Detection of Dermoscopic Structures with Supervised and Weakly Supervised Learning

Bárbara Filipa Ferreira Cardoso
barbara.ferreira.cardoso@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

October 2018

Abstract

Melanoma is the deadliest type of skin cancer, due to its high potential to metastasize. However, when detected at an early stage it has a high cure rate. Dermoscopy is an image acquisition technique used by dermatologists to observe and diagnose skin lesions. This technique allows the observation of dermoscopic structures. These structures are then used in several diagnostic procedures, such as the ABCD rule and the 7-point list. The recent development and public availability of numerous databases of detailed dermatoscopic images and medical annotations have relaunched the development of automated diagnostic systems for skin lesions based on clinical criteria. These systems present a diagnosis of the lesion and return a medical justification, which can be understood by the specialists, using the dermatoscopic structures detected. In this thesis there are presented three different automatic methods to detect and locate four dermatoscopic structures (pigment network, milia-like cysts, negative pigment network and streaks), using the ISIC2017 database: two supervised methods, developed based on local medical annotations (SVM and ANN); and a weakly supervised method that only uses global image annotations (Corr-LDA). The results obtained suggest that the supervised methods present promising results, even in the detection of rare structures. Using the weakly supervised method, the results obtained for pigment network are as good as the ones obtained with supervised methods, which emphasize the high potential of the Corr-LDA. But the algorithm was not able to detect the remaining structures.

Keywords: Dermosocpy, Melanoma, Dermoscopic Structures, Supervised Annotation Methods, Weakly Supervised Annotation Methods, SVM, ANN, Corr-LDA

1. Introduction

Melanoma is the deadliest type of skin cancer, due to its high potential to metastasize in later stages, propagating to to other parts of human body. However, when melanoma is diagnosed at an early stage there are significant improvements in the survival probability [1]. The reason why this happened is that, at an early stage, melanoma is located in the epidermic layer and does not has contact with the blood vessel that are only present in the dermis. Due to this, the lesion does not have metastatic capacity. This is the reason why early melanoma detection is so important [2]. Melanoma in its early stage is called melanoma *in situ*.

Several medical procedures have been proposed to examine and diagnose skin lesions. A common technique used by dermatologists to diagnose this disease is dermoscopy [2]. Dermoscopy is an inspection method which magnifies the size of a lesion up to 100x, allowing a better visualization of several dermoscopic structures that are invisible to the naked eye [3].

There are established medical procedures to analyze dermoscopy images and to help in the diagnosis [4]. Some examples of these methods are: ABCD rule [5], 7-point checklist [6] and pattern analysis [7]. Despite the existence of such methods, the diagnose of a lesion is still a subjective process that relies both on the visual acuity and experience of the dermatologist. Moreover, there are some melanomas that are very similar to other bening lesions, and it is very difficult to distinguish them. Due to these limitations, several computer aided diagnosis (CAD) systems have been proposed, for the automatic analysis and diagnosis of dermoscopy images [8]. These systems have some common steps: image preprocessing, lesion segmentation, feature extraction and classification.

There are two types of CAD systems: Pattern Recognition CAD systems and Clinically Inspired CAD systems, depending on the kind of features extracted in each system. The first ones are usually inspired by ABCD rule and try to extract features that can be related to the four criteria of this rule [8]. Some examples of these features are: asymmetry, shape, color and texture features [8]. Some works that uses these features have been achieving very promising results. However, the used features are expressed in numerical values and it is difficult correlate them with the dermoscopic criteria. Due to this, these methods do not provide comprehensive clinical information that can be used by dermatologists to understand the system classification (benign or malign lesion). This is the reason why dermatologists do not easily accept these type of CAD systems [9]. In addition, the data bases used in different works are not the same and the feature extraction processes are usually poorly described, which does not allow a direct comparison between different systems.

To overcome the lack of clinical information in classic

CAD systems, there have recently been developed some systems that are trying to replace the abstract features by clinical features. The clinically inspired CAD systems try mimic the dermatologist's procedures of analyze and diagnosing a lesion. This means that an additional step, corresponding to the extraction of dermoscopic features from numerical features, is required. Afterwards, the dermoscopic features are used to classify the lesion (e.g. [10]), making some clinical information that can be understood by dermatologists available. This clinical information makes dermoscopic features so important in automatic diagnosis. The systems should provide a set of text labels referring which are the dermoscopic criteria present in the lesion, and associating those labels with specific regions, such that they can be evaluated by physicians [9]. An approach to deal with this problem consists in detecting clinical criteria that can be related with ABCD rule and/or 7-point checklist. Some studies aiming to detected structures such as blue-whitish veil [11], regression areas, pigment networks [12], dots and streaks [13] were already conducted.

This paper describes three different method to detect four dermoscopic structures (pigment network, milia-like cysts, negative pigment network and streaks) that can be used to diagnose a skin lesion by a CAD system. Two of these methods are supervised (SVM and MLP) and another one is a weakly supervised method (Corr-LDA). Until now, the development of automatic methods for the detection of structures has a small developed by each research group for a single structure due to the difficulty in obtaining medical annotations. Recently, a large dataset (ISIC2017 [14]) was published with detailed annotation for each region (superpixel).

This document has the following structure: In section 2, an overall description of the study id presented, as well as the system's architecture; in section 3 a brief description of the two supervised methods used in this thesis (SVM and MLP) are presented; on the other hand, in section 4 a description of the Corr-LDA, a weakly supervised learning algorithm used to detected and locate dermoscopic structures based on global annotations is presented; section 5 presents the implementation methods, results and discussion, and in section 6 the main conclusions of this work are presented.

2. Overall Description

This section succinctly describes the proposed work. The sequential architecture (Fig. 1) of this work is similar to the analysis performed by dermatologists.

2.1. Local and Global Annotations

The data set of dermoscopy images used in this thesis has also available local annotations for all the images. Local annotations are labels that associate one or more clinical criteria with one or more regions (local labels). On the other hand, global annotations are text labels produced for the entire image (global labels). With this type of labels, it is not possible to have informations on what regions present the clinical criteria. The database used in this thesis [14] does not provide global annotations, but these can be obtained from local annotations, i.e, if a region of an image has a certain local annotation, it can be assumed that the image also has this global annotation.

