

Detection of melanoma from dermoscopic images of naevi acquired under uncontrolled conditions

Arthur Tenenhaus¹, Alex Nkengne², Jean-François Horn^{3,4}, Camille Serruys^{3,4}, Alain Giron^{3,4} and Bernard Fertil⁵

- 1- Supelec, Department of Signal Processing & Electronic Systems, Gif-sur-Yvette, France
- 2- Johnson & Johnson Consumer France SAS, Skin Care Research Institute, Issy-les-moulineaux, France
- 3- INSERM, U 678, Paris, France
- 4- Pierre et Marie Curie-Paris University, UMR 678, Paris, France
- 5- LSIS (UMR CNRS 6168), Marseilles, France

Corresponding author:

Arthur Tenenhaus

SUPELEC, Department of Signal Processing & Electronic Systems

Plateau de Moulon, 3 rue Joliot-Curie

91192 Gif-sur-Yvette Cedex, France

Arthur.tenenhaus@supelec.fr

Number of pages: 19

Number of tables: 4

1 Abstract:

Background and objective: Several systems for the diagnosis of melanoma from images of naevi obtained under controlled conditions have demonstrated comparable efficiency with dermatologists. However, their robustness to analyze daily routine images was sometimes questionable. The purpose of this work is to investigate to what extent the automatic melanoma diagnosis may be achieved from the analysis of uncontrolled images of pigmented skin lesions.

Material and methods: Images were acquired during regular practice by two dermatologists using a Reflex® 24X36 cameras combined with HEINE DELTA 10 dermoscopes. The images were then digitalized using a scanner.

In addition, five senior dermatologists were asked to give diagnosis and therapeutic decision (exeresis) for 227 images of naevi, together with opinion about the existence of malignancy predictive features. Meanwhile, a learning by sample classifier for the diagnosis of melanoma was built, which combines image-processing with machine learning techniques.

After an automatic segmentation, geometric and colorimetric parameters were extracted from images and selected according to their efficiency in predicting malignancy features. A diagnosis was subsequently provided based on selected parameters. An extensive comparison of dermatologists' and computer results was subsequently performed.

Results and conclusion: the KL-PLS based classifier shows comparable performances with respect to dermatologists (sensitivity: 95% and specificity: 60%). The algorithm provides an original insight into the clinical knowledge of pigmented skin lesion.

2 Introduction

In 2003, Rosado et al. have reviewed the literature about computer diagnosis of melanoma (1). They found a wide spectrum of methods (thirty) spanning over the 1991-2002 decade. As a general rule, they observed that dermatologists and computers have comparable efficiencies under controlled experimental conditions but they raised doubts about the real value of the computer diagnosis in daily routine. They defined a set of quality requirements to be fulfilled to ascertain the validity of future computer-aided melanoma diagnosis systems. As a matter of fact, they claimed that many past studies could have been potentially affected by different bias, including "selection bias and verification bias" (1). Some experimental designs were found partly jeopardized by the poor, representativeness of data sets and the lack of validation phase. They do not report however differences in efficiency between "good quality" studies and others.

Meanwhile, many systems (Molemax, Fotofinder, Xcell diagnostic, Siascope, ...) have been developed using powerful and dedicated video cameras. Usually performances are granted but a recent study shows that such systems tend to overdiagnose benign melanocytic lesions as potential melanomas (2). We conjecture that the construction of these systems doesn't fulfilled Rosado's requirements. Moreover, these systems have a cost, related to the acquisition material, which does not support their wide diffusion to physicians. They are also cumbersome and rely on plug-in devices that preclude their use for outdoor screening campaigns. It is worth pointing out that the recent evolution of numerical cameras, together with the availability of inexpensive macro ring speedlights and dermoscopic accessories greatly simplify the acquisition of reliable images of skin tumors. Actual quality of numerical photography now allows dermatologists delivering working diagnosis from the mere examination of digital images acquired under "relaxed" conditions (i.e. using standard camera fitted with a Dermoscope-Epiluminescence (D-ELM) apparatus for example) (3).

As a matter of fact, human eyes (and brain) are very efficient in dealing with the analysis of images obtained with variable lighting conditions. Recent advances in statistical learning methods however, allow coping with the intricate process of characterization/classification of complex data especially when variables of interest are blurred by many irrelevant factors of variability (4).

The purpose of this work is to investigate to what extent the automatic melanoma diagnosis may be achieved from the analysis of uncontrolled images of pigmented skin lesions.

The general framework of our approach is built upon the learning paradigm (using supervised learning methods). Several studies have already described automatic systems combining dedicated apparatuses, image processing and statistical learning

theory (5-15). The very satisfying results that were obtained lead us to consider applying this approach to the analysis of digital images acquired under more “relaxed” conditions. Using diagnosis, therapeutic decision and features detection, provided by a panel of senior dermatologists for a set of representative images, our system mimics the malignancy detection process. Learning from an in depth dermatologists’ expertise, our approach systematically combines image processing with statistical learning techniques in each phase of the system, from the segmentation of the tumor and the research of predictive features to the diagnosis and the decision of exeresis (tumor removal). This strategy leads to integrate a prior knowledge based-expert in every phase of the construction of the automated melanoma system.

3 Materials and methods

3.1 General considerations

The diagnosis and associated therapeutic decision (excision essentially) for black skin tumors is a multi-step procedure. According to dermatologist’s “rules of good clinical practice”, a first step consists in detecting malignancy predictive features, including the popular ABCD ones (Asymmetry, Border, Color, Diameter). In a second step, dermatologists combine these features according to their capacity in predicting malignancy.

The automatic characterization of pigmented skin lesion was built following a very similar process. After the segmentation of tumors, parameters of malignancy were extracted. A general model was built thereafter (second step), using the selected parameters as input for the diagnosis of melanoma.

According to Rosado’s recommendations (1), an intensive comparison between dermatologists’ expertise and a melanoma detection system should be required to validate the melanoma database and the melanoma detection system. The performances of our melanoma classification system were thus compared to those of five senior dermatologists.

3.2 Image database

The initial image database used in this study included 900 images of pigmented skin lesions from the dermatology departments of the British Hertfort Hospital and the Louis Mourier Hospital in “Ile de France” (France). Images were acquired with Reflex® 24X36 cameras combined with HEINE DELTA 10 dermoscopes. Multiplicity of locations and operators resulted in a significant level of variability (a welcome effect for the present study). We refer to this variability as “uncontrolled” conditions for image acquisition. Magnification was the only parameter under control whereas illumination conditions were highly variable and no calibration was performed. Images were then digitalized using a x-finity scanner (Optic resolution: 600x2400, color depth: 36 bits). To meet requirements linked to the learning framework, the database had to include a great

variety of tumors with representative examples. As a consequence of the inclusion protocol (unrelated runs of selection), many tumors were quite similar, melanoma were largely minority. As we were seeking to have high sensitivity in detecting melanoma, we included all identified melanoma lesions in the learning database, to favor its detection, the learning database being completed to 227 with randomly selected tumors (Tab 1). Doing so, it appeared that 77 lesions were classified as benign lesions (class 1). In order not to cause any needless distress to the patient, the majority of benign lesions were not surgically excised. Dysplastic lesions (class 2) were represented in the dataset by 118 pigmented lesions. Thirty-two pigmented lesions were categorized as malignant melanomas (class 3). Lesions of classes 2 and 3 were all surgically excised and histopathologically analyzed. For this study, 2 classes were finally considered: histologically-confirmed melanomas on one side, the remaining lesions on the other. For simplicity, this classification is referred to as the “gold standard” diagnosis in this paper.

