAD system that is capable of identifying medical features in the lesions and diagnose them, basing the decision on the detected features. Medical feedback tells us that physicians prefer these kinds of systems, over the ones (*e.g.*, [7,8]) that use abstract image features, such as color histograms or the gray level co-occurrence matrix, to characterize the lesions. Although the latter achieve promising experimental results, dermatologists feel they are not informative enough and lack medical meaning [9].

The development of a clinically inspired system is a challenging task that is hampered by several factors: (i) the lack of datasets with detailed annotations of the medical features and their locations; (ii) the difficulty of distinguishing and detecting some of the medical features; and (iii) the difficulty of using the detected criteria to predict a diagnosis. In this paper we propose a machine learning framework to tackle the aforementioned issues.

1.2. Related work

The number of works that address the development of clinically inspired CAD systems has been growing over the last decade [6,10]. These works can be divided into two categories, according to the medical algorithms that they try to replicate: (i) methods that try to detect global patterns and (ii) methods that try to detect localized dermoscopic criteria.

The first type of method aims at mimicking a medical approach called pattern analysis [11]. This technique consists of inspecting the lesion for the presence of specific patterns associated with texture structures, such as pigment network, streaks, and globules. A lesion is considered to be represented by one of patterns if its characteristic dermoscopic criterion is the predominant one (the

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

http://dx.doi.org/10.1016/j.patcog.2017.04.023
0031-3203/© 2017 Elsevier Ltd. All rights reserved.

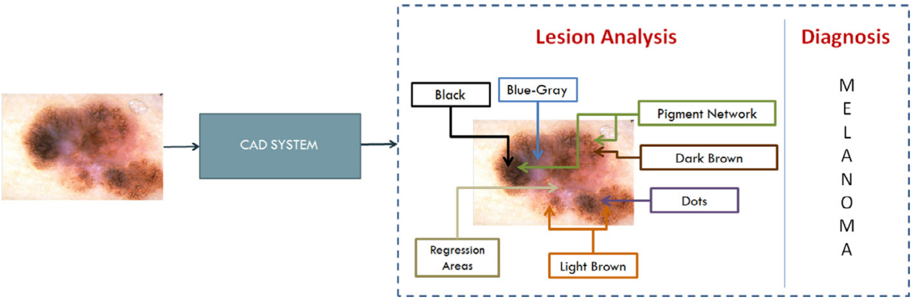


Fig. 1. Desired output of a clinically inspired system.

one which covers the largest area) [5]. Different strategies have been proposed to automatically associate the lesion with a pattern. Some methods characterize the whole lesion using a single feature vector that comprises information about color, texture, and symmetry. This information is then used to either train a different classifier for each of the patterns (e.g. [12,13]), or to train a multiclass algorithm, as proposed in [14]. Alternatively, other methods divide the lesion into small regions and separately characterize each one of them. Then, regions that belong to the same type of pattern are grouped and used to train a classifier [15–17].

The majority of clinically inspired CAD systems try to detect dermoscopic criteria. These methods are based on medical procedures such as the ABCD rule and 7-point checklist. A bibliographic search shows that it is possible to detect a multitude of criteria. Pigment network is one of the most popular criteria due to its medical relevance [18–27]. The remaining texture structures are less popular. Nonetheless, some works can be found on literature regarding the detection of streaks [28–30] and dots/globules [18,31,32].

The detection of color criteria is also an active topic of research (see [33] for a review on this topic). Several works were inspired by the ABCD rule [3] and proposed strategies to detect and/or quantify clinically relevant colors [34–40]. Several of these works start by computing a color palette to represent the possible colors. This palette is estimated using color segmentations provided by experienced dermatologists. New images are divided in small regions, and each region is compared with the templates and associated with the closest one. In order to estimate the color models, a training set of images with segmented colors is required. Moreover, since the assessment of colors is a subjective process that relies on the visual perception of the practitioner, it is necessary to have more than one dermatologist providing the color segmentations for the same lesion. This is a tedious approach. Other methods try to avoid these limitations by applying a clustering algorithm in order to count the number of colors that can be found in the lesion [40,41]. However, these methods do not discriminate which are the colors that are detected.

Other works focus on the detection of color structures that are usually associated with melanoma. Two of the studied criteria are associated with abnormal pigmentation, namely dark blotches [36,42–44] and regions of decreased pigmentation [45–47]. Other relevant color based structures are the white regression areas [22,45,48,49], which are lesion regions that have a scar like aspect, and blue-whitish veil [22,48,50,51], which appears as a gray-blue to whitish-blue diffuse pigmentation. Multiple strategies have been proposed to detect these structures. A common step to most of the approaches is to request dermatologists to identify regions within the lesions where the criterion is present and regions where it is absent. Then, features are extracted from both types of regions and used to train a classifier, namely decision trees [22,48,50] and neural networks [45,49]. An alternative strategy consists of learning a color palette using the region examples, and then use a nearest neighbor approach to label new regions according to the estimated palette [51].

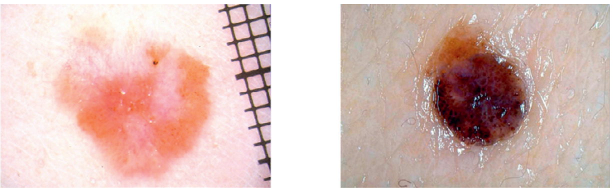
Several problems can be pointed out to the works found in literature. Few of them attempt to use the detected medical criteria to diagnose the lesions [12,14,17,22,34,39,40,45,50], and even fewer attempt to detect more than one structure [22,36,45,48]. Furthermore, a significant number of the mentioned methods require reliable segmentations of the criteria in order to be able to learn a classifier, which are hard to obtain. The lack of segmentations hampers the application of several systems and, in some cases, makes it impossible to detect the criteria (*e.g.*, colors). Recently, different works have tried to deal with this problem and learn from weakly annotated data (the information available for training consists of text labels). One of the works proposed a model for the detection of blue-whitish veil using a multiple-instance framework [52], while another used a probabilistic method called correspondence latent Dirichlet allocation (corr-LDA) to detect colors [53]. Despite their relevance, these focus on the detection of a single criteria, which is clearly insufficient as the diagnostic performed by an expert is based on more than one criterion.

2. Problem statement

Based on the literature, a clinically inspired system must fulfill two requirements [9]. Both of them are exemplified in Fig. 1:

1. Provide relevant clinical information to the dermatologists, *i.e.*, replicate their identification of relevant criteria. The system should provide a set of text labels stating which are the clinical criteria that are present in the lesion and associate those labels with specific regions, such that they can be checked by the physicians.
2. Diagnose the lesions, basing that decision on the detected clinical criteria. This ensures that the features used by the system have a medical meaning, making it possible for the dermatologist to understand and validate the automated diagnosis.

   These requirements raise a set of problems. First, the medical criteria can be very subtle structures. Second, the development of strategies to detect clinical criteria usually requires large datasets of images with detailed information: text annotations stating which are the criteria that can be found in the lesion and corresponding region segmentations. The segmentations are used in the training of several algorithms, as mentioned in Section 1.2. Unfortunately, most datasets lack the segmentations and only provide text labels, as exemplified in Fig. 2, since performing the latter is seen as a time consuming and subjective task by the experts. Finally, it is not easy for a computer to convert the detected criteria into information that can be used to automatically diagnose melanomas.



Colors: Dark brown, light brown, red, and white.
Texture Structures: Dots.
Color Structures: Regression areas.
Diagnosis: Melanoma.

Colors: Dark brown, black, light brown, and blue.
Texture Structures: Pigment network and dots.
Color Structures: Blue-whitish veil.
Diagnosis: Benign.

Fig. 2. Images and annotations provided by dermatologists [5].

Each of the aforementioned problems is addressed in this work. The lack of reliable segmentations associated to the clinical features is tackled using an image annotation approach called corr-LDA [54], where the idea is to learn the joint probability distribution of text labels provided by experts and image regions. Annotation algorithms are trained using weakly annotated data (with text labels only), so it is appropriate to formulate the detection of medical criteria as an annotation problem.

Since some medical criteria can be difficult to detect, it is important to select a subset that play an important role in the diagnosis. The selected criteria correspond to different characteristics of the lesions and will be divided into three classes (as exemplified in Fig. 2): (i) six colors (C) defined in the ABCD rule (dark and light browns, blue-gray, black, white, and red) [3]; (ii) two texture structures (TS), called pigment network and dots/globules [3]; and (iii) two color structures (CS) assessed in the 7-point method [4] - blue-whitish veil and white regression areas. All of these criteria are then used to extract appropriate features for lesion diagnosis.

3. Comparison with other works

There are two main differences between the proposed system and other clinically inspired methods: i) most of the methods focus only on the detection of one or two clinical criteria [6], while the proposed method detects a larger number of criteria that characterize different aspects of the lesion; and ii) few methods try to diagnose melanomas using the clinical features [6], which is performed in this work.

Recently, we applied to the problem of color detection in dermoscopy images [53]. This paper proposes significant changes. First, it extends the detection framework to a wider set of clinical criteria besides color, which is a necessary step, since medical experts use more criteria besides color. The selected criteria cover different properties of the lesions and can be associated to the same region, making them a challenge. The extension to other criteria required several modifications, such as the inclusion of a more sophisticated lesion segmentation strategy and new types of region descriptors. Moreover, this paper also explores different methodologies to combine their information, in order to diagnose the lesion as melanoma or benign. Finally, the experiments were carried on a larger dataset (804 against 482 images).

4. System overview

This section succinctly describes the proposed clinically inspired CAD system. The sequential framework of the system is similar to the analysis performed by dermatologists, i.e., first the system tries to identify the presence or relevant dermoscopic criteria and then performs a diagnosis using this information. Fig. 3 exemplifies the pipeline of the system.

The first step of the system consists of dividing the image into smaller regions (see Fig. 3), each characterized by a feature vector r_n. An image is assumed to be characterized by a set $\mathbf{r} = \{r_1, \ldots, r_N\}$, which comprises the feature vectors from all of the N regions. The segmentation strategy will be discussed in Section 5.1. The features that characterized the regions are discussed in Section 5.2. It is assumed that each of the images can have one or more text labels.

The next step is the identification of the dermoscopic criteria that are present in the lesion. This is a two fold task, as exemplified in Fig. 3, where the criteria are associated with one or more regions (local labels) and text labels are produced for the entire image (global labels). The detection of the criteria is achieved using a probabilistic model called corr-LDA [54], which is estimated using a database of weakly annotated images. The probabilistic formulation of corr-LDA as well as specific aspects of the annotation process will be discussed in Section 5.3.