2.2. Superpixel Segmentation

Having available the local and global annotations, the next step is superpixel segmentation. The goal of superpixel segmentation is to divide the lesion into different regions. These regions are the ones that are annotated by an expert dermatologist. The data provided by the ISIC2017 database also includes the superpixel segmentation. The lesion segmentation that is available is perfomed using the SLIC0 algorithm [15].

A lesion image's superpixels are provided as an integer-valued label map mask image. All superpixel mask images have the same spatial dimensions as their corresponding image. However, to simplify storage, superpixels masks are encoded as 8-bit-per-channel 3-channel RGB PNG images. It was necessary to run an algorithm that allowed decoding these PNG superpixel images into a label map.

2.3. Feature Extraction

After superpixel segmentation, each of the $1, 2, ..., N$ regions is characterized by a feature vector $r_n \in \mathbb{R}^f$. The feature vector is like a description of each superpixel and contains information about color and texture feature that will be related with dermoscopic structures. An image d is characterized by a set of $\mathbf{r}^d = \{r_1^d, ..., r_N^d\} \in \mathbb{R}^{f \times N_d}$. Based on the study [16], the features that is used to describe each superpixel are:

- **Color:** The mean color vector in the HSV space (μ_{HSV}). Since the original dermoscopy images are in RGB space, it is necessary converting each pixel from RGB to HSV space.

- **Texture:** In addition to color, the superpixels are described using texture features. In this work the texture features extracted are: mean contrast (μ_c) of the gray level values in the regions and statistics computed using the directional filters proposed in [12].

The pigment network, as well as the negative pigment network and the streaks, present directional structures whose directions are unknown. For this reason, a filter bank, called directional filters, was adopted. These filters are a set of $N + 1$ filters, computed at different orientations $\theta_i \in [0, \pi], i = 1, ..., N$, with the impulse response h_{θ_i} given by:

$$h_{\theta_i} = G_1(x, y) - G_2(x, y), \tag{1}$$

where G_k is a Gaussian filter:

$$G_k(x, y) = C_k \exp\{-\frac{x'^2}{2\sigma_{xk}^2} - \frac{y'^2}{2\sigma_{yk}^2}\}, k = 1, 2. \tag{2}$$

The step between two consecutive filters h_{θ_i} and $h_{\theta_{i+1}}$ is constant and equal to $\frac{\pi}{N}$. In (2) C_k is a normalization constant and the values x' and y' are related to x and y, respectively, by a rotation of amplitude θ_i:

$$x' = x \cos\theta_i + y \sin\theta_i \tag{3a}$$

$$y' = y \cos\theta_i - x \sin\theta_i. \tag{3b}$$

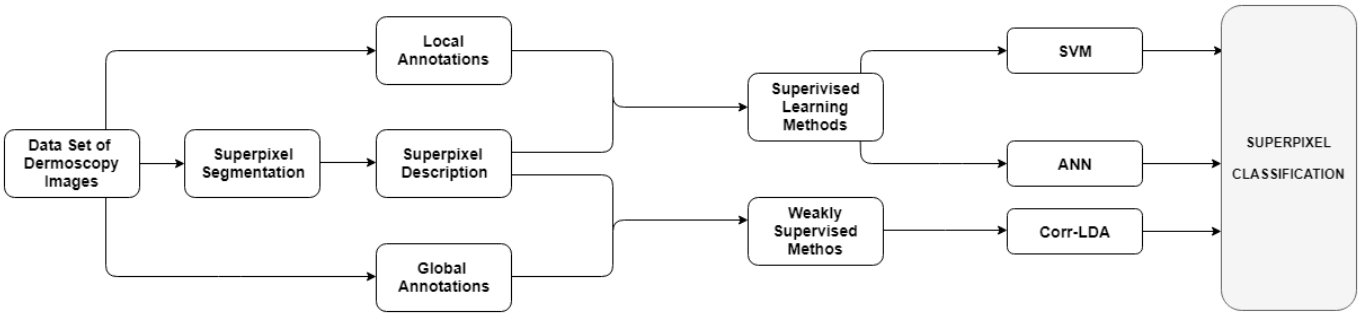


Figure 1: System's Architecture.

The values for the parameters σ_{xk} and σ_{yk} are chosen in such a way that the second filter is highly directional and the first one is less directional or isotropic. Based on image dimension and parameters chosen in [12], creating a function that, given the image dimensions returned the appropriate σ_{xk} and σ_{yk} parameters, was required in order to verify the situation described above.

An image I is filtered by each directional filter. The result image is given by the application of the following convolution:

$$I_i(x, y) = h_{\theta_i} * I(x, y), \tag{4}$$

each image I creates $N + 1$ filtered images, corresponding to each directional analyzed. To combine all directions a selection of the maximum and minimum output at each pixel (x, y) is performed:

$$J(x, y) = \max_i I_i(x.y) \tag{5a}$$

$$L(x, y) = \min_i I_i(x.y). \tag{5b}$$

All superpixels are described by the mean and standard deviation of the maximum and minimum values combined in all the directions: μ_M, σ_M, μ_m e σ_m. The contrast feature (μ_c) is obtained through the image channel of greater entropy, and it corresponds to the difference between the maximum and the minimum value of each region.

2.4. Supervised and Weakly Supervised Learning

The next step in the system's architecture (Fig. 1) is the learning phase. At this point, after supeprixel segmentation and feature extraction, it is necessary to train the algorithms to detect dermoscopic structures. In this work, two types of algorithms are trained to do automatic image annotation (AIA): Supervised Learning and Weakly Supervised Learning Methods.

The first one is trained separately for each of the possible labels, i.e, an annotation problem is created for each possible text label (as in the SVM ans ANN), which is not very practical if there is a large number of possible labels and/or images [17]. The fact that the algorithm is supervised implies that, during the training phase, the segmentation of each clinical criterion in the image, containing the information of the regions in which is observed a certain criterion, is available. This information is introduced into the input.

On the other hand, the general idea of weakly supervised learning algorithms is to introduce hidden variables, which are capable of detecting a probabilistic relation between image regions and text labels [17]. These methods do not need to know the ground truth. An example of this type of AIA is Corr-LDA [18]. This is a probabilist model used to learn the correlation between labels and regions. Corr-LDA has a capacity of generating local annotation and associating them with the different regions of a lesion, from global annotations [18].