Table 1. The learning database: lesions with diagnosis and clinician’s decision

Diagnosis	# case	(%)	Excision
<i>Melanoma</i>	27	(11.9%)	<i>excised</i>
<i>Nodular melanoma</i>	1	(0.45%)	<i>excised</i>
<i>Dubreuilh’s melanoma</i>	4	(1.8%)	<i>excised</i>
<i>Dysplastic nevus</i>	118	(51.2%)	<i>excised</i>
<i>Benign nevus</i>	62	(27.3%)	<i>not excised</i>
<i>Blue benign nevus</i>	2	(0.9%)	<i>excised</i>
<i>Congenital benign nevus</i>	5	(2.2%)	<i>excised</i>
<i>Junctional & dermic benign nevus</i>	7	(3.1%)	<i>excised</i>
<i>Palm-plantar benign nevus</i>	1	(0.45%)	<i>excised</i>
Total	227	100%	(72.7%)

3.3 Dermatological expertise

Five senior dermatologists were asked for their expertise about the 227 selected images. They were presented each tumor both as macroscopic image and dermoscopic image. They subsequently gave their opinion about the presence of ABCD and dermoscopic features (dichotomic answer), their diagnosis (melanoma, dysplastic or benign lesion) and their therapeutic decision (dichotomic answer, excision/ non excision). They also provided hand-made segmentation of tumors. As illustration of the collected data, figure 1 summarizes parts of the collected expertise and level of agreement between dermatologists. Obviously, dermatologists do not necessarily agree about presence of features, diagnosis and therapeutic decision (referred to as “items” in the following). In order to decide about the status of each image with respect to the “item” under investigation, a voting schema has been implemented. For example, asymmetry for a given tumor was given a score of 5 when all dermatologists answered positively to the corresponding question (figure 1). It can be subsequently observed that dermatologists agree on the presence/absence of the asymmetry feature for 54% of tumors (grouping of tumors receiving 0 or 5 votes), whereas they highly disagree for 20% of tumors

(grouping of tumors receiving 2 or 3 votes). Histograms of votes per tumor (left column of figure 1) allow evaluating level of agreement between dermatologists together with the distribution of the “item” among tumors in the database, and, in particular, among melanoma (dark areas). 75 pigmented skin lesions are diagnosed by all dermatologists as asymmetric. Twenty-two of these lesions are melanoma. Using a threshold between 2 and 3, it can be decided (at the majority vote) about the absence/presence of the “item” for each tumor. This value is used thereafter as reference (target) to select relevant parameters in images. In contrast with the asymmetric feature, most of the tumors in the database are labeled “multicolor” by dermatologists, although agreement is far from perfect. Specificity and sensibility of the “item” with respect to melanoma can also be calculated (Figure 1, right column).

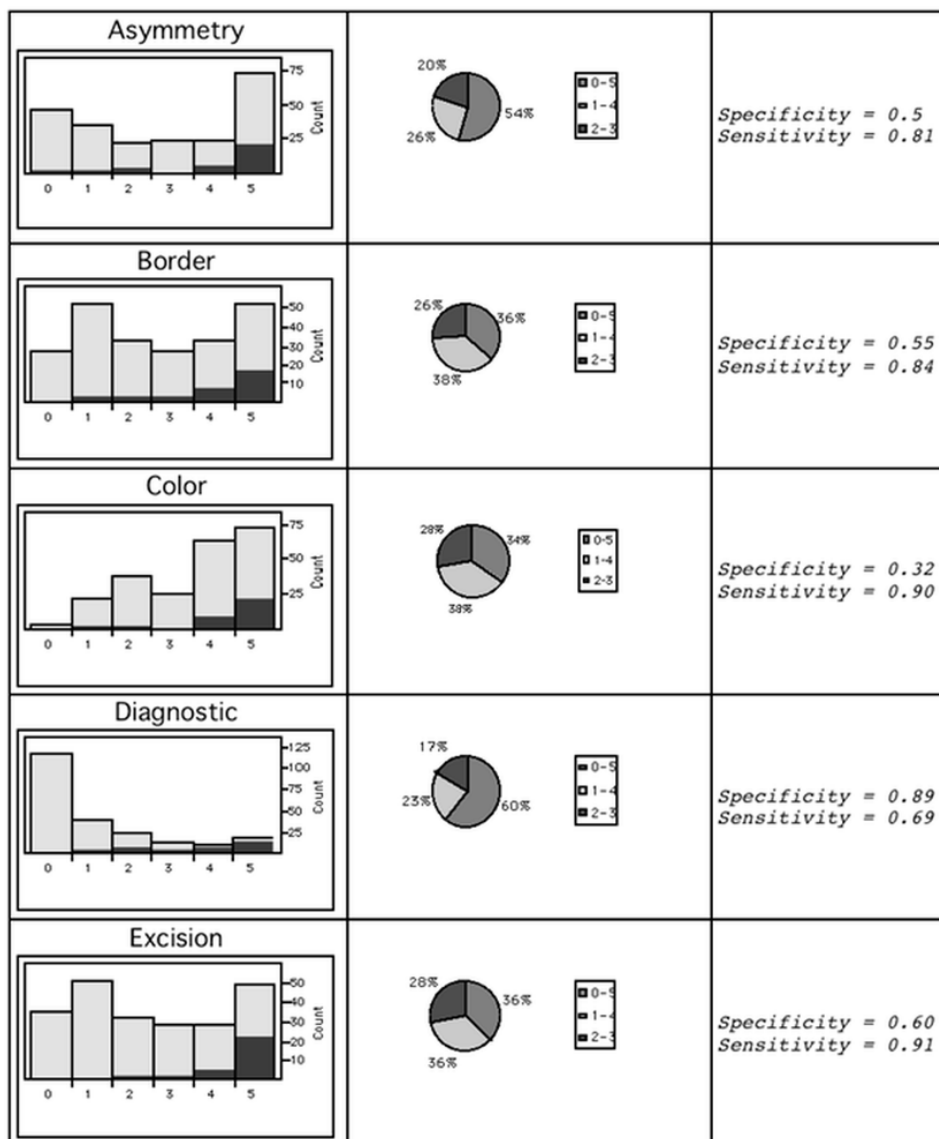


Figure 1. Distribution of the 227 images for some items as a function of the dermatologists' vote (left column). For example, bin 3 of histograms groups images declared with the feature by 3 experts and without it by the 2 others; Grouping of bins (0 and 5, 1 and 4, 2 and 3) allows evaluating dermatologists' agreement (middle column) Sensitivity and specificity with respect to the melanoma diagnosis (right column)

Figure 1 shows that full agreement between dermatologists is high (60%) as far as diagnosis is concerned (12 out of 32 for melanoma), whereas therapeutic decision is more disputed (27 out of 195 for non melanoma tumors). Interestingly, according to the majority vote, 22 melanoma out of 32 are detected and all melanomas but 3 are considered worthy of excision: sensitivity and specificity for diagnosis as well as therapeutic decision are higher if clinicians' advices are pooled.

3.4 Segmentation

Most of the time, visual segmentation of tumor (i.e. localization of tumor boundary) by dermatologist is easy. However, irreducible fuzziness remains occasionally, when transition between lesion and surrounding skin is too smooth. The numerous papers devoted to boundary detection of skin tumors demonstrate that it is still an open problem for computers (16-20). As a matter of fact, lesions have a large range of size, color, texture and are more or less contrasted. In addition, contrast between skin and tumor is highly variable and it may even change along the border. Based on an original multi-scale classification scheme, our approach mimics dermatologists that first look at the image as a whole, and then concentrate on local details to precisely localized the border. We consider segmentation here as a supervised classification task: given a pixel of the image, the objective is to know whether it belongs to the lesion or not. Dermatologists were initially asked to outline border of tumors. The derived masks thus provide targets (output) for the learning phases since every pixel may be labeled inside or outside the tumor.