The final block of the system classifies the lesion as melanoma or benign using information extracted from the detected medical criteria. This task requires the use of a classification algorithm, which is learned using a dataset of dermoscopy images diagnosed by experts. The learning process of the classifier works as follows. First, the estimated corr-LDA models are applied to the training images. Then, new features are extracted from their output. Finally,

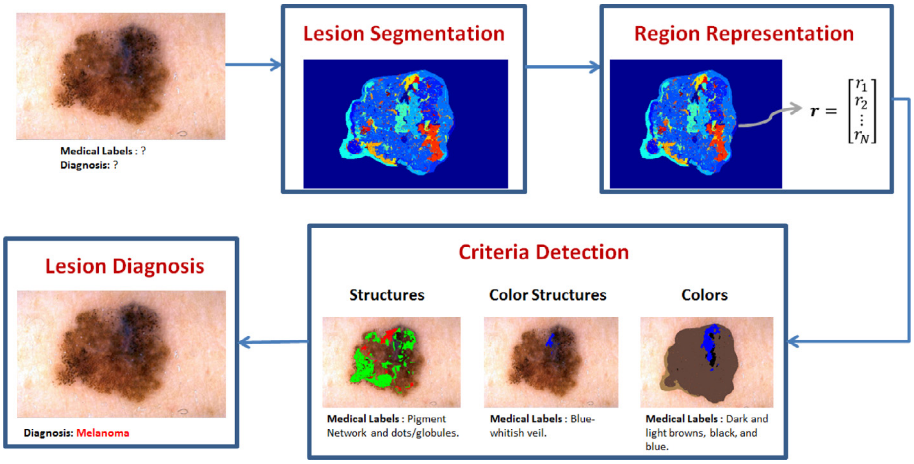

Medical Labels : ?
Diagnosis: ?
Lesion Segmentation
Region Representation
$r = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_N \end{bmatrix}$
Lesion Diagnosis
Diagnosis: Melanoma
Criteria Detection
Structures
Color Structures
Colors
Medical Labels : Pigment Network and dots/globules.
Medical Labels : Blue-whitish veil.
Medical Labels : Dark and light browns, black, and blue.

Fig. 3. Proposed CAD System.

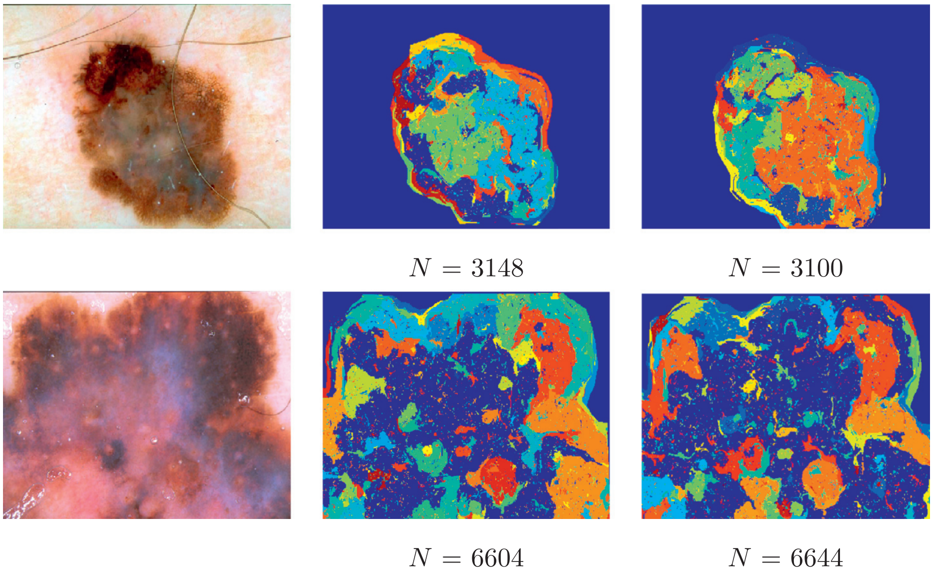


Fig. 4. Image segmentation: original image (left), segmentation using color features (mid), and segmentation using color and texture features (right). Each color label represents a different region. The number of regions obtained by the algorithm (N) is also shown.

these features are used to train the classifier. In the case of new images, their corr-LDA outputs are used as features and then the classifier is applied to predict the diagnosis. Detailed information about the classification approach is provided in Section 5.4.

The main advantage of the proposed system is its ability to interact with the dermatologist, since the automated diagnosis relies on descriptors that are medically inspired and can be checked by clinicians. This allows them to understand and validate the suggested lesion diagnosis. Furthermore, its sequential framework is similar to the analysis performed by an expert: first look for several dermoscopic criteria and then diagnose the lesion . These two characteristics of the proposed system make it valuable for the medical community and make it significantly different from other systems found in literature [9].

5. Proposed system

5.1. Lesion segmentation

The goal of the segmentation block is to divide the lesion into different regions. Previously, segmentation was performed using a regular grid approach [53]. This method is not appropriate for this setting, since a block can contain more than one color or structure. Such output is undesirable and might hamper the training and testing of the detection model. Thus, it is important to use a different segmentation method. Assuming that each dermoscopic criteria has specific color and texture properties, and that the lesion is segmented into regions with homogeneous color and texture, it is possible to ensure that each region will contain only one color or structure. It is also important to have in mind that structures like pigment network may appear with different sizes (scales). Hence, it is also important to ensure that the obtained regions are invariant to scale.

Lesion segmentation is performed using the method proposed in [55]. This is a graph-based algorithm that assumes that each of the pixels in the image is one vertice of a graph, and that neighbor pixels are linked through edges. The weight of each edge is given by the Euclidean distance between the feature vectors of the two neighbor pixels. Connected vertices/pixels are combined in the same region if they are similar, i.e., if their corresponding edge weight is lower than a given threshold δ. The value of δ defines the size of the obtained regions and is influenced by the resolution of the images. It was experimentally found that setting $\delta = \frac{L^d+B^d}{130}$, where $L^d \times B^d$ is the resolution of image d (*i.e.*, respectively the number of columns and rows), led to a good trade-off between the size of the regions and the computation time.

Each pixel in the image is characterized by a feature vector that contains color and texture information. A common strategy to obtain scale invariance is to enforce it during the feature extraction process. In this work, this task is performed using the method proposed by Carson et al. [56]. The idea is to determine the optimal scale for each pixel using a local image property called polarity [56], and then extract its features at that scale.

The color information of a pixel are its L*a*b* components. These components are determined after performing spatial averaging using a Gaussian with $s^2 = s^*(x, y)^2$, where s^* is the ideal scale. The texture information is characterized by the contrast, contrast × polarity, and anisotropy × contrast [56]. The second moment matrix at each pixel is used to compute contrast and anisotropy

$$M_s(x, y) = G_s(x, y) * (\nabla I)(\nabla I)^T, \tag{1}$$

where $s = s^*(x, y)$ is the pixel's best scale. From this matrix, anisotropy and contrast at each pixel (x, y) are respectively defined as:

$$a(x, y) = 1 - \frac{\lambda_2}{\lambda_1} \quad c(x, y) = 2\sqrt{\lambda_1 + \lambda_2}, \tag{2}$$

where λ_1, λ_2 are the eigenvalues of $M_s(x, y)$.

Two segmentations are performed for each image, one using only color features and the other using color and texture features, as exemplified in Fig. 4. The first segmentation is used to train/test the corr-LDA associated with the color criteria, while the other is used to train the models associated with color and texture structures.

5.2. Feature extraction - region representation

After segmenting the lesions, as exemplified in Fig. 4, each of the $1, 2, \ldots, N$ regions is characterized by a feature vector $r_n \in \mathbb{R}^f$. This feature vector comprises information about color, texture or both, depending on the type of dermoscopic criteria. An image d is characterized by a set $\mathbf{r}^d = \{r_1^d, \ldots, r_N^d\} \in \mathbb{R}^{f \times N_d}$ of N^d vectors (recall from Fig. 4 that each lesion is segmented into a different number of regions). Since we do not know which is the best type of features that can be used to describe the regions, combinations of the following descriptors were tested for each type of criterion.

- **Colors:** The mean color vector in the HSV space (μ_{HSV}).
- **Texture structures:** The regions are described using texture features. In this work the tested features are the mean contrast (μ_c) and mean contrast × anisotropy (μ_{ca}), the mean (μ_g) and standard deviation (σ_g) of the gray level values in the region, and statistics computed using the directional filters proposed in [25]. These filters are computed at different orientations $\theta_i \in [0, \pi]$, $i = 0, \ldots, 9$, with the impulse response for direction θ_i given by

$$h_{\theta_i}(x, y) = G_1(x, y) - G_2(x, y), \tag{3}$$

where G_k is a Gaussian filter:

$$G_k(x, y) = C_k \exp\left\{-\frac{x'^2}{2\sigma^2_{x_k}} - \frac{y'^2}{2\sigma^2_{y_k}}\right\}, k = 1, 2. \tag{4}$$

In (4) C_k is a normalization constant and the values of (x', y') are related with (x, y) by a rotation of amplitude θ_i.

$$\begin{aligned} x' &= x\cos\theta_i + y\sin\theta_i, \\ y' &= y\cos\theta_i - x\sin\theta_i. \end{aligned} \tag{5}$$

We compute the output of the directional filters (3) for all the directions and keep the maximum and minimum. The regions are described by the mean and standard deviation of these values (μ_M, σ_M, μ_m, and σ_m).

- **Color structures:** These structures simultaneously exhibit color and texture properties. The color of the regions is characterized using μ_{HSV}, while two texture descriptors are compared: (μ_c, μ_{ca}) and (μ_M, σ_M, μ_m, σ_m).

5.3. Image annotation & detection of medical criteria

5.3.1. Annotation methods - overview

The goal of image annotation methods is to find a relationship between text labels and image features, such that a computer can replicate the annotation process in new images and, on some occasions, associating those concepts with specific regions of the image (semantic segmentation). One of the major challenges of image annotation is the lack of completely annotated data. This means that most of the algorithms 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 [57,58]. This is very similar to the problem discussed in this work: we want to be able to assign dermoscopic criteria to new images and locate them in the lesion, but the training data consists only of text labels. It is possible to divide annotation algorithms into two groups. Those that treat annotation as a multi-class classification problem, and those that want to estimate a probabilistic model to characterize the relationship between image or region features and the text labels [58].

The common approach used by the classification methods is learn a separate classifier for each of the labels, treating the multi-class problem as a set of binary ones. This kind of approach can have a high computational cost. Moreover, several of the classification methods are only suitable to perform annotation at the image level, which means that they do not allow the identification of the regions that are associated with each of the labels (recall that the localization of the dermoscopic criteria is a goal of this work). An exception is the multiple-instance multi-label (MIML) framework that divides the image into regions, being then able to model the relationship between them and the text labels (*e.g.*, [59–62]). However, this methodology has a set of drawbacks. It lacks robustness in the presence of outliers (a single outlier can bias the solution) and it is highly sensitive to initialization. Moreover, these methods are restrictive in the sense that they usually require the definition of the number of instances, which can be accomplished by either breaking the images into a fixed number of regions or by applying the bag-of-features framework before the learning phase [62]. If the latter is performed, the clustering step might introduce errors and create misleading prototypes. Finally, they still rely on decomposing the learning task into a series of single-class multi-instance learning procedures.