3. Dermoscopic Structures Detection with Local Annotations

Two supervised algorithms are trained in order to identify and localize some dermosocpic structures. These methods are support vector machine (SVM) and artificial neural networks (ANN), which are capable of detecting these structures in each superpixel based on local annotations.

3.1. Support Vector Machine (SVM)

SVM is a binary classification algorithm based on supervised learning [19]. Given a set of descriptions (feature vectors) and the region's binary labels, the goal of SVM is to create an hyperplane that separates the training patterns of two classes. Since this problem may have multiple (infinite) solutions, another restriction is added: the separation hyperplane must be the one that has the largest distance to the nearest training pattern of each class. This distance is called margin. This problem can have another formulation: finding the optimal hyperplane that maximezes the margin to the training set.

When a SVM classifier is trained, one of the three situations occurs [20]: linearly separable data, non-linearly separable data and non-linear SVMs.

Sometimes it is not possible to separate the training data using an hyperplane, since the two clouds of training features overlap. To deal with this drawback, a soft-margin formulation is used. In this formulation, a penalty term is added to the optimization problem, which represents the trade-off between increasing the margin size, and ensuring that a training pattern is correctly classified. This hyperparamenter is adjusted during the training phase.

Although the aforementioned formulation assumes that training data are linearly separable, this does not happen in most of the cases. In theses cases, the training data is not linearly separable in the input space. To deal with this difficulty, the patterns are mapped into a high dimension space, called feature space, where the patterns are linearly separable. The strategy used to learn the optimal hyperplane, which can be found in [19], depends on the compu-

tation of inner product between all pairs of feature vectors and mapping them to the feature space is a hard computational problem. The strategy used to deal with this problem is called kernel trick. This solution involves the usages of a kernel function to compute the inner product between the training patterns in a high dimension, without actually having to map them. The most well-known kernel function is the Gaussian Radial Basis Function (RBF), which is the one used in this thesis. But there are other kernel functions that can be found in the literature.

3.2. Multi-layer Perceptron

Multi-layer perceptron (MLP) is a type of artificial neural networks (ANN). ANN is a supervised or unsupervised learning algorithm that can be used both in classification and regression [21]. The development of this algorithm was inspired by the characteristic functioning of the human brain, having some common aspects, such as: single processing unit (neuron), with a high number of interconnections between them; these neurons are organized in layers and have the capacity of learning based on experience.

MLPs are based on a supervised procedure, i.e., the network builds a model based on examples in data with known outputs. At structural level, MLP comprises three layers: input, hidden and output layers (Fig. 2). The information flows from the input layer through the hidden ones and finally reaches the output layer. All neurons from one layer are fully connected to neurons in the adjacent layers. These connections are represented as weights in the computational process. The weights contain the knowledge of the neural network about the problem, i.e, the solution relation.

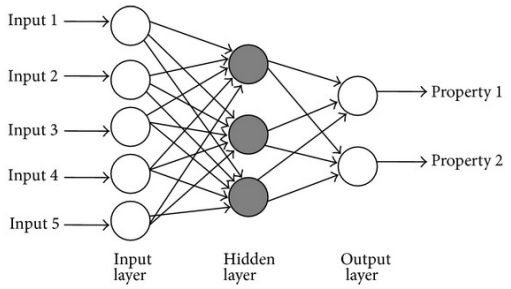


Figure 2: Multi-layer Perceptron (Image from [22]).

The number of neurons in the input layer depends on the number of independent variables in the model. The number of neurons in the output layer is equal to the number of dependent variables, i.e, depends on the predicted output and can be single or multiple. On the other hand, choosing the number of neurons in the hidden layers is not simple. Since this number depends on the complexity of the model, it is a parameter that should be chosen and adjusted during the training phase of the MLP. All neurons have an activation function that determines their output signal, depending on the input one. There are different activation functions that can be used, namely [21]: Logistic Function, Linear Function, Hyperbolic Tangent Function and Rectified Linear Unit Function (RELU).

MLP is trained based on the minimization of a cost criteria (e.g, quadratic cost or entropy) that measures the difference between the network outputs and the desired outputs. Since the cost criteria depends on non-linear network weights, cost optimization has to be done based on numerical optimization methods (e.g, gradient method). These methods allow to adjust the network weights and minimize the cost function [21]. As a result of this reduction, the following responses returned by the MLP will be closer to those desired.

In order to be used to solve classification problems, MLP must be trained, based on a training set. For this, the best network configuration and activation functions must be chosen, based on the validation set: the configuration that is chosen is the one that reproduce the best results. Usually, a neural network with fews hidden neurons produces a greater training error due to the lack of flexibility; many hidden neurons produce a low training error but may present a higher generalization error due to over-fitting. Over-fitting corresponds to situations in which the algorithm fits very well to the already observed data set, but is not able to generate good results when is applied a new set [23].

4. Dermoscopic Structures Detection with Global Annotations

When a dermatologist analyzes a lesion, he notes each image with the clinical criteria that it is present. However, the specialist does not identify the location of each criterion in the image. This type of annotation is called global annotation. In this way, this thesis intends to develop a method that is capable of generating local annotations from global annotations. The method that will be use to achieve this goal is the Corr-LDA [14]. Finally, the results obtained with this weakly supervised method will be compared with those obtained by the supervised methods proposed in the section 3.

4.1. Correspondence Latent Dirichlet Allocation (Corr-LDA)

The goal of this image annotation method is to find a relationship between text labels and image features. Corr-LDA is a generative model that first creates the patch features and then generates annotation words conditioned on the image regions [18].

One dataset comprises D dermoscopy images, each one of the d images is divided into N small non-overlapping regions (superpixels). Each superpixel is characterized by a feature vector r_n. An image d is characterized by a set $\mathbf{r} = \{r_1, ..., r_N\} \in \mathbb{R}^{f \times N}$ of N vectors. For each image d there are a set of global text labels provided by dermatologists. The text labels belong to the set $w \in \{w_1, ..., w_M\}$, where w_m is the i-th label of the dictionary w.