First, tumors are localized on low resolution images (1/16) using a classifier that takes 3x3-pixel windows (48x48 pixels of the initial image), mean intensity of central pixels (assumed to be in the tumor) and mean intensity of outer pixels (assumed to be outside the tumor) as inputs. A coarse segmentation is then achieved using local information (9x9-pixel windows from image with a higher resolution) and global information obtained from the preceding step (including the mean intensities of the lesion and the skin). The classification process is reiterated down to the initial image resolution (Figure 2). Note that the classification does not concern all pixels at each step. Only pixels localized in the vicinity of the border defined at the preceding step are analyzed. The size of the strip defining the vicinity is directly linked to the size of the previously used screening window. Doing so, many artifacts of segmentation that would have appeared at various levels of resolution are eliminated. Grey level images are processed throughout this phase, since no significant differences were observed between segmentations achieved by dermatologists with full color images and grey level images.

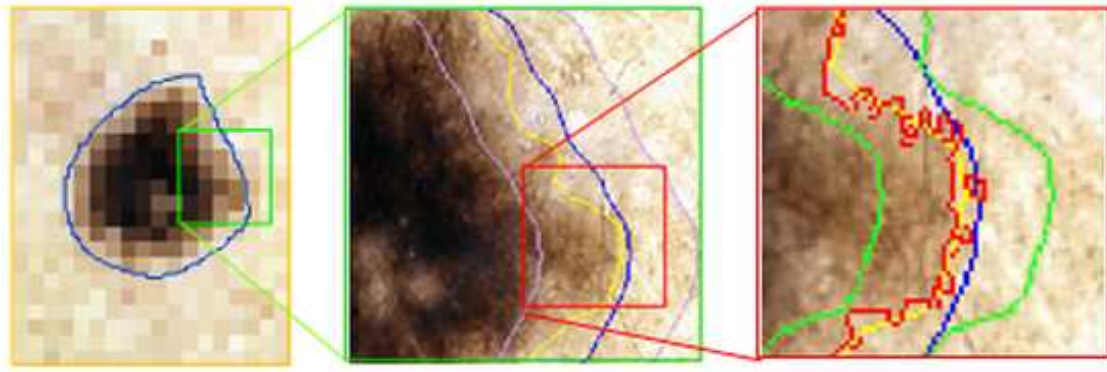


Figure 2. Multi-scale approach illustration based on a three-level resolution analysis: the location of the lesion is observed for resolution $1/16$ (left panel), a coarse segmentation is obtained for resolution $1/2$ (central panel), the final one is obtained for resolution $1/1$ (right panel). The neighborhood is set to 9×9 windows for the last 2 steps.

About one hundred images were used as learning dataset (featuring 100 000 windows) and 80 images as validation dataset. Classifications were performed using logistic regression. Incidentally, the segmentation provides the first -and highly valuable- malignancy parameter: the size of the tumor.

Three criteria are used to compare binary masks resulting from the multi-scale segmentation process with those drawn by dermatologists.

1. Distance between contours:

Let $A = \{a_1, \dots, a_{N_a}\}$ and $B = \{b_1, \dots, b_{N_b}\}$ the contours defined by the masks under comparison, (where we note a_1, \dots, a_{N_a} and N_a the pixels and length of contours A).

The distance between 2 contours A and B may be defined as (3):

$$D(A,B) = 1/2 \times \left(\frac{1}{N_a} \sum_{i=1}^{N_a} \min_{j=1}^{N_b} [d(a_i, b_j)] + \frac{1}{N_b} \sum_{j=1}^{N_b} \min_{i=1}^{N_a} [d(a_i, b_j)] \right) \quad (3)$$

where $d(a_i, b_j)$ is the Euclidean distance between the a_i and b_j .

2. The sensitivity may be defined as the percentage of the pixels within the mask to be tested that lies within the mask of reference.
3. The specificity may be defined as the percentage of the pixels outside the mask to be tested that lies outside the mask of reference.

3.5 Detection and validation of malignancy features

Within the scope of this study, several parameters in relation with ABCD rules were extracted from the images of tumors. Parameters presented below are selected for their

ability to express the feature under investigation (asymmetry for example), after examination of the classification results.

In medical diagnosis applications, sensitivity and specificity are far more relevant than accuracy because of different misclassification costs. Moreover, most medical datasets include only a few hundred cases with an unbalanced distribution of classes. In order to address these issues, many medical applications need cost-sensitive algorithms. In this study, we use a measure (introduced by Sboner to compare the diagnosis of classifiers (11)) that may be useful for comparing classifier and dermatologists' expertises. Given that the ideal classifier has both sensitivity and specificity equal to 1.0, Sboner defines the distance of a real classifier from the ideal one in this way:

$$d_{quality} = \sqrt{(1 - sensitivity)^2 + (1 - specificity)^2} \quad (2)$$

Considering that $d_{quality}$ allows balancing sensitivity and specificity, it is used here to evaluate the ability of the various extracted parameters to capture dermatologist's expertise about malignancy features

3.5.1 Asymmetry

Asymmetry is an important characteristic for differentiating malignant and benign pigmented skin lesions. In fact, asymmetry as defined by dermatologists is a complex mixture of shape and texture. Symmetry of shape can be estimated from the overlapping of the tumor surface after rotation around a given symmetry axis. Two symmetry axes are considered to account for the complexity of shape displayed by tumors. They were calculated with the Hough transform algorithm. Asymmetry (for each axis) is expressed by the percentage of overlapping pixels. Similarly, 2 symmetry axes are considered to qualify texture distribution over the tumor. As observed by Schmid-Saugeon, the first axes of a Principal Components Analysis (PCA) of the pixels of the lesion weighted by their intensity are good candidates for that task (12). Texture symmetry indexes are taken as the quadratic error averages between the intensity of overlapping pixels after rotation around the symmetry axes. Dermatologist' asymmetry feature is consequently described by a 4-dimensional vector. Figure 3 depicts texture and shape symmetry axes for a given lesion.

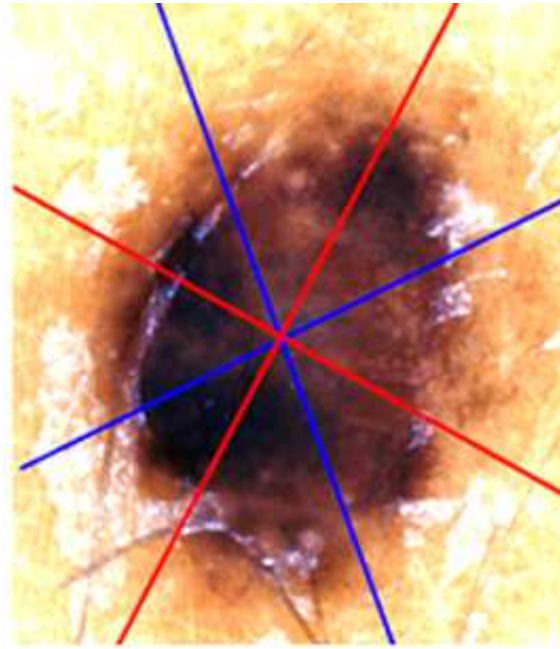


Figure 3. Texture (blue) and shape (red) symmetry axes

3.5.2 Colorimetric study

Pigmentation of lesion is generally not uniform: the presence of several colors may contribute predicting malignancy (the “variegated color” feature). Usually, dermatologists locate zones of homogeneous color and then determine the “number of colors” in the lesion, by considering the number and size of homogeneous zones. They refer to brown or black with shades of red, white, or blue to qualify colors in lesions. Color information is considered as an essential parameter: it has been proposed as the main input for melanoma classification in recent publications (14, 21, 22). To proceed similarly as dermatologists, two unsupervised classification methods are used to describe color heterogeneity: Kohonen map and K-means clustering.