An alternative to classification methods are the ones that use a probabilistic formulation to model the co-occurrence of image features and labels. The general idea is to use a Bayesian framework to estimate the posterior distribution of each of the admissible labels, given the observation of features from the image, defined as $p(w_m|\mathbf{r})$, where w_m is the mth possible annotation and $\mathbf{r}$ is the set of image features. Some probabilistic methods are not suitable for our work, since they work at the image level and do not allow region labeling (*e.g.* [63,64]). Other methods divide the image into regions, characterize each of them by a feature vector, and finally try to translate them into text labels. This last step is accomplished by first clustering the region features into a set of centroids and then computing the co-occurrence between them and each of the admissible text labels (*e.g.*, [65–67]). This allows the association between text labels and regions. However, assuming that each region inherits all of the labels associated with the entire image is not appropriate, and the clustering step may introduce errors in the model. Another type of region-based probabilistic methods are those that define the relationship between regions features and text labels through the use of hidden variables. An example is the corr-LDA algorithm [54], which belongs to the family of generative methods. This method assumes that there is a set of hidden variables called *topics* that are simultaneously associated with a distribution over region features and possible text labels. Under this assumption, it is possible to estimate a joint distribution of text labels and features and, consequently, the desired labeling probabilities $p(w_m|\mathbf{r})$. The downside of corr-LDA is the need to define the number of topics, but it does not suffer from the limitations of other aforementioned methods, making it an appropriate choice for this work.

5.3.2. corr-LDA formulation

corr-LDA assumes that an image can be described by a Dirichlet distribution, θ, of parameter α. This distribution is defined over a set of K latent variables, z, called topics. The topics are the core of the corr-LDA model, since each topic z_k is simultaneously associated with: (i) a distribution of region features $p(r|z_k, \Omega_k)$, with parameters Ω_k; and (ii) a multinomial distribution of the possible text labels $p(w|z_k, \beta_k)$, with parameter β_k. This formulation makes it possible to establish an indirect relationship between the region features and the text labels using topics.

Given the model parameters, the different components (topics, features, and labels) are sequentially generated, as summarized below [54]:

1. For an image d, sample a topic distribution $\theta \sim$ Dirichlet(α).
2. For each of the N^d image regions:
   (a) Sample a topic $z_n \sim$ Multinomial(θ).
   (b) Sample a region descriptor $r_n \sim p(r|z_n, \Omega)$ from a distribution conditioned on z_n.
3. For each of the M^d global labels w_m:
   (a) Sample a region indexing variable $y_m \sim \text{Unif}(1, \ldots, N)$.
   (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.

During the training phase, all of the model parameters α, $\Omega = \{\Omega_1, \ldots, \Omega_k\}$, and $\beta = \{\beta_1, \ldots, \beta_k\}$ have to be estimated, using a training set of D weakly annotated images. The traditional approach is to use a Maximum Likelihood formulation, which requires the computation of

$$p(\mathbf{r},\mathbf{w}|\alpha,\beta,\Omega)=\int_\theta\sum_z\sum_y p(\mathbf{r},\mathbf{w},\theta,\mathbf{z},\mathbf{y}|\alpha,\beta,\Omega), \tag{6}$$

This expression does not have an analytic solution. Blei and Jordan [68] address this issue using a variational method to estimate the parameters. This approach starts by introducing a new set of independent variational parameters, each associated with a distribution over a specific hidden variable of the original model. The variational parameters, here identified as (γ, ϕ, λ), allow the definition of a factorized distribution of the hidden variables

$$q(\theta,\mathbf{z},\mathbf{y})=q(\theta|\gamma).\left(\prod_{n=1}^{N}q(z_n|\phi_n)\right).\left(\prod_{m=1}^{M}q(y_m|\lambda_m)\right). \tag{7}$$

The factorized distribution can be introduced in the original log-likelihood using Jensen's inequality:

$$\begin{aligned}&\log p(\mathbf{r},\mathbf{w}|\alpha,\beta,\Omega)\\&\quad=\log\int_\theta\sum_{\mathbf{z}}\sum_{\mathbf{y}}p(\mathbf{r},\mathbf{w},\theta,\mathbf{z},\mathbf{y}|\alpha,\beta,\Omega)d\theta\\&\quad=\log\int_\theta\sum_{\mathbf{z}}\sum_{\mathbf{y}}\frac{p(\mathbf{r},\mathbf{w},\theta,\mathbf{z},\mathbf{y}|\alpha,\beta,\Omega)q(\theta,\mathbf{z},\mathbf{y})}{q(\theta,\mathbf{z},\mathbf{y})}d\theta\\&\quad\geq\mathbb{E}_q[\log p(\mathbf{r},\mathbf{w},\theta,\mathbf{z},\mathbf{y}|\alpha,\beta,\Omega)]-\mathbb{E}_q[\log q(\theta,\mathbf{z},\mathbf{y})],\end{aligned} \tag{8}$$

where $\mathbb{E}_q$ is the expected value according to the variational distribution $q(\theta$, **z, y**). The right side of the equation gives the lower bound $\mathcal{L}(\gamma,\phi,\lambda;\alpha,\beta,\Omega)$ of the log-likelihood, and can be decomposed as follows

$$\begin{aligned}\mathcal{L}(\gamma,\phi,\lambda;\alpha,\beta,\Omega)=&\ \mathbb{E}_q[\log p(\theta|\alpha)]+\mathbb{E}_q[\log p(\mathbf{z}|\theta)]\\&+\mathbb{E}_q[\log p(\mathbf{r}|\mathbf{z},\Omega)]+\mathbb{E}_q[\log p(\mathbf{y}|N)]\\&+\mathbb{E}_q[\log p(\mathbf{w}|\mathbf{y},\mathbf{z},\beta)]-\mathbb{E}_q[\log q(\theta|\gamma)]\\&-\mathbb{E}_q[\log q(\mathbf{z}|\phi)]-\mathbb{E}_q[\log q(\mathbf{y}|\lambda)].\end{aligned} \tag{9}$$

Each of the terms in (9) can be expanded into explicit functions of the model (α, β, Ω) and variational (γ, ϕ, λ) parameters. For the expanded form of $\mathcal{L}(\gamma,\phi,\lambda;\alpha,\beta,\Omega)$, the reader is referred to [54].

In the end, the problem is transformed into one of finding the best set of parameters that maximizes $\mathcal{L}$, and can be solved using a variational Expectation-Maximization (EM) algorithm:

- **E-step:** Estimate the variational parameters for each image d in the training set. The update equations of the parameters are obtained by taking derivatives of $\mathcal{L}$ with respect to each of the parameters and setting them to zero.
- **M-Step:** Estimate the model parameters by maximizing the overall lower bound $\mathcal{L}(\mathbb{D})$, with respect to (α, β, Ω)

$$\mathcal{L}(\mathbb{D})=\sum_{d=1}^{D}\mathcal{L}^d(\gamma^d,\phi^d,\lambda^d;\alpha,\beta,\Omega), \tag{10}$$

where $\mathbb{D}$ is the training set of D images and $\mathcal{L}^d$ is the lower bound computed for training image d.

These two steps are performed until $\mathcal{L}(\mathbb{D})$ converges. The update equations of the variational (γ, ϕ, λ) and model (α, β) parameters are the same as those proposed in [54,68]. The update equations for Ω depend on the distributions $p(r_n|z_k, \Omega_k)$. In [54] these distributions are defined as multivariate Gaussian. However, this kind of distribution is not suitable to model all types of features. An example are features that comprise periodic angular information, such as the Hue channel of the HSV color space. Therefore, two distributions are applied in this work, according to the features r_n used to describe the regions. If μ_{HSV} is included in the feature vector, then $p(r_n|z_k, \Omega_k)$ is a von-Mises multivariate Gaussian distribution [69]

$$p(r_n|z_k,\Omega_k)=\nu(\mathrm{H}_n|\tau_k,\varepsilon_k)\ .\ G(r_n'|\mu_k,\Sigma_k), \tag{11}$$

where G is a multivariate Gaussian, and r_n' defined the feature vector of region n without the feature corresponding to the H channel. ν is a von-Mises distribution

$$\nu(\mathrm{H}_n|z_n,\tau,\varepsilon)=\frac{1}{2\pi I_0(\varepsilon)}e^{\varepsilon\cos(\mathrm{H}_n-\tau)}, \tag{12}$$

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 τ. In this case, Ω_k comprises four parameters $(\mu_k, \Sigma_k, \tau_k, \varepsilon_k)$ that must be estimated. Otherwise, $p(r_n|z_n, \Omega_k)$ is a Gaussian distribution. All update equations can be found in Appendix A.1.

5.3.3. Region and image labeling

corr-LDA has a greedy formulation, where all the labels compete to annotate a region and only one is selected. In dermoscopy, more than one label can be associated with the same region (e.g., Fig. 1, the color labels *dark brown* and *black* share regions with the texture label *pigment network*). To tackle this issue we propose two strategies: (i) train three corr-LDA models, one for each class of labels; and (ii) train a corr-LDA model for color and combine all the structures into a single model.

To annotate the regions of a new image it is necessary to compute the following probability for each of the possible labels w

$$p(w|r_n)\propto\sum_{z_k}q(z_k|\phi_n)p(w|z_k,\beta), \tag{13}$$

where ϕ_n is the topic-related variational parameter of region n and $q(z_k|\phi_n)$ is a multinomial distribution. The label w with the highest probability is then selected.

The image labeling process is based on classification. In the case of the color model, the image is annotated with a color label if following area ratio is above an estimated threshold

$$\delta_c=\frac{A^c_{regions}}{A_{lesion}}, \tag{14}$$

where $A^c_{regions}$ is the total area of the regions annotated with color c and A_{lesion} is the area of the lesion.

In the case of texture and color structures, the idea is to use a set of binary classifiers to predict the label. Each of the classifiers is trained to predict one of the possible labels. The features used by the classifiers are the outputs of corr-LDA: the image label probability

$$p(w|\mathbf{r})\propto\sum_{n=1}^{N}\sum_{z_k}q(z_k|\phi_n)p(w|z_k,\beta), \tag{15}$$

which is computed for all the labels, and the average number of regions per topic $\eta \in \mathbb{R}^K$, where each position k is given by

$$\eta_k=\alpha_k-\gamma_k. \tag{16}$$

The latter descriptor is based on the assumption that the k-th position of variational parameter γ corresponds approximately to the kth position of model parameter α plus the expected number of patch features that were generated by the kth topic [68]. Two classification algorithms are tested in this work: random forests and SVM.

5.4. Lesion diagnosis

The previous sections described the strategy used to obtain a medical description of the lesions associated to specific regions. This section will address the question of how to use the detected medical information to diagnose the lesions as melanoma or benign. Two main problems have to be solved at this stage: (i) How to convert the annotations into an appropriate descriptor that can be used by machine learning algorithms? (ii) How to combine the information of the different corr-LDA models in order to obtain a final diagnosis?

The first problem can be tackled using the outputs of corr-LDA, since these convey medical information. Three descriptors are selected:

- **Present/Absent criteria (i):** Binary vectors $\mathbf{f}_C \in \mathbb{R}^6$, $\mathbf{f}_{TS} \in \mathbb{R}^2$, and $\mathbf{f}_{CS} \in \mathbb{R}^2$. The ith position of each of these vectors is equal to 1 if the image was annotated with the ith label, and 0 otherwise.
- **Label distribution (ii):** Computation of the conditional probabilities $p(w|\mathbf{r})$ (15), which provide an approximation of the distribution of each label in a given lesion.
- **Average number of regions per topics (iii):** Computation of $\eta \in \mathbb{R}^K$ for each of the trained corr-LDA models.