Based on this information, the model must allow the computation of the following probabilities: the distribution of a label given a single region $p(w_m|r_n)$, which can be use to region labeling; and the distribution of a label given the entire lesion $p(w_m|\mathbf{r})$, which is used to obtain the global labels.

The probabilistic formulation of Corr-LDA defines that, for an image d, N feature vectors to characterized each image regions are generated. Each one of these descriptors are generated conditioned on a hidden variable (topic) z_n,

being $\mathbf{z} = \{z_1, ...z_N\}$ the set of topics that was used to obtain the image d. Finally, for each of the M global annotations, one of the region is selected and a corresponding annotation w_m is obtain conditioned on the topic that was used to generate the region descriptor. The selection of the image region is performed using a latent indexing variable y_m that takes values between 1 and N.

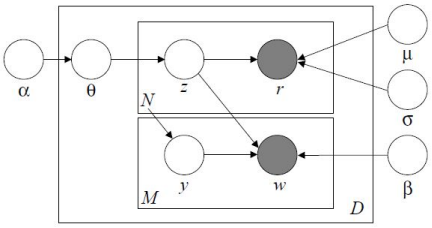


Figure 3: Coor-LDA representation (Image from [18]).

The full generative process and the parameter involved are present in the Figure 3. The generative process can be summarized as follows [18]:

1. For each image d (from a set of D images), sample a topic distribution $\theta \sim Dirichlet(\alpha)$.
2. For each of the N image regions described by r_n:
   (a) Sample a topic $z_n \sim Multinomial(\theta)$.
   (b) Sample a region descriptor $r_n \sim p(r \mid z_n, \mu, \sigma)$ from s distribution conditioned on z_n.
3. For each of the M labels w_m:
   (a) Sample an indexing variable $y_m \sim Uniform(1, ..., N)$.
   (b) Sample an annotation $w_m \sim p(w \mid y_m, \mathbf{z}, \beta)$ from a multinomial distribution conditioned on the z_{y_m} topic.

During the training phase, all of the model parameters α, Ω and β are estimated, using a training set of weakly annotated images. The common way to do this is to use a Maximum Likelihood formulation. All the steps of this formulation can be seen in [16], [18] and in the main document of this thesis. These steps end with the application of a variational Expectation-Maximization (EM) algorithm.

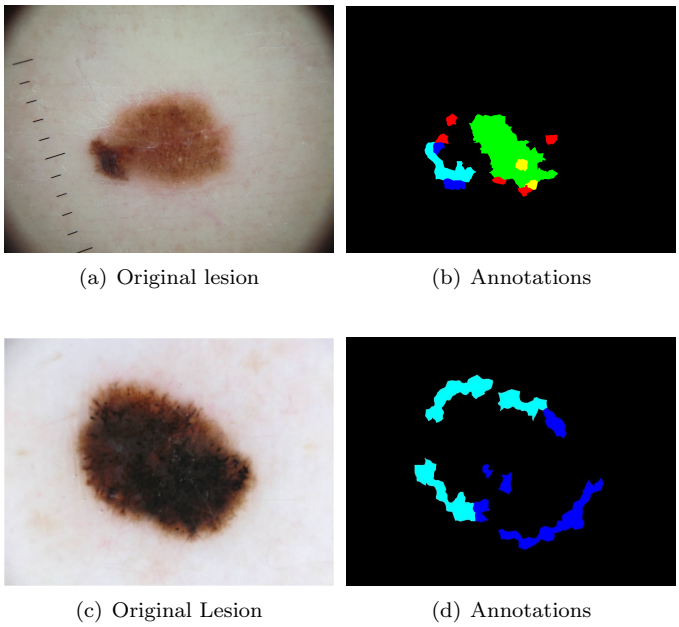
5. Implementation and Experimental Results

In this section the dataset and the evaluation metrics used in this work are presented. In addition, the way that the algorithms were implemented, the hyperparameters chosen using the validation set, and the results obtained for the test set, are also presented.

5.1. Dataset and evaluation metrics

The proposed algorithms were trained using a dataset of 2000 dermoscopy images from the ISIC2017 [14] dataset. These dataset includes a test set with 600 images and a validation set with 150 images.

All the images available in the database are segmented into superpixels, which are approximately homogenous local regions, which form a partition of the image. The database also provides the medical annotations that indicate the dermatoscopic structures present in each superpixel, from a total of 4 possible structures (pigment network, milia-like cysts, negative pigment network and streaks) - Fig. 4. This medical information is exhaustive and difficult to obtain. This type of information was availabel on a large scale for the first time with the publication of the ISIC2017 database.



(a) Original lesion (b) Annotations

(c) Original Lesion (d) Annotations

Figure 4: Examples of dermoscopy images from ISIC 2017 dataset - Original lesion and superpixel annotations: Red - Milia kile cysts; Green - Negative Pigmnet Network; Yellow - Milia like cysts and Negative Pigment Network; Blue - Pigment Network; Light Blue - Streaks.

Table 1 shows the distribution of the different dermoscopic structure in each of the image sets. It is possible to note that the number of superpixels associated with each of the structures is very unbalanced.

Table 1: Superpixels distribution and percentage in each images dataset.

		% of Superpixels				
Set	**# Superpixels total**	**Pigment Network**	**Milia like cysts**	**Negative Pigment Network**	**Streaks**	**Without structure**
Training	460272	16.92	1.01	0.71	0.46	81.12
Test	193730	10.38	0.66	1.12	0.07	87.82
Validtion	31946	10.41	1.02	1.03	0.04	87.56

The pigment network is the structure that is most represented, being the one that is present in a higher percentage of superpixels. On the other hand, negative pigment network, milia-like cyst and streaks are structures that are present in a small percentage of superpixels. The values presented in Table 1 also suggest that there are a large percentage (80-90 %) of superpixels where none of the structures are present. This matter is reflected in the classification systems, since there is a great difference between the classes, i.e., if the detection of each structure is considered as a binary problem, there are a large number of examples that belong to class 0 (without structure) and few examples that belong to class 1.