Kohonen map

Kohonen map is used here to account for variation in colors in a given tumor relatively to the spectrum of colors found in the tumor database. It provides an absolute categorization of colors. Five randomly selected pixels of each tumor of the database are used to get a 5x5 Kohonen map in the RGB space. The map is subsequently representative of the diversity of colors found in lesions. Let us consider the projection of all pixels of a pigmented skin lesion onto the map. In case of a single color lesion, most of the pixels are projected in the same region, whereas pixels of multicolor pigmented skin lesions are projected all over the map (Figure 4). The proportion of pixels projected onto each of the 25 neurons of the map thus provides a 25-dimensional vector to be used for the coding of the “color variegation” feature.

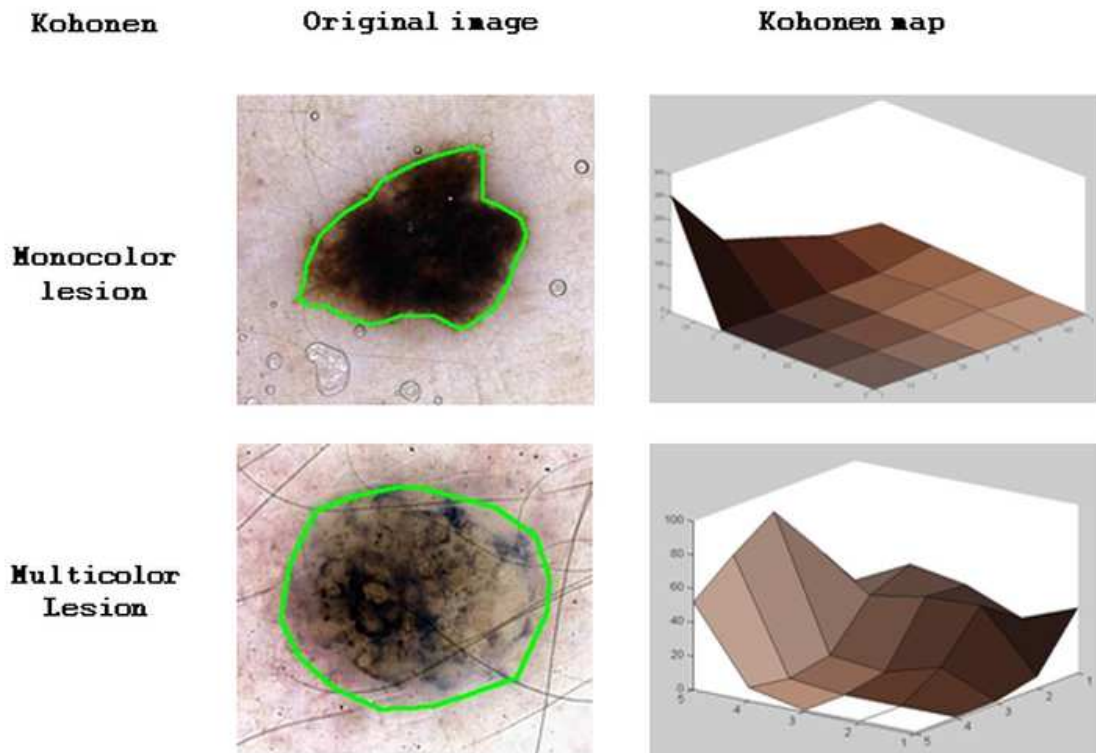


Figure 4. Kohonen method for the evaluation of the color variegation feature

Kmeans clustering

Since we accept not strictly controlling image acquisition (including calibration), it is worth considering an alternative approach of color analysis, based on the specific properties of each lesion and related image. The K-means clustering allows adapting color analysis to each case. It provides a relative categorization of colors. In the K-Means clustering, a set of n objects in m -dimensions is given. The goal is to arrange these objects into K clusters, with each cluster having a representative prototype, usually chosen as the centroid of the objects in the cluster. When applied to the classification of the pixels in a tumor, the dimension of the points is $m = 3$ (the Red, Green and Blue channels) and the optimal number of clusters should be the number of colors detected by the dermatologists. Considering that dermatologists rarely observed more than 4 colors in a given tumor, it was decided to run a 4-Means. As a consequence of the 4-Means, images of lesions are interpreted with 4 colors, showing areas covered with each color (Figure 5). The sample size of clusters (4 variables), together with the color of K centroids (3 x 4 variables) therefore constitute the features of interest expected to characterize the color variegation feature. Thus, the 4-Means step provides a 16-dimensional vector.

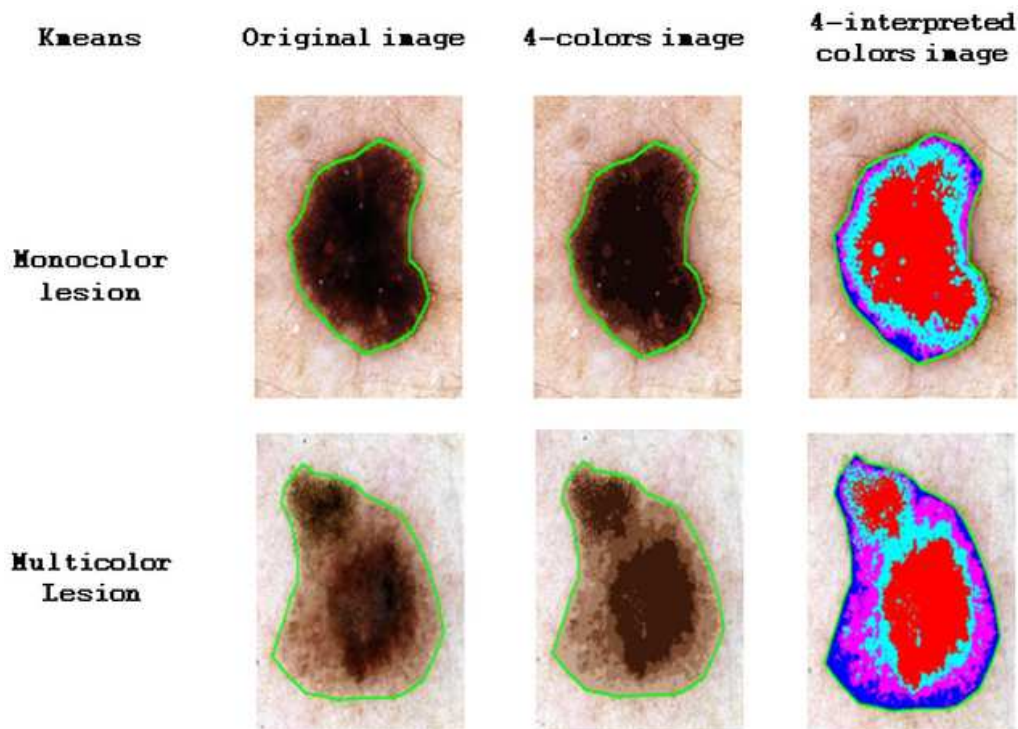


Figure 5. Kmeans for the evaluation of the color variegation feature

Blue Veil

The Hue component in the HSV space (Hue, Saturation and Value (brightness)) codes for the pixel coloration; it ranges from 0 to 360. The detection of a blue coloration can be easily coded by the proportion of pixels in the blue interval (200-250).