The easiest strategy to combine all of these features would be to concatenate all of them into a single feature vector (early fusion). However, it has been shown that this kind of approach is not appropriate to deal with features associated with different properties of an image [70]. Therefore, a late fusion approach is used to tackle this problem. This framework combines the outputs of different classifiers, each one trained using the information of one of the corr-LDA models, and uses it as input for the last classifier that predicts the final diagnosis [70].

Three classification algorithms are tested as candidates for the first line of classifiers: SVM with RBF kernel (SVM-RBF), random forests, and k-nearest neighbor (kNN).Both SVM-RBF and kNN have been widely used in dermoscopy image analysis, as reported in [6,71]. The choice of the RBF kernel for SVM is also based on the results reported in the literature, where according to [71], more than a half of the CAD systems use this kernel. Moreover, other well-known kernels have some limitations: the linear kernel cannot deal with nonlinearly separable tasks, while the polynomial kernel is less stable and requires the tuning of more hyperparameters. Decision trees are also a popular classification algorithm, mainly because they are interpretable [6]. However, they have a tendency to overfit the training data, which motivated us to use random forests, instead of another tree classifier. All of the aforementioned classifiers provide a score for the lesion in the interval [0, 1] and the lesion is classified as melanoma if the score is greater than 0.5. The scores of the different classifiers are combined using two different strategies: logistic regression (LR) and median rule (MR) [72].

6. Experimental results

6.1. Dataset and evaluation metrics

The experiments were performed using a dataset of 804 images selected from the EDRA atlas [5]. This is a multi-source database acquired at different hospitals. The experimental dataset contains 241 examples of melanoma and 563 examples of benign lesions belonging to the following classes: blue nevi, Clark nevi, Spitz nevi, combined nevi, congenital nevi, and dermal nevi. All of the images were analyzed by several experts during a consensus meeting. Each image is associated with a set of text labels stating which are the observed criteria. The total number of medical annotations is 10 (6 colors, 2 texture structures, and 2 color structures). For computational purposes, an additional label defined as *other structures/no structures* was added to the systems to deal with lesions that: (i) did not exhibit any of the assessed texture and color structures: or (ii) exhibited more clinical criteria besides the ones considered in this work. Labels associated with texture and color structures are available for all the images. However, color labels are only available for a subset of 344 images. Tables 1 and 2 show the number of lesions that are labeled with each criterion.

All of the images were pre-processed in order to remove acquisition artifacts and skin hair as described in [25] and their colors were normalized as proposed in [73]. Manual segmentations were used to separate the lesions from healthy skin. The algorithms used in this work have all been implemented in MATLAB 2015b®.

The detection of the medical criteria was evaluated using two metrics: precision (Pre), recall (Re), and $F1$ score. These metrics are used to compare the global labels provided by the automatic system against those of the experts. Lesion diagnosis is evaluated using sensitivity (SE), specificity (SP) and the area under the curve (AUC) value. All of the aforementioned metrics were computed using stratified 10-fold nested cross validation, where the images were separated into 10 folds, each with approximately the same number of nevi and melanomas. It was also ensured that each fold contained approximately the same number of each type of dermoscopic criteria. From these folds, 9 are kept for training and validation (selection of hyperparameters of both the corr-LDA model and the different classifiers) and the 10th fold is used for testing. The testing process is repeated ten times, each time with a different fold, while the training-validation processes are performed nine times for each testing fold. Each time a different fold is kept out for validation. This process ensures that the choice of the best hyperparameters is independent of the test set.

6.2. Detection of medical criteria - results

Different experiments were conducted to optimize this block. The goal of the experiments was: (i) assess which is the best subset of region features (recall Section 5.2); (ii) select the best classifier to obtain the global labels (SVM or random forests); and (iii) compare the performance of training a corr-LDA model for texture structures and another for color structures against training a model that comprises all the structures.

In each of the experiments the number of topics of the corr-LDA models was tuned $K \in \{50, 70, \ldots, 300\}$, as well as the specific hyperparameters of the classifiers. In the case of random forests we specified the number of trees $T \in \{1, 2, \ldots, 75\}$, while in SVM we tuned the width of the RBF kernel $\rho \in \{2^{-12}, 2^{-5}, \ldots, 2^{12}\}$ and the penalization cost $C \in \{2^{-6}, 2^{-4}, \ldots, 2^{12}\}$ given to the soft margin.

Table 3 shows the performance of the best annotation models, as well as the best configurations. The first three rows show the results obtained after training a model for each type of criteria. Most of the colors were detected with good scores. However, the performance scores for the red and white colors are lower than for the rest. Unfortunately, few examples of these colors are available on the dataset, which justifies the worst results for these two colors. Regarding the texture structures' model (2nd row), it is possible to see that the best results were obtained when all of the features are used (μ_g, σ_g, μ_c, μ_{ca}, μ_M, σ_M, μ_m, σ_m). This leads to RE scores above 80% for both of the criteria using random forests. The color structures' model (3rd row) is the one that achieved the worse scores. This was expected as the number of examples of each of the color structures is smaller than those of the other criteria. Interestingly, the best overall results are obtained when we combine all of the structures in a single model (4th row), leading to significant improvements in the detection of blue-whitish veil and regression areas. Although it is not possible to state which are the optimal number of topics K and trees T that led to these results (recall that we are using nested-cross validation), it is possible to show their tendency across the different test sets using box plots. This information is provided in Fig. 5. Notice that the box plot of corr-LDA (2nd row) includes the results for the color model. The

Table 1
Number of color labels per class of lesion.

Type of Lesion	Color					
	Dark Brown	Light Brown	Blue-Gray	Black	Red	White
Melanoma (#142)	134	115	103	100	17	10
Blue Nevi (#24)	6	11	24	0	0	0
Combined Nevi (#11)	9	9	11	2	1	0
Congenital Nevi (#6)	6	6	2	1	0	0
Dermal Nevi (#28)	24	19	15	3	3	3
Spitz Nevi (#74)	64	33	55	52	6	0
Clark Nevi (#59)	292	273	84	104	12	11
Total (#344)	303	247	226	179	31	15

Table 2
Number of texture and color structures labels per class of lesion.

Type of Lesion	Structure			
	Pigment Network	Dots/Globules	Blue-Whitish Veil	White Regression Areas
Melanoma (#241)	157	153	117	69
Blue Nevi (#27)	1	1	4	0
Combined Nevi (#12)	4	9	11	0
Congenital Nevi (#13)	7	11	0	0
Dermal Nevi (#32)	5	24	2	0
Spitz Nevi (#78)	28	43	35	2
Clark Nevi (#401)	302	251	9	39
Total (#804)	504	492	178	110

Table 3
Detection results and best configurations for the four annotation models - * identifies the results of the model that combines color and texture structures. In **bold** we highlight the best results.

Criteria (#Images)	Precision	Recall	F1	Best Configuration
Blue-Gray(#226)	87.6%± 4.3%	94.2%± 5.5%	90.8%± 3.8%	μ_{HSV}
Dark-Brown(#303)	95.7%± 3.4%	95.7%± 3.3%	95.7%± 2.4%	
Light-Brown (#247)	89.1%± 4.3%	92.7%± 3.5%	90.9%± 2.9%	
Black (#179)	81.5%± 10.6%	88.8%± 7.8%	85.0%± 7.7%	
Red (#31)	79.3%± 14.6%	74.2%± 21.2%	76.7%± 15.5%	
White (#15)	63.6%± 29.5%	93.3%± 16.7%	75.6%± 24.1%	
Pigment Network (#504)	**77.6%± 6.7%**	**88.9% ± 5.8%**	**82.9%± 4.0%**	μ_g, σ_g, μ_c, μ_{ca} μ_M, σ_M, μ_m, σ_m
Dots/Globules (#492)	71.8% ± 7.1%	83.2% ± 5.5%	77.1% ± 5.6%	Random Forests
Blue-Whitish Veil (#178)	75.4% ± 9.6%	68.5% ± 7.6%	71.8% ± 7.1%	μ_{HSV}, μ_c, μ_{ca}
Regression Areas (#110)	60.8% ± 9.9%	51.3% ± 9.8%	55.6% ± 9.7%	Random Forests
Pigment Network* (#504)	78.5% ± 6.8%	86.1% ± 4.5%	82.1% ± 3.8%	μ_c, μ_{ca}, μ_M, σ_M,
Dots/Globules* (#492)	**72.8%** ± **6.2%**	**83.7%** ± **5.1%**	**77.9%** ± **4.7%**	μ_m, σ_m,
Blue-Whitish Veil* (#178)	**82.8%** ± **7.5%**	**68.1%** ± **8.6%**	**74.7%** ± **5.8%**	μ_{HSV},
Regression Areas* (#110)	**63.9%** ± **10.9%**	**58.8%** ± **9.0%**	**61.2%** ± **10.1%**	Random Forests

selection of the optimal hyperparameter for the random forests is straightforward and depends only on the structure to be detected, while the selection of the optimal number of topics is selected as the one that leads to the best combined results for all the criteria considered in the model.

Figs. 6 and 7 show some examples of the output of the criteria detection block. Although we do not have ground-truth segmentation for the criteria (recall that the model was trained using text labels only), the region labeling proposed by the model seems to provide a correct interpretation of the lesion. Nonetheless, it is possible to see that in images (c) and (d) the local labeling of pigment network and dots/globules extend to other regions that clearly are not associated with those structures. This problem was tackled when the four structures are characterized by a single model (see Fig. 8).

Table 4 shows the results obtained by state-of-the-art methods regarding the detection of pigment network [26,74] and blue-whitish veil [51,52]. All of these strategies have been developed and tested using images from the EDRA database [5], which is also used in our work. We also show the results of the pigment network algorithm described in [25] applied to our dataset of images. This is the only case where it is possible to establish a true

Table 4
Comparison of structure detection methods. "*" identifies the results obtained with the combined structures model.