In order to evaluate the results obtained from the proposed algorithms, the same metrics were used for all the algorithms and for all the structures. The three algorithms proposed return a binary classification for each type of structure (superpixel with structure, '1', or superpixel without structure, '0'). This information is later organized into a confusion matrix, resulting in one confusion matrix for each structure and for each test performed. A confusion matrix makes it easy to compare the results obtained by the classifier with the real medical annotations and allows the classification of each superpixel as true positive (TP), true negative (TN), false positive (FP) or false negative (FN). These parameters are used to calculate the sensitivity (SE) and the specificity (SP). Sensitivity is defined as the percentage of superpixels for which each structure was correctly identified. Specificity is the percentage of superpixels for which each structure was correctly non detected. These two metrics are defined as [24]:

$$SE = \frac{\#TP}{\#TP + \#FN}, \tag{6}$$

$$SP = \frac{\#TN}{\#TN + \#FP}. \tag{7}$$

Since it is not defined which of these two metrics is most important, the desirable situation would be having the highest possible value of each of them. However, increasing one of these metrics usually means decreasing the other. For this reason, it is also chosen to evaluate the algorithms the balanced-accuracy (BACC), which for a binary classification problem is no more than the arithmetic mean between sensitivity and specificity.

5.2. Results obtained with Supervised Learning Algorithms

This section presents a description of how the supervised learning methods used (SVM and ANN) were implemented using the eight characteristics extracted from lesion images. The results obtained with these algorithms are also presented in this section.

5.2.1 SVM

The Support Vector Machine is a supervised learning algorithm that can be used in pattern recognition, which allows the detection of dermatoscopic structures. The dermoscopic structure detection was formulated as four independent binary classification problems (one for each structure), each of which was solved using an SVM classifier. The software used for this purpose was MATLAB2017b. To train each binary classifier the *fitcsvm* function was used. Since the training data was non separable, it was necessary to use a non-linear SVM. Thus, a Gaussian Radial Base (RBF) function was used as kernel function.

Based on Table 1, it is possible to note that the data is unbalanced, i.e, there are structures that are presented in a considerable percentage of superpixel, while others were associated with a small number of superpixels. To deal with this problem, we associate a class dependent weight to each structure. The value of weight assigned to each structure depends on the number of superpixels where this structure are presented, being inversely proportional to this value.

To evaluate the effect of the weights in the classification problem, five SVM models were trained for each structure, varying the weights in each model. Thus, we trained a model with the weight, P, inversely proportional to the number of superpixels with a given structure in the training set and with the weight P by adding and subtracting from it 20 % and 40 % of this value ($0.6P, 0.8P, P, 1.2P$ and $1.4P$). The P value was calculated from the training set as follows:

$$P_i = \frac{\text{\# Superpixels Total}}{\text{\# Superpixels where structure i is present}} \tag{8}$$

Table 2 presents the value of the weights P considered for each structure, as well as the values obtained for the variations of weights considered.

Table 2: Weights assigned to each structure.

	Dermoscopic Structure			
Weight	**Pigment Network**	**Milia-like cysts**	**Negative Pigment Network**	**Streaks**
0.6P	4	59	85	131
0.8P	5	79	114	174
P	6	99	142	218
1.2P	7	119	170	262
1.4P	8	139	199	305

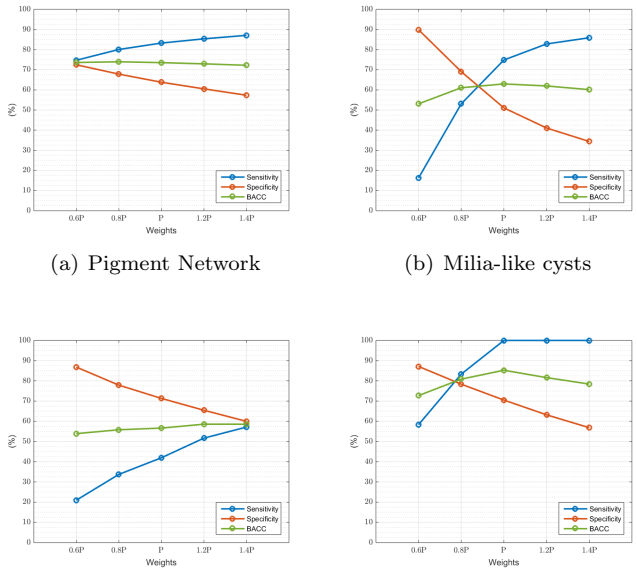
SVM is a method whose complexity increases with the number of training data. Although it is not necessary to map all the input characteristics to the feature space, it is necessary to calculate the internal product among all the training patterns, which is computationally complex. Thus, it was not possible to train a SVM classifier with all training patterns (460272). To deal with this fact, we choose to divide the training set into 9 disjoint subsets and calculate a SVM classifier for each of them.

The test and validation set was applied to each classifier, with this it was obtained 9 different annotations for each superpixel. To obtain the final annotation, the most voted annotation in each superpixel was chosen. It was used an odd number (9) of classifiers to avoid having equality in the number of votes in each class (0 or 1). A confusion matrix is constructed after obtaining of the most voted annotation, and, consequently, the metrics that was used to method evaluation.

The results obtained for the validation set are presented in Fig. 5 . These graphics allows the evaluation of the weights attributed to each structure in the SVM performance.

Figure 5 suggest that the BACC value (green line) is practically constant with the weights variation for all structures. This result confirms the trade-off between sensitivity and specificity.

Another detail that should be emphasized is that the pigment network is the only structure whose sensitivity values are always higher than the specificity for the considered weight variation. On the other hand, for the negative pigment network the sensitivity value is never higher than the specificity, in the range of weights considered. For the remaining structures, the sensitivity values are only superior to those of the specificity from a given weight: for



(a) Pigment Network (b) Milia-like cysts

(c) Negative Pigment Network (d) Streaks

Figure 5: Sensitivity, Specificity and BACC graphs, measured in the validation set, in function of the weights, obtained from SVM classifier.

milia cysts it is from the weight P and for the streaks only from the weight 0.8P.

Based on Table 1, the pigment network is the structure that is present in a greater percentage of superpixels in the training set. Due to this, this strusture is the one that presents a greater variety of examples that can be learning by SVM classifier, which may justify the fact that pigment network present sensitivity values higher than those of specificity, for all weights considered. Although a weight formulation was used, in the remaining structures the examples in the training set have low variation and diversification. This may justify the fact that the sensitivity values are only higher than those of specificity from a sufficiently high weight, since lower weights continue to give greater importance to the examples without structures.