The color features are consequently described by a 42-dimensional vector

3.6 Classification

The Kernel Logistic PLS (KL-PLS) (23), which has proven to be competitive with other state-of-the-art supervised approaches such as Support Vector Machines [Boser, 1992], is used throughout this paper. KL-PLS is a supervised nonlinear dimensionality reduction method for binary classification. KL-PLS combines the flexibility of the empirical kernel map (24) with the supervised properties of the PLS logistic regression (25). The main idea behind KL-PLS is to look for a discriminant space spanned by the KL-PLS components (latent variables), where a simple model, such as the logistic regression may become efficient for classification. Considering that the probability of finding a simple decision rule (an hyperplane in the data space for example) increases with the dimension of the space, we use here the kernel matrix to map the data into a higher-dimensional space. The melanoma detection system based on KL-PLS is built using all the 47 selected features as input (each lesion is characterized by five geometric and 42 colorimetric features.) and the histological diagnosis as output. KL-PLS was also applied to the features validation procedure.

For comparison purposes, a logistic regression and a radial basis function (RBF) neural network (with a Gaussian radial kernel and trained with the standard sequential algorithm) were also used to detect melanoma from the 47 selected features.

3.6.1 Statistical validation:

Statistical validation is required in the framework of learning methods, especially when powerful classification models are concerned and when the size of the database is “relatively” small. Cross-validation is usually the good approach in such a situation. It consists in splitting the database into 2 subsets: a training dataset used to build the classifier and a testing dataset devoted to evaluation. Classification results consequently only concern data not seen previously by the classifier. Throughout this paper, a “leave-one-out” cross-validation technique is used. Let n be the size of the database. According to the “leave-one-out” scheme, n classifiers are built, using $n-1$ images, by discarding successively one image at a time. The efficiency of classification is subsequently observed by collecting results of classification for the discarded images. All classification results presented below are obtained that way.

3.6.2 Dermatologists' model:

According to Rosado's recommendations (1), an intensive comparison between dermatologists' expertise and a melanoma detection system is required to validate the melanoma database and the melanoma detection system. “Performances of Dermatologists”, are evaluated in order to describe their ability to diagnose melanoma from their observations of predictive features (Features are size of the tumor, asymmetry, color variegation, border regularity).

3.7 Results

3.7.1 Segmentation

As far as the different levels of segmentation are concerned, whatever the resolution level, classification rate is about 75%, at the pixel level,. Thanks to the global information that is propagated from one level of resolution to the next one, the method appears robust to brightness and dominant color variations between images.

The comparison of ten lesions segmented by four dermatologists provided the reference for the three criteria described in the method paragraph. It was observed that the maximum distance between contours was 69, the minimum sensitivity was 89% and the minimum specificity was 68%. The automatic segmentation of 86% of the images in our database meets these requirements. Segmentation errors essentially occur when: i) the lesion covers a large part of the image and reaches its borders, ii) there is a considerable artifact on the image (nail, oil bubble), iii) illumination conditions are really poor and/or when the contrast between tumor and surrounding skin is low (Figure 6). Although these situations can be automatically detected and proposed for manual correction, all masks were systematically validated by visual examination in this work. Despite the high diversity of analyzed tumors, the multi-scale segmentation of skin lesions

demonstrated its robustness and appears very reliable. Note that the segmentation mask directly provides a multi-dimensional vector of characteristics, among them size, perimeter, compactness ... that could have been used for the classification tasks to come. However, considering that the size of the tumor is a much more important parameter than the others, it was the only parameter given to the classifier in this study.

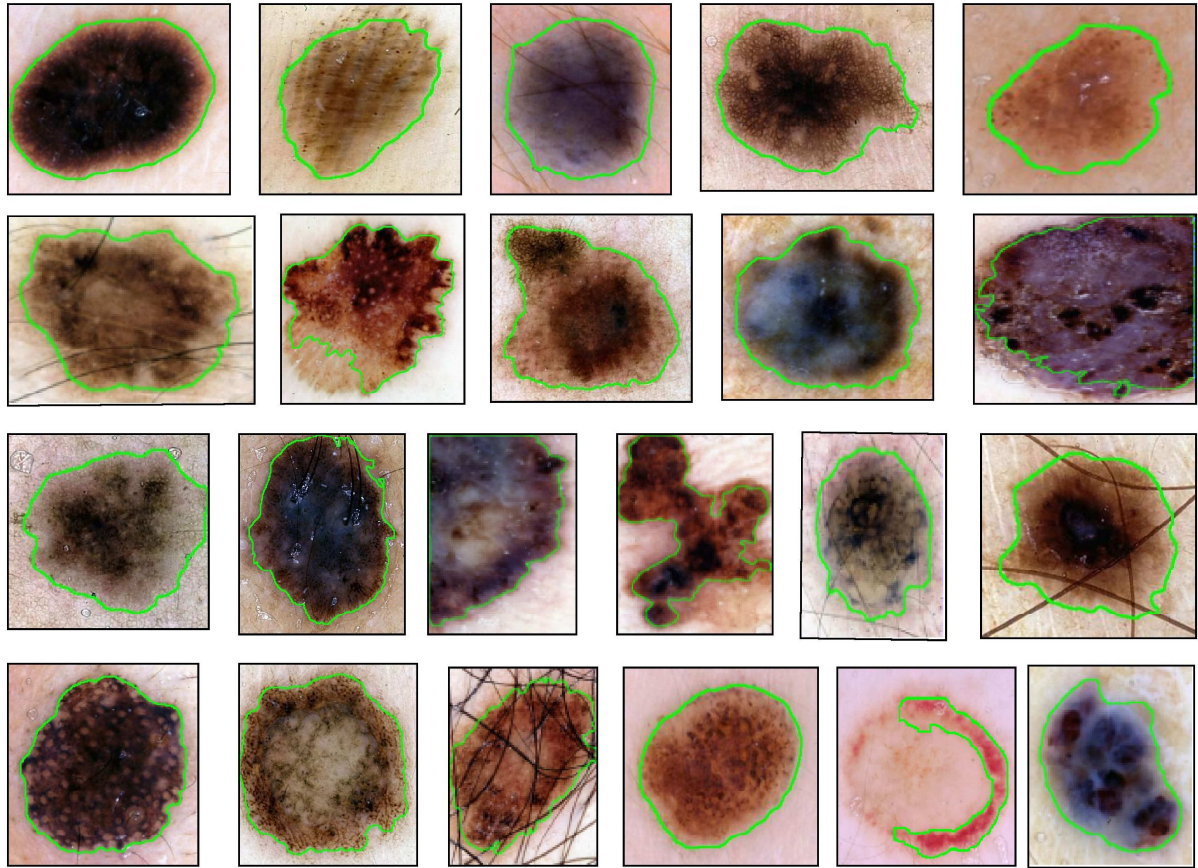


Figure 6. Several Segmentations of naevi

3.7.2 Features evaluation

Asymmetry

All 227 tumors are checked for the asymmetric feature by dermatologists. As previously described, a majority decision provides the “gold standard” about the feature. The automatic detection of asymmetry is based on a four-dimensional vector. A KL-PLS based classifier (Gaussian kernel parameter = $1e-2$, number of retained KL-PLS components = 1) has been built using this vector as input and the asymmetry gold standard feature as output. It allows evaluating the efficiency of the vector for the detection of the feature, as compared to dermatologists. Asymmetry is found by dermatologists in 54% of the tumors. Level of agreement is high as it can be seen from the characteristically 2-mode histogram and corresponding pie graph (figure 1). Therefore, sensitivity, specificity and $d_{quality}$ are high for most of the dermatologists (table 2). In contrast, the classifier performs moderately well, suggesting that there is room for improvement. It is clear that the asymmetric feature does not merely result

from simple geometric considerations (26), (12). However, since the classifier efficiency reaches 73% (only 62 tumors out of 227 are misclassified with respect to the asymmetric feature), the 4-dimensional vector may be found useful for the detection of melanoma. It seems interesting to note that dermatologist 3 is at variance with his colleagues: the high sensitivity and low specificity he gets indicate that he considers many more tumors as asymmetric, including those found positive by his colleagues.