Criteria	Recall	Precision	F1	Method (#Images)
Pigment Network	83.8%	79.2%	81.4%	[25] (#504/804)
	82.3%	82.3%	82.1%	[74] (#275/436)
	86.0%	79.6%	82.7%	[26](#100/220)
	88.9%	77.6%	82.9%	Proposed (#504/804)
	86.1%*	78.5%*	82.1%*	Proposed* (#504/804)
Blue-whitish veil	70.0%	71.0%	70.5%	[51](#173/223)
	68.1%	63.8%	65.9%	[52](#198/855)
	68.5%	75.4%	71.8%	Proposed(#178/804)
	68.1%*	82.8%*	77.9%*	Proposed*(#178/804)

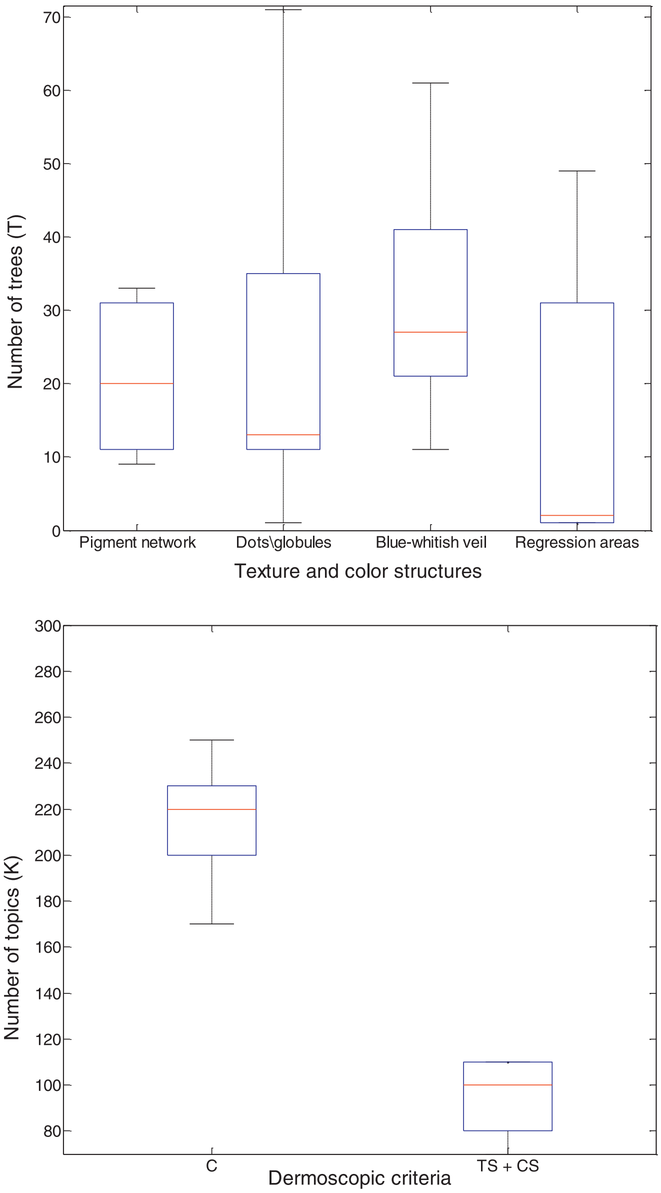
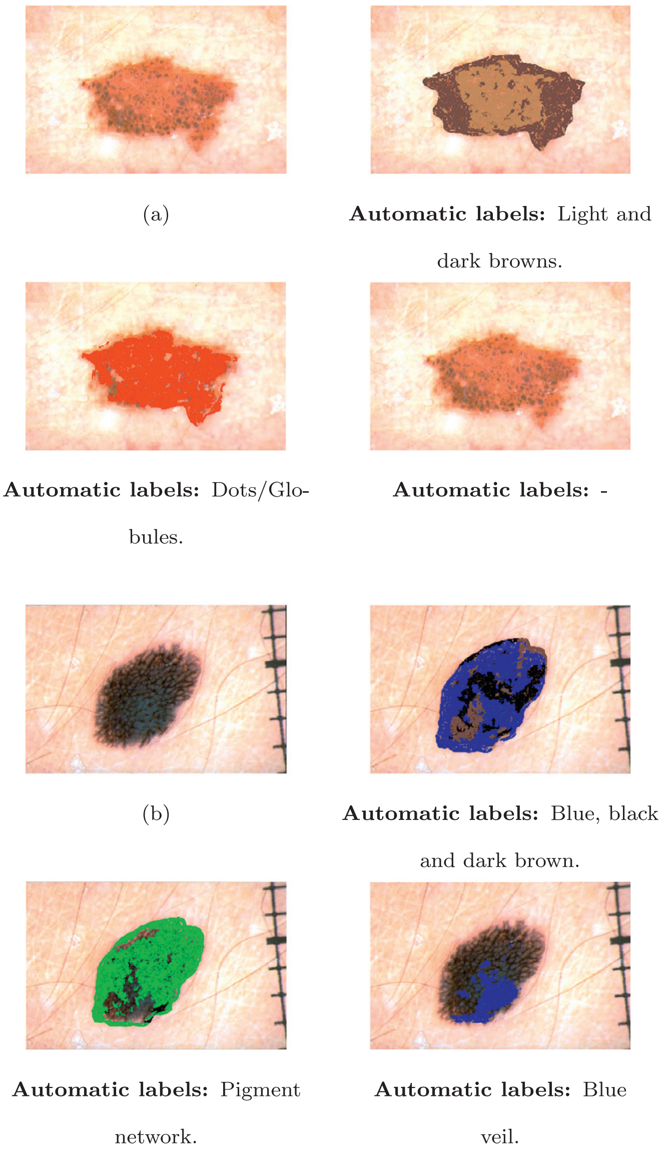


Fig. 5. Box plots showing the range of the selected optimal hyperparameters: the number of trees T of random forests, regarding the detection of the four structures (1st row); and the number of topics K of corr-LDA, regarding the color and all structures models (2nd row).



(a)

Automatic labels: Light and dark browns.

Automatic labels: Dots/Globules.

Automatic labels: -

(b)

Automatic labels: Blue, black and dark brown.

Automatic labels: Pigment network.

Automatic labels: Blue veil.

Fig. 6. Original image, region and image annotation obtained with the color, texture structures, and color structures models (from top left to bottom right). The color scheme is green for pigment network, red for dots/globules, blue for blue whitish veil, yellow for regression areas. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

comparison between the two methods, because the training and test sets are exactly the same. Nonetheless, comparing our results with those of [26,51,52,74] still provides us with relevant information and allows us to check if our results are similar to those obtained by other methods. Our method compares favorably with all of the aforementioned approaches. Particularly interesting is the comparison between the proposed method and the one described in [51], since the later uses ground truth segmentations to train the model. The proposed system is also able to detect multiple structures, while the state-of-the-art methods only focus on one structure.

6.3. *Lesion diagnosis - results*

Three experiments were performed on this stage: (i) assess the performance of the different medical criteria and their combination; (ii) compare the diagnostic accuracy of SVM-RBF, random forests, and kNN; and (iii) compare two late fusion strategies (LR and MR).

All of the experiments were carried on using the outputs of the best corr-LDA models. The number of trees for the random forests algorithm was set to be $T \in \{1, 2, \ldots, 200\}$, the hyperparameters of SVM-RBF are set to be $\rho \in \{2^{-12}, 2^{-5}, \ldots, 2^{12}\}$ and $C \in \{2^{-6}, 2^{-4}, \ldots, 2^{12}\}$, and the number U of neighbors of kNN is searched in the set $U \in \{1, 3, \ldots, 101\}$.

Table 5 shows the best results for the experiments with the different criteria and their combinations, Fig. 9 shows the ROC curves of the different classifiers per type of criteria, as well as the results obtained. As expected, the combination of all the criteria improves the scores. The random forests and kNN algorithms are the ones that achieve the best and worse classification scores, respectively. The low sensitivity values achieved by kNN might be justified by the unbalance between the number of melanomas and benign lesions. Due to the characteristics of late fusion, it was also possible

Table 5
Results for melanoma diagnosis using SVM-RBF, random forests (RF), and kNN. "All*" refers to the combination of the colors with color + texture structures model. **Bold** highlights the best results for each classifier.

Criteria	SVM	RF	kNN
Colors	$SE = 72.6\% \pm 9.6\%$ $SP = 71.4\% \pm 7.6\%$ $AUC = 75.9\% \pm 7.3\%$	$SE = 84.6\% \pm 5.5\%$ $SP = 61.1\% \pm 7.7\%$ $AUC = 78.6\% \pm 4.4\%$	$SE = 54.7\% \pm 6.6\%$ $SP = 85.1\% \pm 4.6\%$ $AUC = 74.9\% \pm 6.2\%$
Texture Structures	$SE = 87.1\% \pm 6.7\%$ $SP = 60.0\% \pm 8.3\%$ $AUC = 78.4\% \pm 7.1\%$	$SE = 75.1\% \pm 9.1\%$ $SP = 75.1\% \pm 7.6\%$ $AUC = 79.7\% \pm 7.5\%$	$SE = 58.9\% \pm 10.0\%$ $SP = 79.6\% \pm 7.4\%$ $AUC = 72.5\% \pm 4.8\%$
Color Structures	$SE = 83.4\% \pm 9.0\%$ $SP = 70.9\% \pm 5.8\%$ $AUC = 81.5\% \pm 5.8\%$	$SE = 82.1\% \pm 3.5\%$ $SP = 72.5\% \pm 5.5\%$ $AUC = 82.1\% \pm 4.9\%$	$SE = 60.1\% \pm 7.8\%$ $SP = 82.9\% \pm 4.5\%$ $AUC = 76.7\% \pm 6.9\%$
Color Structures & Texture Structures	$SE = 90.9\% \pm 4.3\%$ $SP = 56.0\% \pm 6.6\%$ $AUC = 79.8\% \pm 4.4\%$	$SE = 80.9\% \pm 8.9\%$ $SP = 74.8\% \pm 6.6\%$ $AUC = 82.8\% \pm 5.8\%$	$SE = 70.1\% \pm 8.5\%$ $SP = 77.3\% \pm 6.1\%$ $AUC = 76.0\% \pm 6.3\%$
All	$\mathbf{SE = 84.2\% \pm 6.8\%}$ $\mathbf{SP = 72.1\% \pm 5.6\%}$ $\mathbf{AUC = 85.1\% \pm 4.3\%}$	$\mathbf{SE = 81.3\% \pm 5.3\%}$ $\mathbf{SP = 74.8\% \pm 6.5\%}$ $\mathbf{AUC = 85.4\% \pm 4.2\%}$	$\mathbf{SE = 55.6\% \pm 7.8\%}$ $\mathbf{SP = 86.1\% \pm 4.6\%}$ $\mathbf{AUC = 79.4\% \pm 5.4\%}$
All*	$\mathbf{SE = 85.4\% \pm 7.9\%}$ $\mathbf{SP = 65.4\% \pm 6.1\%}$ $\mathbf{AUC = 83.8\% \pm 6.0\%}$	$\mathbf{SE = 82.3\% \pm 7.7\%}$ $\mathbf{SP = 73.2\% \pm 8.3\%}$ $\mathbf{AUC = 84.3\% \pm 4.9\%}$	$\mathbf{SE = 57.6\% \pm 9.1\%}$ $\mathbf{SP = 81.7\% \pm 6.4\%}$ $\mathbf{AUC = 79.3\% \pm 6.4\%}$

Table 6
Results for melanoma diagnosis using criteria and classifier fusion. "All*" refers to the combination of the colors with color + texture structures model. **Bold** highlights the best results.