In order to evaluate which structure presented the best results, the weight with the highest BACC of each of the four graphs presented above was chosen, corresponding to each dermoscopic structure studied. After the selection of this hyperparameter, the test set was applied to the SVM classifier trained with its weight. The results obtained are shown in Table 3.

Table 3: Best results obtained with the best hyperparameters and test set for each structure using SVM.

Structure	Weight	Sensitivity (%)	Specificity (%)	BACC (%)
Pigment Network	$0.8P$	84.63	69.18	76.91
Milia-like cysts	P	62.65	60.31	61.48
Negative Pigment Network	$1.2P$	67.62	70.80	69.21
Streaks	P	71.43	72.98	72.20

Analyzing Table 3, it is possible to conclude that pigment network is the only structure that presents better results for a weight smaller than P. Furthermore, pigment network is the structure that presents the highest BACC value of 76.91 %. For this reason, it is possible to conclude that this structure is the one which is more easily correctly classified. This result was expectable, since pigment network is the structure with the highest number of examples in the training set. Streaks is the second structure to be more easily detected, presenting a BACC of 72.20 %, followed by negative pigment network and milia-like cysts. Thus, it is possible to conclude that SVM classifiers trained for detecting structures with a small number os examples in the training set have lower performance. However, even for these structures the results obtained are quite promising.

5.2.2 ANN

As mentioned in Section 3.2, it is necessary to train several neural networks and evaluate which one produces the best results to select the number of layers and hidden neurons for a given problem. Thus, using MATLAB2017b, several parameters were optimized. In a first step, the activation function was chosen for the hidden layers neurons and the number of epochs necessary to obtain the lowest cost value. Subsequently, the best configuration and the most adequate weights assigned to each of the dermatoscopic structures were chosen.

The two activation function considered were the Rectified Linear Unit (RELU) -$F(S) = \max(0, S)$ - and the hyperbolic tangent (tanh) - $F(S) = \frac{e^S - e^{-S}}{e^S + e^{-S}}$. These activation functions were studied using the following hidden layers configuration: [30], [30,30] and [30,30,30].

The best results obtained for each ANN trained for each structure are represented in Table 4.

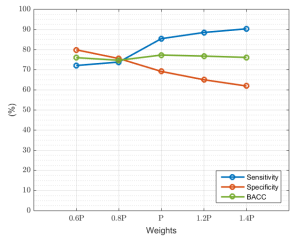
Table 4: Activation Function and number of ephocs that produce best results in validation set for each structure.

Dermoscopic Structure	Activation Function	Number of Epochs
Pigment Network	RELU	< 1000
Milia-like cysts	RELU	< 1000
Negative Pigment Network	RELU	< 1000
Streaks	tanh	< 1000

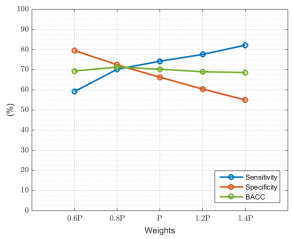
After defining the most appropriate activation function for each structure, as well as the number of epochs necessary to obtain a better network performance, the hidden layers configuration and the weights assigned to each structure were studied. Thus, for each dermoscopic structure networks with nine different configurations were trained. The number of hidden neurons considered were 10, 30 and 50, and the number of hidden layers were 1, 2 and 3. The ANN trained not only included all the configurations variation, but also the weights assigned to each structure. Thus, all the weights of Table 2 were also considered. In this way, it was obtained 45 ANN (9 configuration $\times$ 5 weights variation) for each dermoscopic structure.

Based on the results obtained for each structure (that can be found in the main document of this thesis), the best hidden layers configuration was chosen as the one that generated higher BACC values. After choosing the best hidden layer configuration, it was possible to study the weights influence in the ANN performance (Fig. 6).

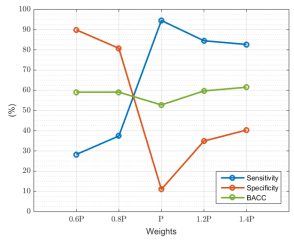
Based on Figure 6, it is possible to confirm the trade-off between sensitivity and specificity. The arguments used to



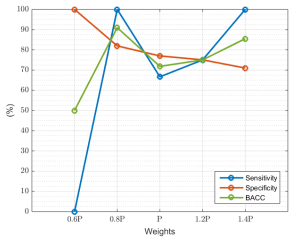
(a) Pigment Network, Configuration 10,10,10



(b) Milia-like cysts, Configuration - 30



(c) Negative Pigment Network, Configuration - 50,50,50



(d) Streaks, Configuration - 50

Figure 6: Sensitivity, Specificity and BACC graphs, measured in the validation set, in function of the weights, obtained using ANN.

justify the graphics behavior for each structure using ANN are the same ones used to justify the graphics obtained with SVM.

Another fact that needs our attention is the best hidden layer configuration of each structure, that is present in each graph description. As it is possible confirm the most complex structures, such as pigment network and negative pigment network are the ones that require configurations with more layers (3 layers) to obtain best results in the validation set. Namely, the negative pigment network is the structure that needs the most complex configuration, among all the configuration tested, to obtain the best performance.

As it was done for the case where SVM classifier was used, the weights influence was also evaluated. To do this, the weight (in the graphics of Fig. 6) with the highest BACC values was chosen for each dermoscopic structure and the test set was applied to the corresponding neural network. The results obtained are shown in Table 5.

Table 5: Best results obtained with the best hyperparameters and test set for each structure using MLP.

Structure	Configuration	Weight	Sensitivity (%)	Specificity (%)	BACC (%)
Pigment Network	[10,10,10]	P	85.37	66.92	76.15
Milia-like cysts	[30]	$0.8P$	55.80	78.72	67.26
Negative Pigment Network	[30,30,30]	$1.4P$	85.84	38.30	62.07
Streaks	[50]	$0.8P$	65.00	83.08	74.04

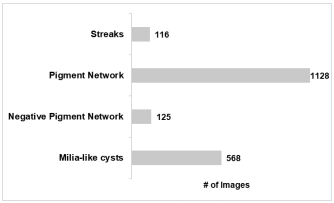
Analyzing Table 5, it is possible to see that the dermoscopic structure with the highest BACC value is, as in the case of the SVM, the pigment network, followed by the streaks, whose value is also high. Although streaks are present in a small number of superpixels, their structure is similar to the pigment network, being present in well defined directions, and being easily detected due to the use of directional filters to obtain the characteristics of each region. However, based on the results presented, it is possible to conclude that, also in the case of ANN, the structures present in a smaller percentage of superpixels in the training set (milia-like cysts and negative pigment network) are more difficult to detect, presenting a lower BACC value.