Table 2. Classifier and dermatologists' efficiencies for the detection of asymmetry. Sensitivity, specificity and $d_{quality}$ are calculated with respect to the majority decision gold standard.

Asymmetry	Sensitivity / Specificity	$d_{quality}$
Dermatologist 2	0.94 / 0.87	0.14
Dermatologist 5	0.84 / 0.88	0.2
Dermatologist 1	0.84 / 0.88	0.2
Dermatologist 4	0.78 / 0.98	0.22
Classifier	0.73 / 0.72	0.38
Dermatologist 3	0.98 / 0.61	0.39

Color variegation

All 227 tumors are checked for the presence of the color variegation feature by dermatologists. As previously described, a majority decision provides the "gold standard" about the feature. The automatic detection of color variegation is based on a 42-dimensional vector. A KL-PLS based classifier (Gaussian kernel parameter = $1e-5$, number of retained KL-PLS components = 10) is built using this vector as input and the color variegation gold standard feature as output. Color variegation is found by dermatologists in 66% of the tumors, but level of agreement is low as it can be seen from figure 1. Sensitivity, specificity and $d_{quality}$ for clinicians are lower than for asymmetry (table 3). In addition, the variability of specificity among dermatologists is high. In fact, specificity is inversely correlated with the proportion of tumors classified as multicolor by physicians. The level of triggering for the color variegation feature essentially explains the variation of specificity among dermatologists. The classifier performs properly, suggesting that the 42-dimensional vector captures the desired information.

Table 3. Color: Classifier and dermatologists' performances: Sensitivity, specificity and $d_{quality}$ are calculated with respect to the majority decision gold standard.

Color	Sensitivity/Specificity	$d_{quality}$
Dermatologist 2	0.84 / 0.95	0.17
Dermatologist 4	0.91 / 0.77	0.24
classifier	0.78 / 0.67	0.39
Dermatologist 5	0.59 / 0.94	0.44
Dermatologist 3	0.97 / 0.54	0.46
Dermatologist 1	0.97 / 0.26	0.74

3.7.3 Diagnosis and therapeutic decision

Dermatologists

The study of dermatologists' performances requires considering several factors. As far as the diagnosis is concerned, sensitivity and specificity express the efficiency of the clinicians (on this limited database), but also the trade-off they believe acceptable with respect to the risk for a false diagnosis (table. 4). Depending on their level of confidence, they may privilege sensitivity over specificity. The opposite may also be true since it is a "risk-free" trial (Here, $d_{quality}$ can't be used to compare performances). The prior frequency of melanoma they meet usually in their daily practice may also play a role. Unfortunately, we cannot conclude on this point since it was not checked during the investigation. All these factors play a role for the therapeutic decision, although much smaller. As a matter of fact, we can expect the therapeutic decision to get a higher sensitivity (as far as the prediction of melanoma is concerned), together with a lower specificity, since the class of doubtful lesions worthy of an excision encompass melanoma.

Table 4. Diagnosis and therapeutic decision: Dermatologists' performances. Sensitivity and specificity are calculated with respect to the "gold standard" diagnosis.

Diagnosis and therapeutic decision	Sensitivity/Specificity (Diagnosis)	Sensitivity/Specificity (Therapeutic decision)
Dermatologist 1	0.62 / 0.90	0.84 / 0.63
Dermatologist 2	0.78 / 0.85	0.93 / 0.63
Dermatologist 3	0.59 / 0.71	0.84 / 0.39
Dermatologist 4	0.81 / 0.90	0.84 / 0.55
Dermatologist 5	0.71 / 0.80	0.87 / 0.63

Variability of performance is high, dermatologists with a melanoma-specific hospital activity having the best performance, both for the diagnosis and the therapeutic decision.

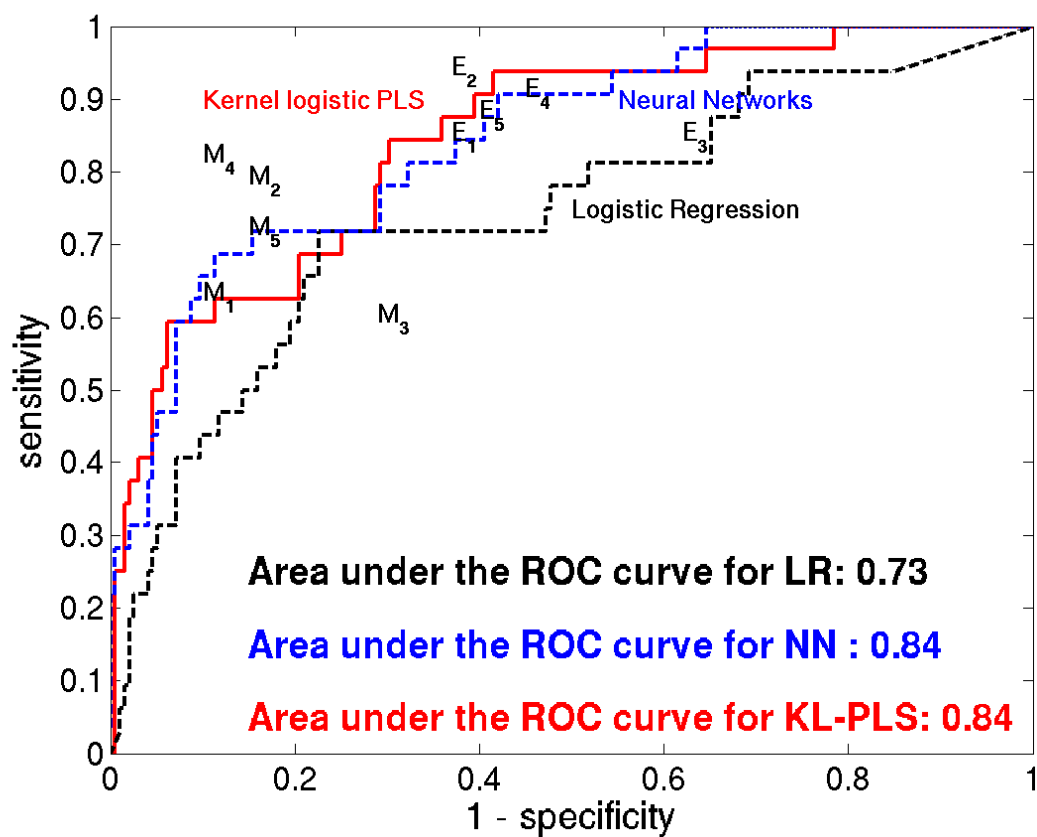
Combining dermatologists' diagnoses by means of a voting scheme allows evaluating the efficiency of the group of experts. The result is remarkable, considering in particular that 31 out of 32 melanomas are detected (sensitivity = 0.97) while the cost remains low (specificity = 0.60). It appears that several opinions are better than one. Such a result was not granted, given the false positives that also cumulate within the voting scheme.