Criteria	SVM + kNN	RF+kNN	SVM + RF	SVM + RF + kNN
Colors	$SE = 62.6\% \pm 9.6\%$ $SP = 80.8\% \pm 6.8\%$ $AUC = 77.8\% \pm 6.5\%$	$SE = 67.6\% \pm 7.4\%$ $SP = 75.5\% \pm 5.5\%$ $AUC = 79.7\% \pm 4.8\%$	$SE = 76.7\% \pm 8.7\%$ $SP = 66.8\% \pm 6.9\%$ $AUC = 80.7\% \pm 5.5\%$	$SE = 70.9\% \pm 8.9\%$ $SP = 76.9\% \pm 5.4\%$ $AUC = 80.2\% \pm 5.5\%$
Texture Structures	$SE = 64.3\% \pm 9.8\%$ $SP = 77.3\% \pm 6.5\%$ $AUC = 77.6\% \pm 5.4\%$	$SE = 66.4\% \pm 8.1\%$ $SP = 76.4\% \pm 7.8\%$ $AUC = 78.2\% \pm 6.5\%$	$SE = 81.7\% \pm 6.9\%$ $SP = 66.1\% \pm 7.5\%$ $AUC = 81.1\% \pm 7.0\%$	$SE = 69.7\% \pm 9.5\%$ $SP = 76.2\% \pm 6.2\%$ $AUC = 79.7\% \pm 6.3\%$
Color Structures	$SE = 70.1\% \pm 9.3\%$ $SP = 78.7\% \pm 5.3\%$ $AUC = 81.9\% \pm 6.3\%$	$SE = 68.0\% \pm 9.1\%$ $SP = 80.5\% \pm 6.5\%$ $AUC = 82.0\% \pm 5.7\%$	$SE = 84.6\% \pm 5.3\%$ $SP = 73.4\% \pm 5.2\%$ $AUC = 85.0\% \pm 5.3\%$	$SE = 78.4\% \pm 6.5\%$ $SP = 78.9\% \pm 7.4\%$ $AUC = 84.2\% \pm 5.5\%$
Color Structures & Texture Structures	$SE = 76.3\% \pm 9.3\%$ $SP = 71.1\% \pm 7.7\%$ $AUC = 80.9\% \pm 5.8\%$	$SE = 71.8\% \pm 6.5\%$ $SP = 77.3\% \pm 6.6\%$ $AUC = 82.0\% \pm 5.4\%$	$SE = 83.4\% \pm 9.4\%$ $SP = 68.1\% \pm 8.4\%$ $AUC = 83.8\% \pm 4.8\%$	$SE = 79.2\% \pm 8.6\%$ $SP = 72.7\% \pm 7.3\%$ $AUC = 83.3\% \pm 5.0\%$
All	$SE = 67.6\% \pm 7.7\%$ $SP = 83.8\% \pm 5.4\%$ $AUC = 83.6\% \pm 4.6\%$	$SE = 68.4\% \pm 11.1\%$ $SP = 79.6\% \pm 6.4\%$ $AUC = 84.2\% \pm 4.6\%$	$SE = \mathbf{84.6}\% \pm \mathbf{6.8}\%$ $SP = \mathbf{74.2}\% \pm \mathbf{6.9}\%$ $AUC = \mathbf{86.9}\% \pm \mathbf{4.1}\%$	$SE = 75.9\% \pm 7.8\%$ $SP = 79.4\% \pm 5.7\%$ $AUC = 85.8\% \pm 4.4\%$
All*	$SE = 69.2\% \pm 11.4\%$ $SP = 80.5\% \pm 6.2\%$ $AUC = 82.6\% \pm 6.2\%$	$SE = 71.7\% \pm 7.9\%$ $SP = 76.9\% \pm 5.3\%$ $AUC = 83.2\% \pm 4.5\%$	$SE = \mathbf{85.8}\% \pm \mathbf{8.2}\%$ $SP = \mathbf{71.1}\% \pm \mathbf{9.2}\%$ $AUC = \mathbf{85.9}\% \pm \mathbf{4.5}\%$	$SE = 76.7\% \pm 8.5\%$ $SP = 76.2\% \pm 6.6\%$ $AUC = 84.2\% \pm 5.0\%$

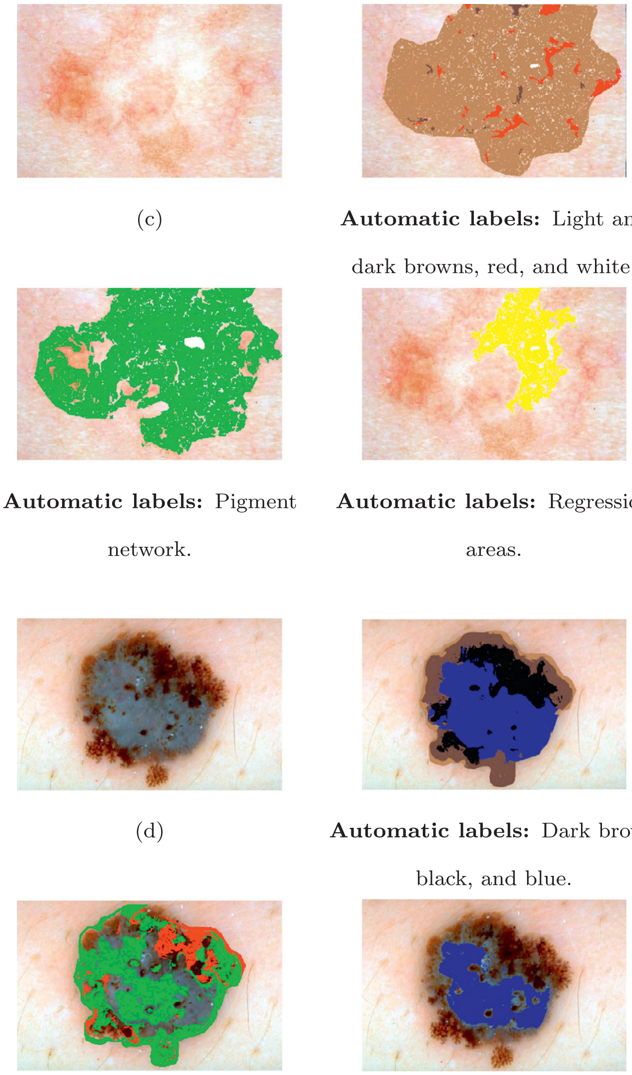
to combine the outputs of the different classifier [70]. These fusion results can be seen in Table 6 and the ROC curves for the combination of criteria and classifiers can be seen in Fig. 10. The best classification scores are obtained by combining SVM and random forests. The results in both tables and figures were achieved using MR as the fusion strategy. Although this method outperformed LR in all the experiments, the latter also led to reasonable classification scores: $SE = 79.2\%$, $SP = 75.6\%$, and $AUC = 83.6\%$. However the results of MR ($SE = 84.6\%$, $SP = 74.2\%$ and $SE = 85.8\%$, $SP = 71.1\%$), are clearly better. It should be stressed that these results were obtained using a challenging multi-source dataset (acquired at three different facilities). The range of the optimal hyperparameters values of the classifiers that achieved the best classification scores can be seen in Figs. 11 and 12.

Table 7 shows the comparison of the melanoma diagnosis results obtained in this work against the ones reported in [75], where abstract image features were combined in order to diagnose melanomas. The proposed method compares favorably, suggesting that it is possible to replace the more traditional pattern recognition features by ones that can be associated with medical information, without losing classification power.

Table 7
Comparison of melanoma diagnosis methods.

Dataset	Sensitivity	Specificity	Method
EDRA	83.0%	76.0%	[75]
	84.6%	74.2%	Proposed
PH^2	98.0%	90.0%	[75]
	100.0%	88.2%	Proposed

We have also evaluated the generalization power of our method using the publicly available PH^2 dataset that contains 200 images (40 melanomas) [76]. To be able to apply our methodology to this new dataset, we first trained two corr-LDA models (one for colors and the other for color+texture structures) using all of the EDRA images. The number of topics K of each model was selected



(c)

Automatic labels: Light and dark browns, red, and white.

Automatic labels: Pigment network.

Automatic labels: Regression areas.

(d)

Automatic labels: Dark brown black, and blue.

Automatic labels: Pigment network and dots/globules.

Automatic labels: Blue veil.

Fig. 7. Original image, region and image annotation obtained with the color, texture structures, and color structures models (from top left to bottom right). The color scheme is green for pigment network, red for dots/globules, blue for blue whitish veil, yellow for regression areas. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

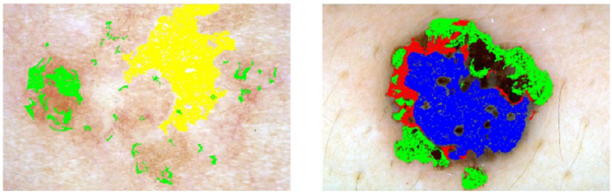


Fig. 8. Output of the corr-LDA that combines color and texture structures for lesions (c) and (d). The color scheme is green for pigment network, red for dots/globules, blue for blue whitish veil, yellow for regression areas. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 8
Computational times for the best CAD system.

Step	Time (seconds)
Lesion segmentation + region representation	10.2s
Detection of medical criteria	0.58s
Diagnosis	0.0273s

based on the information provided by the box-plots in Fig. 5, *i.e.*, we selected the median number of topics of each model (the red line). This means that a color model was trained using $K = 220$ and the color+texture structures model was trained using $K = 110$. Both of these models were then applied to the PH^2 images, in order to extract the medical features. Finally, the images were classified as melanoma or benign using the MR fusion of SVM and random forests. The classifiers were trained and tested using a 10-fold nested cross validation method and exactly the same folds as the ones used in [75], which allowed us to establish a fair comparison with the results obtained using abstract features. In order to deal with any color changes introduced by different acquisition setups, all of the PH^2 images were pre-processed as described in Section 6.1. The obtained results can be seen in Table 7. These are fairly similar, with a slight increase in the sensitivity and decrease in specificity, showing us that the learned models can be applied to other images to extract relevant information. Moreover, this reinforces the idea that it is possible to extract clinically relevant information, without losing classification power.

Table 8 shows the average computational time for dermoscopy images of size 768 × 512, using the best CAD system configuration. These results show that the algorithm is able to diagnose the image in less than a minute and that the bottleneck of the system are the steps associated with the lesion segmentation and region representation (respectively described in Sections 5.1 and 5.2.)

7. Conclusions

This paper discusses the development of a system for skin lesion diagnosis that is inspired by clinical practice, and tries to mimic the different steps of medical analysis. It starts by identifying regions in the lesion that have similar color and texture properties. These regions are segmented and annotated with one or more medical criteria. The experimental results are promising and show that the proposed framework can be used to identify colors, pigment network, dots/globules, blue-whitish veil, and white regression areas. Finally, all of the detected criteria are combined in order to obtain a diagnosis, achieving a sensitivity of 84.6% and a specificity of 74.2%.