5.3. Results obtained with Weakly Supervised Learning Algorithm

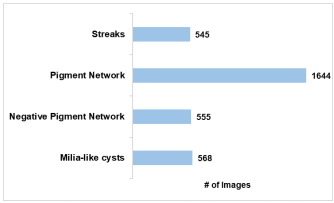
In order to apply a weakly supervised learning method, based on the image global annotations, the Cor-LDA algorithm [18] and the ISIC2017 database [14] were used. This algorithm has already been used by C. Barata *et al.* [16] for color detection in dermoscopic images and was implemented in MATLAB2017b. Thus, using the global annotations and the features extracted for each image region, it was possible to train this algorithm and find the best parameters for the model.

The topic number K is a very important parameter in this model, because it influences the distribution of Dirichlet θ. For this reason, the topic number was varied to test its effect on the results obtained. The topic number used ranged from 100 to 300, from 50 in 50 units (100, 150, 200, 250 and 300).

The ISIC2017 database was also used to train this model and, as mentioned before it is a unbalanced database. Due to this, and considering that it is impossible to introduce a weight formulation in this algorithm, the training images associated to the less represented classes were repeated. Thus, the Corr-LDA was trained with at least 500 images for each structure. Figure 7 illustrates the numerically changes that were made in the database.



(a) Original training set.



(b) Adapted training set.

Figure 7: Images number of training set where each structure is present.

It should also be noted that images that do not have any of the structures present were not discarded in order to allow the algorithm to recognize negative cases. Thus, the initial training set has 2000 images and with the changes performed it became composed of 2814 images. Before this change was made the algorithm did not produce favorable results, since the training set had, in fact, very few images where negative pigment network and streaks were present.

The Corr-LDA is a probabilistic model that uses conditional probabilities and, therefore, cannot have as input a global null legend, i.e, the model cannot receive as input a global label that does not have present any of the four structures in analysis. However, as can be seen from Table 1, there is a high percentage of superpixels (about 80 % in each of the three sets of images) that does not have any

of the four structures. For this reason, it was necessary to add a new structure that represented 'other structure'.

To test the results obtained by the algorithm when applying the test and the validation set a threshold was used. The result obtained by the Corr-LDA corresponds to an array whose size is equal to the superpixel total number of each set where the probabilities of the five structures are present (the four structures that we want to detect and the 'other structure'), $p(w_m|r_n)$. If the structures were equiprobable, the probability of each structure would be 0.2 ($1/5 = 0.2$). Thus, in order to detect which structures were most likely to be identified in each superpixel, several steps were made: choosing structures with a probability greater than 0.2 and increasing the probability threshold to 0.3, 0.4 and 0.5. This formulation allows to discard structures with very low probabilities and to assign more than one structure to each image superpixel, which is what happens in ground truth. Therefore, the hyperparameters that were evaluated using the Corr-LDA were the topics number (5) and the probability threshold (4). As a result, five sets of parameters wew obtained, one for each topic number, to which it was applied the validation set with the different probabilities threshold. The best probability threshold found for each structure is present in Table 6.

Table 6: Best threshold results obtained for each structure using Corr-LDA.

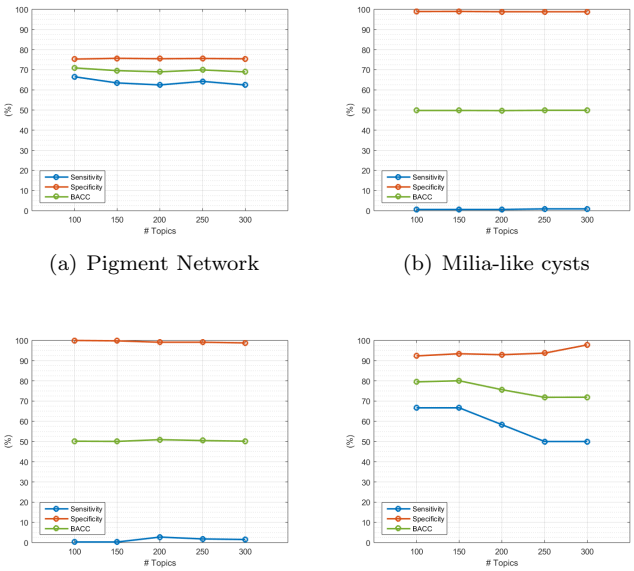
Dermoscopic Structure	**Threshold**
Pigment Network	0.3
Milia-like cysts	>0.4
Negative Pigment Network	0.2
Streaks	0.2

Based on the thresholds present on Table 6, the correspondent topic number was select and its influence one the algorithm performance was studied. To do this, the validation set was used to choose the best topic number. The results obtained are present in Figure 8.

In a first analysis, and contrary to what was expected, it is possible to verify that the variation in topics' number does not produce significant changes in the sensitivity, specificity and BACC values. It is possible to see (Fig. 8 (b) and (c)) that the milia-like cysts and the negative pigment network are practically undetectable. The graphics of these structures present sensitivity values very close to 0 %, which indicates the existence of many false negatives cases, i.e, superpixels that have the structure are classified as not having it.

The graphics corresponding to the pigment network (Fig. 8(a)) show very good results, with BACC values between 70 % and 80 %. Thus, for the case of pigment network detection, the weakly annotated algorithm achieves a very similar performance to the methods based on a very dense set of local annotations (supervised methods). This result emphasizes the high potential of this type of methods that use a small quantity of annotations to detect dermoscopic structures. The streaks structure also presents promising results, although it is a structure with few examples in the training set.