Classifiers

The classifiers were built using a 47-dimensional vector as input and the gold standard diagnosis as output. Under appropriate conditions, KL-PLS (Gaussian kernel parameter = $1e-3$, number of retained KL-PLS components = 3), neural network (number of RBF neurons in each class: 7 for class 1 (grouping benign and dysplastic) and 5 for class 2 (melanoma)) and the logistic regression provide, for each tested lesion, a probability to

be a melanoma. ROC curves are built from these probabilities (figure 7). KL-PLS and RBF have comparable performances with respect to dermatologists, whatever the criteria. Area under the ROC curves is a measure of the quality of prediction. It can be seen from the AUC value (0.84 for both for KL-PLS and RBF) that these classifiers are effective over the whole range of tumors. As examples for KL-PLS, there are only 6% (respectively 2%) false positive in the 60% (respectively 40%) first naevi diagnosed as melanoma.

It is worth pointing out that RBF performances severely degrade as the input space dimension grows. We strongly believe that KL-PLS models will outperform RBFs on high-dimensional data especially when new parameters come into interest. Incidentally, the fine-tuning of RBFs is not straightforward.



F

figure 7. Roc Curves of KL-PLS, Neural Networks and Logistic Regression and corresponding AUCs (areas under the curve)

By virtue of the KL-PLS approach, the number of parameters is reduced from 47 (the size of the input vector) to 3. These three new parameters, which are a complex combination of the initial parameters, sum up the discriminating information within the input vector. Similarities between lesions can thus be observed in 3-D plot. Alternatively, the 3-D space spanned by the three KL-PLS components can be projected into a 2-D plot (using a Linear Discriminant Analysis (LDA) applied on the 3 KL-PLS components and the 3 classes (benign, dysplastic, melanoma)) allowing to find an “optimal angle of view”

from a class separability point of view. This strategy allows to observe similarities with neighboring lesions in the database and to visually justify the classification decision (Figure 7).

It is worth noticing that the benign naevi spread over a large sector of the plot, indicating that they display rather different patterns. In contrast, dysplastic tumors are more concentrated: they display more similar patterns, as a likely consequence of their clinical restricted definition. Melanoma can be very different one from another: they appear on several different locations in the 2D plot and partly overlap with other classes. However, it must be kept in mind that the classification is performed in a 3D space where overlapping is reduced.

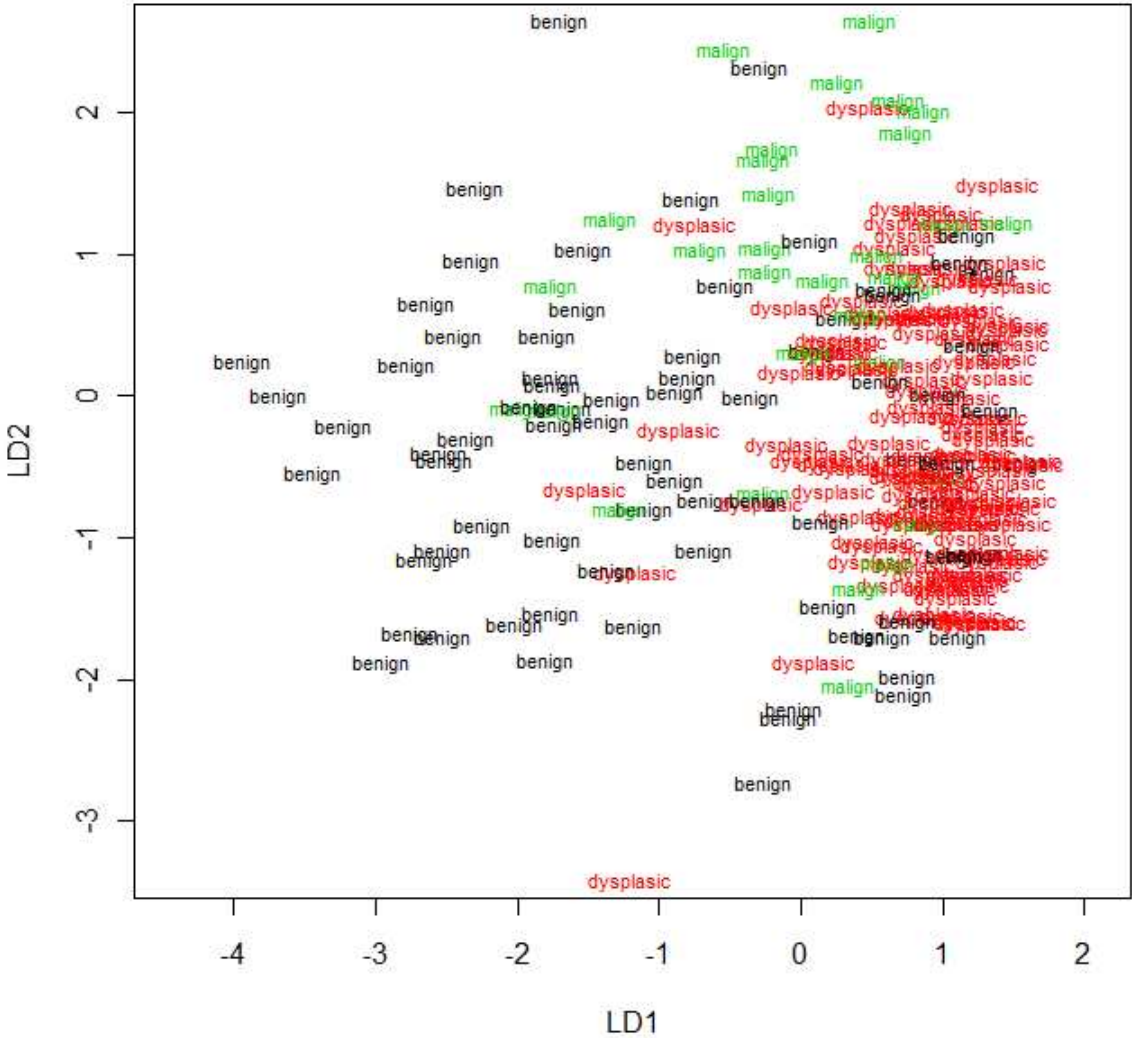


Figure 8. Projection of the database onto the two Linear Discriminant Axis (see text).

4 Discussion and conclusion

In this paper, a new system for analyzing digital images of skin lesions has been presented. The proposed approach allows, exclusively by statistical learning techniques, testing and evaluating the discriminating ability of several features for the recognition of melanoma. We acknowledge that the basic quality requirements suggested by Rosado (1) are essential to the prediction of the value of such computer systems under real world conditions. The project was directed according to Rosado's requirements: in particular, the validation protocol (using independent "new tumors") and an extensive comparison with human diagnosis were carefully conducted. For sake of robustness and flexibility, the system has been specifically designed for the analysis of digital images of pigmented skin lesions acquired under "uncontrolled" conditions. Variability of lesions, shapes, colors, textures, variability of structures (pigmented network, globules...), presence of artifacts (hairs, shadow...), and finally variability in the quality of images are usually considered factors making the task of an automatic classifier difficult. It is remarkable that dermatologists can get rid of these disturbing factors and give good agreement for diagnoses of tumors in our database. We believe that the clinical experience (based on the learning by sample paradigm) they gain during their daily practice is the key to their success. Similarly, by using powerful statistical learning techniques, we have been able to successfully analyze a great variability of tumors. Results presented here are very promising. Sensitivity and specificity of the computer-based diagnosis of melanoma does not greatly differ from the diagnostic capability of the five dermatologists. Currently, KL-PLS gives an estimate of the probability that a lesion belongs to each one of the two classes (melanoma versus non melanoma lesions). Although therapeutic decision was not the target during the learning phase, the classifier also performs properly with respect to this endpoint as it reaches dermatologists' levels in the therapeutic decision area. This property supports the existence of a continuum of malignity for naevi, ranging from benign lesion to atypic lesion and finally to melanoma.