The system presented in this work is different from any other found in literature. First, it uses an image annotation model learned from text labels, to identify multiple dermoscopic criteria and obtain a medical description of the lesion. Second it is capable of combining information from multiple types of medical criteria to obtain a diagnosis. The proposed system compares favorably with a CAD system that uses abstract image features, which is an evidence that it is possible to develop systems that use features with medical meaning, without compromising the classification performance. Moreover, it was also shown that the learned models can be applied to different datasets. We hope that the proposed methodology opens a new direction of progress in the analysis of dermoscopy images. Future directions for this work include: (i) extension to the analysis of non-melanocytic lesions; (ii) evaluation on a real world scenario, *i.e*, during routine clinical practice; (iii) inclusion of other features with medical meaning, such as shape features; and (iv) allowing medical

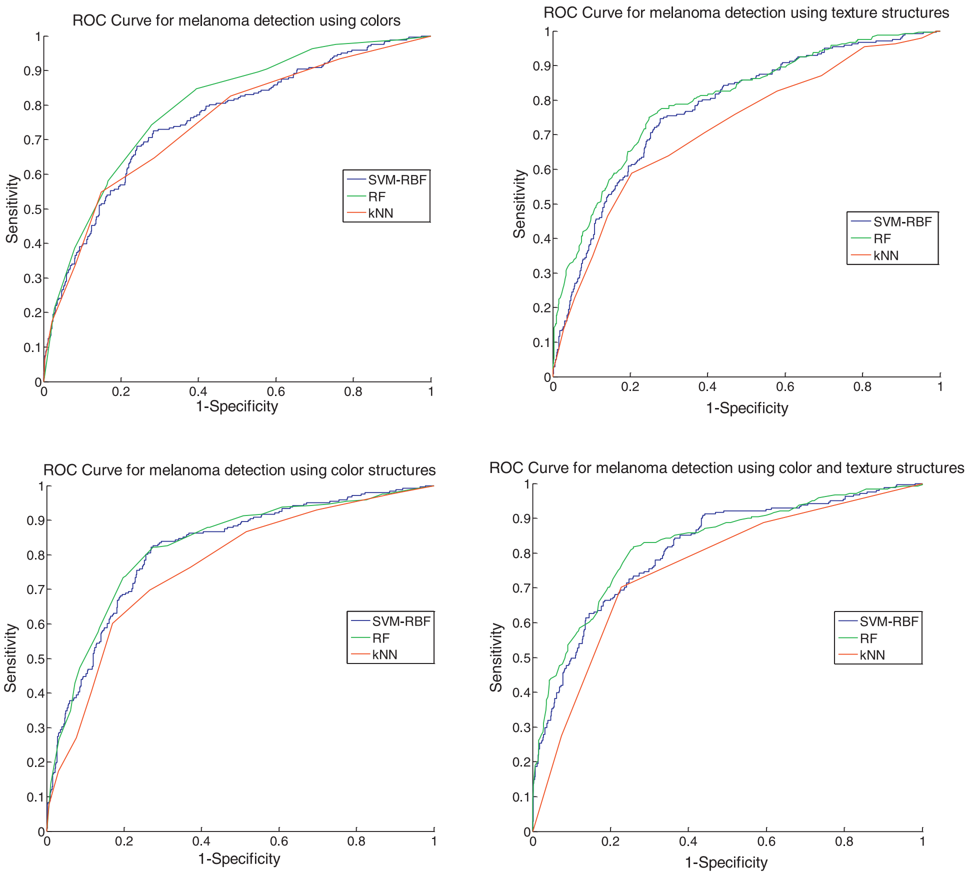


Fig. 9. ROC curves for SVM-RBF, random forests (RF), and kNN in the task of melanoma diagnosis using each type of criteria and the combination of color and texture structures.

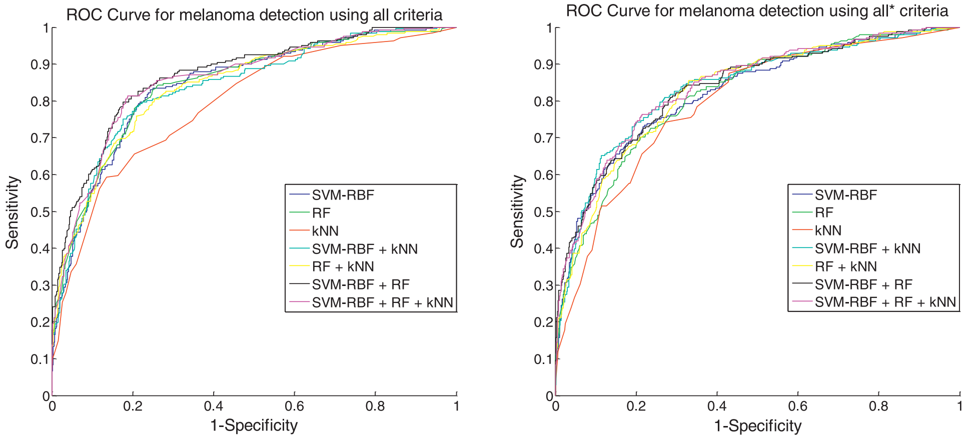


Fig. 10. ROC curves for SVM-RBF, random forests (RF), kNN, and their fusion in the task of melanoma diagnosis using the all the medical criteria: "all*" refers to the combination of the colors with color + texture structures model.

feedback, making it possible for the system to learn from a wrong decision.

Acknowledgments

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

Appendix A

A1. Variational EM - update equations

•E-Step

The update equations for the variational parameters (γ^d, ϕ^d, λ^d) are the following

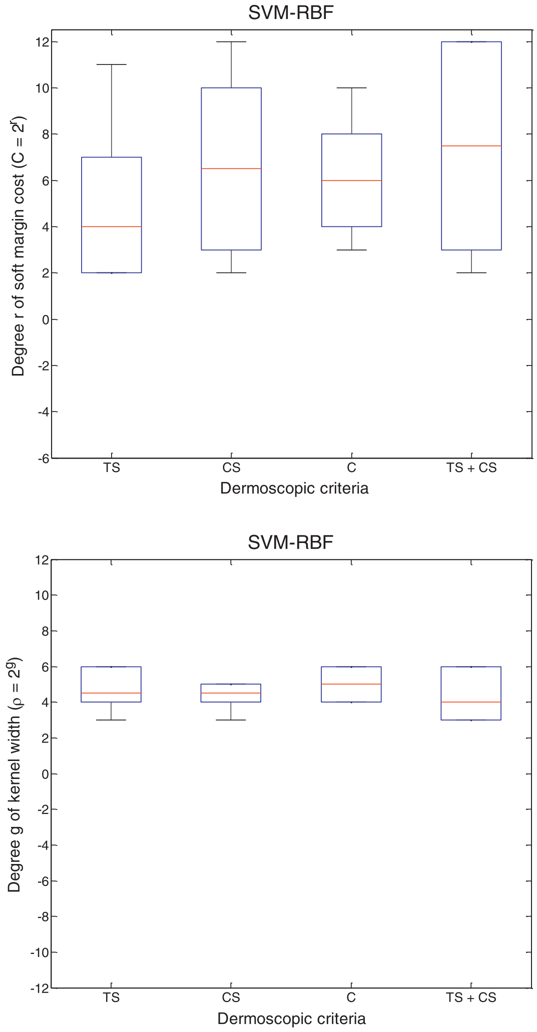


Fig. 11. Box plots showing the range of the selected hyperparameters for SVM (C - 1st and ρ 2nd row).

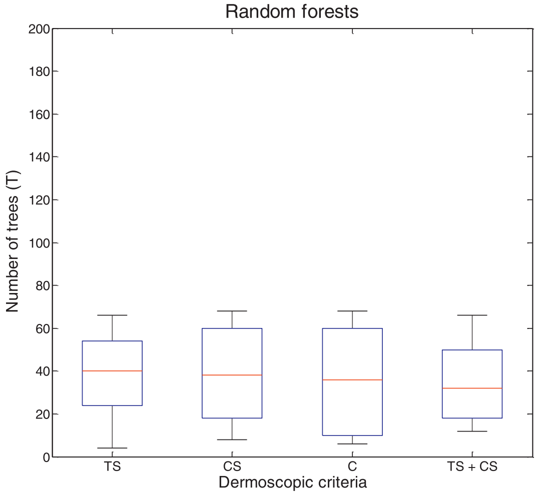


Fig. 12. Box plots showing the range of the selected hyperparameters random forests (T - 1st row) in the task of melanoma diagnosis.

$$\phi^d_{nk} \propto p(r^d_n|z_n=k,\Omega)\exp\{E_q[\log q(\theta_k|\gamma^d)]\}.$$

$$\cdot \ \exp\left\{\sum_{m=1}^{M}\lambda^d_{mn}\log p(w^d_m|y_m=n,z_m=i,\beta)\right\} \quad (A.1)$$

$$\lambda^d_{mn} \propto \exp\left\{\sum_{k=1}^{K}\phi^d_{nk}\log p(w^d_m|y_m=n,z_m=i,\beta)\right\}, \quad (A.2)$$

$$\gamma^d_k = \alpha_k + \sum_{n=1}^{N_d}\phi^d_{nk}. \quad (A.3)$$

These parameters must be estimated by the order that they are presented here.

•M-Step The parameter β that relates the text labels w_m with the topic k is updated as follows

$$\beta_{km} \propto \sum_{d=1}^{D} w^d_m \sum_{n=1}^{N_d}\phi^d_{nk}\lambda^d_{mn}. \quad (A.4)$$

It is not possible to obtain an exact update equation for the Dirichlet parameter α. Therefore, Blei and Jordan propose the use of the Newton–Raphon's method [68] to obtain an estimate of this parameter.

When the traditional formulation of corr-LDA is used, each of the k multivariate parameters $\Omega_k = (\mu_k, \Sigma_k)$ is computed as follows:

$$\mu_k = \frac{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi^d_{nk}r^d_n}{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi^d_{nk}}, \quad (A.5)$$

$$\Sigma_k = \frac{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi^d_{nk}(r^d_n-\mu_k)(r^d_n-\mu_k)^T}{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi^d_{nk}}. \quad (A.6)$$

When the von Mises–Gaussian distributions are used to model the regions' features the update equations for $\Omega_k = (\mu_k, \Sigma_k, \tau_k, \varepsilon_k)$ are as follows. The parameters μ and Σ are update as in (A.5) and (A.6), but using r'_n (feature vector without the H channel information). The remaining parameters are updated using the following equations

$$\tau_k = \tan^{-1}\left(\frac{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi^d_{nk}\sin \mathrm{H}^d_n}{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi^d_{nk}\cos \mathrm{H}^d_n}\right), \quad (A.7)$$

An analytical computation of the parameter ε_k is not possible. Different approximations have been proposed to tackle this issue. This work uses the approach described in [77], which makes use of the Newton–Raphson's method to obtain an approximation. This method requires a few iterations t of the following equation:

$$\varepsilon^t_k = \varepsilon^{t-1}_k - \frac{A(\varepsilon^{t-1}_k) - \bar{R}}{1 - A(\varepsilon^{t-1}_k)}, \quad (A.8)$$

where,

$$A(\varepsilon^{t-1}_k) = \frac{I_1(\varepsilon^{t-1}_k)}{I_0(\varepsilon^{t-1}_k)}, \quad (A.9)$$

and the variable $\bar{R}$ is defined as

$$\bar{R} = \frac{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi^d_{nk}\cos([\mathrm{H}]^d_n - \tau_k)}{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi^d_{nk}}. \quad (A.10)$$

In the first iteration ε^0_k is set to be [77]

$$\varepsilon^0_k = \frac{\bar{R} - \bar{R}^3}{1 - \bar{R}^2}. \quad (A.11)$$

The update equation is applied until convergence is reached.

References

[1] 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.

[2] Cancer facts and figures 2016, American Cancer Society, 2016.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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. Imaging Graph. 31 (6) (2007) 362–373.