As it was done for the two supervised methods, also for the Corr-LDA the topics with higher BACC in each of the



(a) Pigment Network (b) Milia-like cysts

(c) Negative Pigment Network (d) Streaks

Figure 8: Sensitivity, Specificity and BACC graphs, measured in the validation set, in function of the weights, obtained using Corr-LDA.

structures were chosen and the method performance for the best hyperparamentes using the test set was evaluated. The results obtained are shown in Table 7.

Table 7: Best results obtained with the best hyperparameters and test set for each structure using Corr-LDA.

Structure	**# Topic**	**Sensitivity** (%)	**Specificity** (%)	**BACC** (%)
Pigment Network	100	73.34	76.00	74.67
Milia-like cysts	300	0.62	98.60	49.61
Negative Pigment Network	200	3.37	99.33	51.35
Streaks	150	25.71	94.55	60.13

As obtained in supervised algorithms, the pigment network is the structure more easily detected, followed by the streaks. This can, again, be explained by the use of directional filters in the extraction of characteristics. However, milia-like cysts and negative pigment network present a very low performance since they are structures with few examples and, being this a weakly supervised method, it has even more difficulty in detecting these structures than the supervised ones.

6. Conclusions

The automatic detection of dermatoscopic structures aid the diagnosis of skin lesions and provides an auxiliary tool to the specialist physician. The proposed methods to detect four dermosocpic structures produced very promising results.

Comparing the results that were obtained with supervised learning algorithms, to others that use the same database, it is possible to see that the results found in this work are very promising (see Table 8).

Although the supervised learning method has a best performance in the detection of the four structures, it is possible to see that the Corr-LDA, a weakly supervised method,

Table 8: Results obtained for the detection of dermatoscopic structures using the ISIC 2017 database.

Author, Ref.	Method	Sensitivity (%)	Specificity (%)	AUC*/BACC
Yuexiang Li , [25]	Deep Learning Network	66.50	91.50	0.833*
Jeremy Kawahara, [26]	Fully Convolutional Networks	54.20	98.10	0.895*
Our work	SVM	71.58	68.32	69.95
Our work	ANN	73.00	66.76	69.88

produce results as good as those obtained by the supervised methods for pigment network detection. This is the most important conclusion of this thesis and with this, it is possible to show the high potential of this type of methods that only need global annotations of the dermoscopy images to detects dermoscopic structures.

References

[1] C. McCourt, O. Dolan, and G. Gormley. Malignant melanoma: a pictorial review. *The Ulster medical journal*, 83(2):103, 2014.

[2] Dermoscopy tutorial, February 2018.

[3] L. Smith and S. MacNeil. State of the art in non-invasive imaging of cutaneous melanoma. *Skin Research and Technology*, 17(3):257–269, 2011.

[4] G. G. Rezze, B. C. Soares de Sá, and R. I. Neves. Dermoscopy: the pattern analysis. *Anais Brasileiros de Dermatologia*, 81(3):261–268, 2006.

[5] F. Nachbar, W. Stolz, et al. The abcd rule of dermatoscopy: high prospective value in the diagnosis of doubtful melanocytic skin lesions. *Journal of the American Academy of Dermatology*, 30(4), 1994.

[6] G. Argenziano, G. Fabbrocini, et al. 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. *Archives of dermatology*, 134(12):1563–1570, 1998.

[7] H. Pehamberger, A. Steiner, and K. Wolff. In vivo epiluminescence microscopy of pigmented skin lesions. i. pattern analysis of pigmented skin lesions. *Journal of the American Academy of Dermatology*, 17(4):571–583, 1987.

[8] K. Korotkov and R. Garcia. Computerized analysis of pigmented skin lesions: a review. *Artificial intelligence in medicine*, 56(2):69–90, 2012.

[9] S. Dreiseitl and M. Binder. Do physicians value decision support? a look at the effect of decision support systems on physician opinion. *Artificial intelligence in medicine*, 33(1):25–30, 2005.

[10] C. Barata, M. E. Celebi, and J. S. Marques. Development of a clinically oriented system for melanoma diagnosis. *Pattern Recognition*, 69:270–285, 2017.

[11] M. E. Celebi, H. Iyatomi, et al. Automatic detection of blue-white veil and related structures in dermoscopy images. *Computerized Medical Imaging and Graphics*, 32(8):670–677, 2008.

[12] C. Barata, J. S. Marques, and J. Rozeira. A system for the detection of pigment network in dermoscopy images using directional filters. *IEEE transactions on biomedical engineering*, 59(10):2744–2754, 2012.

[13] M. Sadeghi, T. K. Lee, et al. Detection and analysis of irregular streaks in dermoscopic images of skin lesions. *IEEE Trans. Med. Imaging*, 32(5), 2013.

[14] N. C. Codella, D. Gutman, M. E. Celebi, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *Biomedical Imaging (ISBI 2018)*. IEEE, 2018.

[15] R. Achanta, A. Shaji, et al. Slic superpixels. epfl technical report no. 149300, 2010.

[16] C. Barata, M. E. Celebi, et al. Clinically inspired analysis of dermoscopy images using a generative model. *Computer Vision and Image Understanding*, 151:124–137, 2016.

[17] G. Carneiro, A. B. Chan, et al. Supervised learning of semantic classes for image annotation and retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 29(3):394–410, 2007.

[18] D. M. Fi and M. I. Jordan. Modeling annotated data. *Special Interest Group on Information Retrieval*, pages 127–134, 2003.

[19] T. Fletcher. Support vector machines explained. *Tutorial paper*, 2009.

[20] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[21] K. L. Priddy and P. E. Keller. *Artificial neural networks: an introduction*, volume 68. SPIE press, 2005.

[22] J. Swarbrick. *Encyclopedia of pharmaceutical technology*. CRC Press, 2013.

[23] S. Lawrence, C. L. Giles, et al. Lessons in neural network training: Overfitting may be harder than expected. In *Proceeding of the Fourteenth National Conference of Artificial Intelligence*, 1997.

[24] T. Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[25] Y. Li and L. Shen. Skin lesion analysis towards melanoma detection using deep learning network. *Sensors*, 18(2):556, 2018.

[26] Jeremy Kawahara and Ghassan Hamarneh. Fully convolutional networks to detect clinical dermoscopic features. *arXiv preprint arXiv:1703.04559*, 2017.