It is interesting to note that essentially some features linked to the ABCD rules are selected for this study but our results show that the diagnosis obtained from such a limited set is already worth consideration (the KL-PLS based classifier reaches performances useful for a screening campaign (sensitivity: 95% and specificity: 60%)). However, it must be pointed out that many other parameters might be added thanks to KL-PLS, as this classifier is specifically designed for the management of high dimensional data. In particular, texture information such as pigmented network, thickness of the tumor, the Argenziano's 7-point check list (27), ... and finally clinical data ought to be considered.

Our study specifically shows that an automatic learning based system can reach the part of dermatologist' expertise that deal with image examination. The inclusion of some other descriptors is clearly required in order to improve performances. In particular some technical developments may be realized for the preprocessing of the images. As noted by Schmid-Saugeon, for example, asymmetry is still an open question (12). It

should be noted that in clinical practice dermatologists have at their disposal other information such as patient's age, skin type, localization. We expect that, by using this kind of information, dermatologist's and automated system performances should further improve. It is worth pointing out that KL-PLS can manage missing data, which is an important factor when clinical data is of concern.

The relatively small number of tumors analyzed in this study may be considered as a drawback for the system that was developed. It is obvious that a maximum of examples are welcome to reach full expertise. We have seen that the number of samples is not a limiting factor for KL-PLS. As a matter of fact, we are currently working with a group of dermatologists to increase the size of the database. But the validation protocol used in this work (leave one out) already shows that at least for current cases, the diagnosis is reliable. Such a result would not have been obtained if the database has not been representative, at least partly, of the diversity of the tumors encountered in daily practice.

The KL-PLS algorithm provides an original insight into the clinical knowledge of pigmented skin lesion. It offers a measure of probability, some rules of classification (via the feature models) and nearest case visualization. Indeed, based on the construction of latent components, it is possible to display each lesion onto the first components and permit retrieval and visualization of the most similar cases. By mimicking the medical reasoning to a certain degree, our system allows dermatologists directly compare unknown cases with known skin lesions. It is clear that the effective deployment of the automatic diagnosis of melanoma in clinical practice may greatly benefit from the partial understanding of its internal mechanisms.

References

1. Rosado B, Menzies S, Harbauer A, Pehamberger H, Wolff K, Binder M, et al. Accuracy of Computer Diagnosis of Melanoma A Quantitative Meta-analysis. *Archives of Dermatology*. 2003; 139(3): 361-7.
2. Perrinaud A, Gaide O, French L, Saurat J, Marghoob A, Braun R. Can automated dermoscopy image analysis instruments provide added benefit for the dermatologist? A study comparing the results of three systems. *British Journal of Dermatology*. 2007; 157(5): 926-33.
3. Guitera-Rovel P, Vestergaard M. [Diagnosis tools for cutaneous melanoma.]. *Ann Dermatol Venereol*, 2008: 828.
4. Hastie T, Tibshirani R, J. F. *The Element of Statistical Learning*. Ney York: Springer; 2001.
5. Bischof LM, Talbot H, Breen E, Lovell D, Chan D, Stone G, et al. Automated melanoma diagnosis system. *SPIE*, 1999: 130.
6. Ercal F, Chawla A, Lee H, Moss RH. Neural network diagnosis of malignant melanoma from color images. *IEEE transactions on biomedical engineering*. 1994; 41(9): 837-45.
7. Ganster H, Pinz P, Rohrer R, Wildling E, Binder M, Kittler H. Automated melanoma recognition. *IEEE Transactions on Medical Imaging*. 2001; 20(3): 233-9.
8. Hoffmann K, Gambichler T, Rick A, Kreutz M, Anschuetz M, Grunendick T, et al. Diagnostic and neural analysis of skin cancer (DANAOS). A multicentre study for collection and computer-aided analysis of data from pigmented skin lesions using digital dermoscopy. *British Journal of Dermatology*. 2003; 149(4): 801-9.
9. Kittler H, Pehamberger H, Wolff K, Binder M. Diagnostic accuracy of dermoscopy. *Lancet Oncol*. 2002; 3(3): 159-65.
10. Maglogiannis IG, Zafiropoulos EP. Characterization of digital medical images utilizing support vector machines. *feedback*. 2005.
11. Sboner A, Eccher C, Blanzieri E, Bauer P, Cristofolini M, Zumiani G, et al. A multiple classifier system for early melanoma diagnosis. *Artificial Intelligence In Medicine*. 2003; 27(1): 29-44.
12. Schmid-Saugeon P, Guillod J, Thiran JP. Towards a computer-aided diagnosis system for pigmented skin lesions. *Computerized Medical Imaging and Graphics*. 2003; 27(1): 65-78.
13. She Z, Liu Y, Damatoa A. Combination of features from skin pattern and ABCD analysis for lesion classification. *Skin Research and Technology*. 2007; 13(1): 25-33.
14. Stanley RJ, Stoecker WV, Moss RH. A relative color approach to color discrimination for malignant melanoma detection in dermoscopy images. *Skin Research and Technology*. 2007; 13(1): 62-72.
15. Tomatis S, Carrara M, Bono A, Bartoli C, Lualdi M, Tragni G, et al. Automated melanoma detection with a novel multispectral imaging system: results of a prospective study. *Physics in Medicine and Biology*. 2005; 50(8): 1675-87.
16. Do Hyun Chung Sapiro G. Segmenting skin lesions with partial-differential-equations-based image processing algorithms. *IEEE transactions on Medical Imaging*. 2000; 19(7): 763-7.
17. Emre Celebi M, Kingravi HA, Iyatomi H, Alp Aslandogan Y, Stoecker WV, Moss RH, et al. Border detection in dermoscopy images using statistical region merging. *Skin Research and Technology*. 2008; (0).
18. Erkol B, Moss RH, Joe Stanley R, Stoecker WV, Hvatum E. Automatic lesion boundary detection in dermoscopy images using gradient vector flow snakes. *Skin Research and Technology*. 2005; 11(1): 17-26.
19. Xu L, Jackowski M, Goshtasby A, Roseman D, Bines S, Yu C, et al. Segmentation of skin cancer images. *Image and Vision Computing*. 1999; 17(1): 65-74.
20. Zhang Z, Stoecker WV, Moss RH. Border detection on digitized skin tumor images. *IEEE Transactions on Medical Imaging*. 2000; 19(11): 1128-43.

21. Cheng Y, Swamisai R, Umbaugh SE, Moss RH, Stoecker WV, Teegala S, et al. Skin lesion classification using relative color features. *Skin Research and Technology*. 2008; 14(1): 53-64.
22. Seidenari S, Pellacani G, Grana C. Colors in atypical nevi: a computer description reproducing clinical assessment. *Skin Research and Technology*. 2005; 11(1): 36-41.
23. Tenenhaus A, Giron A, Viennet E, Béra M, Saporta G, Fertil B. Kernel logistic PLS: A tool for supervised nonlinear dimensionality reduction and binary classification. *Computational Statistics and Data Analysis*. 2007; 51(9): 4083-100.
24. Schölkopf B, Smola AJ. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press; 2002.
25. Bastien P, Vinzi VE, Tenenhaus M. PLS generalised linear regression. *Computational Statistics and Data Analysis*. 2005; 48(1): 17-46.
26. Menzies SW, Crotty KA, McCarthy WH, Ingvar C. *An Atlas of Surface Microscopy of Pigmented Skin Lesions*. McGraw-Hill Professional Publishing; 1996.
27. Argenziano G, Fabbrocini G, Carli P, De Giorgi V, Sammarco E, Delfino M. 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*. 1998; 134(12): 1563-70.