[8] 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. Imaging Graph. 32 (7) (2008) 566–579.

[9] 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.

[10] M.E. Celebi, T. Mendonça, J.S. Marques, Dermoscopy Image Analysis, 10, CRC Press, 2015.

[11] H. Pehamberger, A. Steiner, K. Wolff, In vivo epiluminescence microscopy of pigmented skin lesions. I. Pattern analysis of pigmented skin lesions, J. Am. Acad. Dermatol. 17 (4) (1987) 571–583.

[12] H. Iyatomi, H. Oka, M.E. Celebi, K. Ogawa, G. Argenziano, H.P. Soyer, H. Koga, T. Saida, K. Ohara, M. Tanaka, Computer-based classification of dermoscopy images of melanocytic lesions on acral volar skin, J.Invest. Dermatol. 128 (8) (2008) 2049–2054.

[13] S. Yang, B. Oh, S. Hahm, K.Y. Chung, B.U. Lee, Ridge and furrow pattern classification for acral lentiginous melanoma using dermoscopic images, Biomed Signal Process Control (2016).

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

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

[16] M. Sadeghi, T.K. Lee, D. McLean, H. Lui, M.S. Atkins, Global pattern analysis and classification of dermoscopic images using textons, SPIE Medical Imaging, International Society for Optics and Photonics, 2012. 83144X–83144X

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

[18] M. Fleming, C. Steger, J. Zhang, J. Gao, A. Cognetta, I. Pollak, C. Dyer, Techniques for a structural analysis of dermatoscopic imagery, Comput. Med. Imaging Graph. (5) (1998) 375–389.

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

[20] C. Grana, R. Cucchiara, G. Pellacani, S. Seidenari, Line detection and texture characterization of network patterns., in: In ICPR'06: Proceedings of the 18th International Conference on Pattern Recognition, 2006.

[21] 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: 2006 IEEE International Workshop on Medical Measurement and Applications (MeMea), IEEE, 2006, pp. 63–67.

[22] G. Di Leo, A. Paolillo, P. Sommella, G. Fabbrocini, Automatic diagnosis of melanoma: a software system based on the 7-point check-list, in: 2010 43rd Hawaii International Conference on System Sciences (HICSS), IEEE, 2010, pp. 1–10.

[23] 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. Imaging Graph. 35 (2) (2011) 137–143.

[24] P. Wighton, T.K. Lee, H. Lui, D.I. McLean, M.S. Atkins, Generalizing common tasks in automated skin lesion diagnosis, IEEE Trans. Inf. Technol. Biomed. 15 (4) (2011) 622–629.

[25] 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.

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

[27] M. Machado, J. Pereira, R.F. Pinto, Classification of reticular pattern and streaks in dermoscopic images based on texture analysis, J. Med. Imaging 2 (4) (2015).

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

[29] 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. Imaging 32 (2013) 849–861.

[30] K. Delibasis, K. Kottari, I. Maglogiannis, Automated detection of streaks in dermoscopy images, in: Artificial Intelligence Applications and Innovations, Springer, 2015, pp. 45–60.

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

[32] I. Maglogiannis, K.K. Delibasis, Enhancing classification accuracy utilizing globules and dots features in digital dermoscopy, Comput. Methods Programs Biomed. 118 (2) (2015) 124–133.

[33] A. Madooei, M. Drew, Incorporating colour information for computer-aided diagnosis of melanoma from dermoscopy images: a retrospective survey and critical analysis, Int. J. Biomed. Imaging 2016 (2016).

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

[35] A. Sboner, P. Bauer, G. Zumiani, C. Eccher, E. Blanzieri, S. Forti, M. Cristofolini, Clinical validation of an automated system for supporting the early diagnosis of melanoma, Skin Res. Technol. 10 (3) (2004) 184–192.

[36] G. Pellacani, C. Grana, S. Seidenari, Automated description of colours in polarized-light surface microscopy images of melanocytic lesions, Melanoma Res. 14 (2) (2004) 125–130.

[37] 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: Computational Modelling of Objects Represented in Images - CompImage 2012, 2012, pp. 103–108.

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

[39] 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, Computer. Med. Imaging Graph. 38 (5) (2014) 403–410.

[40] 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.

[41] 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.

[42] A. Murali, W. Stoecker, R. Moss, Detection of solid pigment in dermatoscopy images using texture analysis, Skin Res. Technol. 6 (4) (2000) 193–198.

[43] W. Stoecker, K. Gupta, R. Stanley, R. Moss, B. Shrestha, Detection of asymmetric blotches (asymmetric structureless areas) in dermoscopy images of malignant melanoma using relative color, Skin Res. Technol. 11 (3) (2005) 179–184.

[44] V.K. Madasu, B. Lovell, Blotch detection in pigmented skin lesions using fuzzy co-clustering and texture segmentation, in: Digital Image Computing: Techniques and Applications (DICTA), IEEE, 2009, pp. 25–31.

[45] A. Dalal, R. Moss, R. Stanley, W. Stoecker, K. Gupta, D. Calcara, J. Xu, B. Shrestha, R. Drugge, J. Malters, Concentric decile segmentation of white and hypopigmented areas in dermoscopy images of skin lesions allows discrimination of malignant melanoma, Comput. Med. Imaging Graph. 35 (2) (2011) 148–154.

[46] J.L.G. Arroyo, B.G. Zapirain, Hypopigmentation pattern recognition in dermoscopy images for melanoma detection, J. Med. Imaging Health Inform. 5 (8) (2015) 1875–1879.

[47] S. Kaya, M. Bayraktar, S. Kockara, M. Mete, T. Halic, H.E. Field, H.K. Wong, Abrupt skin lesion border cutoff measurement for malignancy detection in dermoscopy images, BMC Bioinform. 17 (13) (2016) 367.

[48] 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: Systems, Signals and Devices, 2009. SSD'09. 6th International Multi-Conference on, IEEE, 2009, pp. 1–6.

[49] W. Stoecker, M. Wronkiewiecz, R. Chowdhury, R. Stanley, J. Xu, A. Bangert, B. Shrestha, D. Calcara, H. Rabinovitz, M. Oliviero, Detection of granularity in dermoscopy images of malignant melanoma using color and texture features, Comput. Med. Imaging Graph. 35 (2) (2011) 144–147.

[50] 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. Imaging Graph. 32 (2008) 670–677.

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

[52] A. Madooei, M. Drew, H. Hajimirsadeghi, Learning to detect blue-white structures in dermoscopy images with weak supervision, CoRR abs/1506.09179 (2015).

[53] C. Barata, M.E. Celebi, J.S. Marques, J. Rozeira, Clinically inspired analysis of dermoscopy images using a generative model, Comput. Vision Image Understanding 151 (2016) 124–137.

[54] D. Blei, M. Jordan, Modeling annotated data, in: 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2003, pp. 127–134.
[55] P.F. Felzenszwalb, D.P. Huttenlocher, Efficient graph-based image segmentation, Int. J. Comput. Vision 59 (2) (2004) 167–181.
[56] C. Carson, S. Belongie, H. Greenspan, J. Malik, Blobworld: image segmentation using expectation-maximization and its application to image querying, IEEE Trans. Pattern Anal. Mach. Intell. 24 (8) (2002) 1026–1038.
[57] D. Zhang, M.M. Islam, G. Lu, A review on automatic image annotation techniques, Pattern Recognit. 45 (1) (2012) 346–362.
[58] M.L. Zhang, Z.H. Zhou, A review on multi-label learning algorithms, IEEE Trans. Knowl. Data Eng. 26 (8) (2014) 1819–1837.
[59] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1627–1645.
[60] Z.H. Zhou, M.L. Zhang, S.J. Huang, Y.F. Li, Multi-instance multi-label learning, Artif. Intell. 176 (1) (2012) 2291–2320.
[61] R. Hong, M. Wang, Y. Gao, D. Tao, X. Li, X. Wu, Image annotation by multiple-instance learning with discriminative feature mapping and selection, IEEE Trans. Cybern. 44 (5) (2014) 669–680.
[62] R. Cabral, F. De la Torre, J.P. Costeira, A. Bernardino, Matrix completion for weakly-supervised multi-label image classification, IEEE Trans. Pattern Anal. Mach. Intell. 37 (1) (2015) 121–135.
[63] V. Lavrenko, R. Manmatha, J. Jeon, A model for learning the semantics of pictures, in: Advances in Neural Information Processing Systems, 2003, p. None.
[64] S.L. Feng, R. Manmatha, V. Lavrenko, Multiple Bernoulli relevance models for image and video annotation, in: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2, IEEE, 2004, pp. II–1002.
[65] Y. Mori, H. Takahashi, R. Oka, Image-to-word transformation based on dividing and vector quantizing images with words, in: First International Workshop on Multimedia Intelligent Storage and Retrieval Management, Citeseer, 1999, pp. 1–9.
[66] K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. Blei, M. Jordan, Matching words and pictures, J. Mach. Learn. Res. 3 (2003) 1107–1135.
[67] P. Duygulu, K. Barnard, J. de Freitas, D. Forsyth, Object recognition as machine translation: learning a lexicon for a fixed image vocabulary, in: Computer VisionECCV, Springer, 2002, pp. 97–112.
[68] D. Blei, A. Ng, M. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[69] 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.
[70] L. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms, Wiley, 2004.
[71] R. Oliveira, J. Papa, A. Pereira, J. Tavares, Computational methods for pigmented skin lesion classification in images: review and future trends, Neural Compu. Appl. (2016) 1–24.
[72] J. Kittler, M. Hatef, R.W. Duin, J. Matas, On combining classifiers, IEEE Trans. Pattern Anal. Mach. Intell. 20 (3) (1998) 226–239.
[73] C. Barata, M.E. Celebi, J.S. Marques, Improving dermoscopy image classification using color constancy, IEEE J. Biomed. Health Inform. 19 (3) (2015) 1146–1152.
[74] M. Sadeghi, M. Razmara, P. Wighton, T.K. Lee, M.S. Atkins, Modeling the dermoscopic structure pigment network using a clinically inspired feature set, in: International Workshop on Medical Imaging and Virtual Reality, Springer, 2010, pp. 467–474.
[75] C. Barata, M. Celebi, J. Marques, Melanoma detection algorithm based on feature fusion, in: 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2015, pp. 2653–2656.
[76] 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: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2013, pp. 5437–5440.
[77] 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.

Catarina Barata received B.Sc. and M.Sc. degrees in Biomedical Engineering, and Ph.D. degree in Electrical and Computer Engineering from Instituto Superior Técnico, University of Lisbon, in 2009, 2011, and 2017 respectively. Her interests include medical imaging, image analysis, pattern recognition, and machine learning.

M. Emre Celebi received his B.Sc. degree in Computer Engineering from METU (Turkey). He received his M.Sc./Ph.D. degree in Computer Science & Engineering from the University of Texas at Arlington. He is currently a Professor and the Chair of the Department of Computer Science at the University of Central Arkansas.

Jorge S. Marques received E.E., M.Sc., and Ph.D. degrees in electrical and computer engineering from Instituto Superior Técnico, University of Lisbon, where he is Associate Professor. His interests include image analysis, machine learning, and statistical pattern recognition.