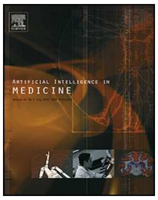
Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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



Full length Article

Performance of a dermoscopy-based computer vision system for the diagnosis of pigmented skin lesions compared with visual evaluation by experienced dermatologists

Maciel Zortea[a,*], Thomas R. Schopf[b], Kevin Thon[b], Marc Geilhufe[a], Kristian Hindberg[a], Herbert Kirchesch[c], Kajsa Møllersen[b], Jörn Schulz[a], Stein Olav Skrøvseth[b], Fred Godtliebsen[a]

[a] Department of Mathematics and Statistics, University of Tromsø, 9037 Tromsø, Norway
[b] Norwegian Centre for Integrated Care and Telemedicine, University Hospital of North Norway, 9038 Tromsø, Norway
[c] Dermatology Office, Venloer Straße 107, 50259 Pulheim, Germany

ARTICLE INFO

Article history:
Received 1 July 2013
Received in revised form
28 November 2013
Accepted 29 November 2013

Keywords:
Computer-aided diagnosis
Supervised classification
Skin cancer detection
Melanoma
Pigmented skin lesions
Dermoscopy

ABSTRACT

Background: It is often difficult to differentiate early melanomas from benign melanocytic nevi even by expert dermatologists, and the task is even more challenging for primary care physicians untrained in dermatology and dermoscopy. A computer system can provide an objective and quantitative evaluation of skin lesions, reducing subjectivity in the diagnosis.
Objective: Our objective is to make a low-cost computer aided diagnostic tool applicable in primary care based on a consumer grade camera with attached dermatoscope, and compare its performance to that of experienced dermatologists.
Methods and materials: We propose several new image-derived features computed from automatically segmented dermoscopic pictures. These are related to the asymmetry, color, border, geometry, and texture of skin lesions. The diagnostic accuracy of the system is compared with that of three dermatologists.
Results: With a data set of 206 skin lesions, 169 benign and 37 melanomas, the classifier was able to provide competitive sensitivity (86%) and specificity (52%) scores compared with the sensitivity (85%) and specificity (48%) of the most accurate dermatologist using only dermoscopic images.
Conclusion: We show that simple statistical classifiers can be trained to provide a recommendation on whether a pigmented skin lesion requires biopsy to exclude skin cancer with a performance that is comparable to and exceeds that of experienced dermatologists.

1. Introduction

The incidence and mortality rates of melanoma in Caucasian populations have been increasing for many decades [1]. Though only accounting for one tenth of the new cases of skin cancer, melanoma are associated with more than 90% of the skin cancer deaths [2]. If detected at an early stage, the prognosis for the patient is excellent because the patient can be cured by simple excision of the tumor. However, early diagnosis is very challenging as melanomas are easily confused with benign skin lesions.

Dermoscopy is a method that allows doctors to examine structures in the skin that are not visible to the naked eye. When practiced by experts, dermoscopy improves the diagnostic accuracy of pigmented skin lesions (PSL) [3–5]. Several methods have been developed to help clinicians interpret the structures revealed through dermoscopy [6]. Well known algorithms include the ABCD rule of dermoscopy [7], the Menzies method [8], the Three-point checklist [9], the 7-point checklist [10], the CASH algorithm for dermoscopy [11], the Chaos and Clues algorithm [12], the BLINCK algorithm [13], and Pattern Analysis [14]. However, intensive and time consuming training is required to become an expert in dermoscopy. Furthermore, dermoscopy has its limitations, especially in the diagnosis of early melanoma [15]. In the early stages of the disease it may look like a common mole. Often there are no specific dermatoscopic features of melanoma, or the features appear subtle and are easily overlooked.

There have been efforts to develop computer programs to diagnose melanoma based on lesion images. Roughly, these studies follow intuitive steps in a standard pattern recognition processing chain: (a) image segmentation to separate the lesion area from the background skin, (b) extraction of image features for classification purposes, and (c) final classification using statistical methods.

* Corresponding author. Tel.: +47 45683608.
E-mail addresses: mzortea@gmail.com, maciel.zortea@uit.no (M. Zortea).

0933-3657/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.artmed.2013.11.006

A wide range of ideas have been utilized in these three steps; see Korotkov and Garcia [16] for an overview and categorizations. Reporting sensitivity and specificity, Rosado et al. [17] presented a thorough overview of state-of-the-art methods at the time, and Dolianitis et al. [18] compared the diagnostic accuracy of four dermoscopy algorithms in the hands of 61 relatively inexpert medical practitioners in Australia. Day and Barbour [19] attempted to reproduce algorithmically the perceptions of dermatologists as to whether a lesion should be excised or not. Comparing results between different systems is difficult because results are very sensitive to the data set used for validation, and a major problem is the lack of publicly available databases of dermoscopic images. For a fair and representative comparison, a data set with a large number of examples of all types of lesions and all types of features expected to be encountered in clinical practice should be made available.

Following this, an interesting research question is:

> Assume identical information is made available to both computers and doctors for the same set of skin lesion images. Then, how does the accuracy of the computer system compare with the accuracy of the doctors?

An answer to the above question would make it easier to objectively assess the performance of any new or existing methods, and would provide an indication of how difficult the lesion images in the data sets used in the experiments were to diagnose. In a data set with a clear distinction between classes, high accuracy would be no surprise. Despite this being a conceptually rather simple experiment to conduct, the study could be demanding because it would require substantial effort by dermatologists to evaluate a large number of lesion images. Also, a more difficult question to answer is whether the data set is sufficiently representative. To be so, it needs to approximate the variability of cases found in a true clinical setting, including the prior information regarding the occurrence of each type of lesion.

Several studies have been reported where the diagnostic accuracy of a computer system is directly compared with human diagnosis. The diagnostic accuracy of the computer systems is generally not significantly different from that of human experts (for an overview, see Rosado et al. [17]). Most studies tend to compare the performance of their system exclusively with histopathological diagnosis, leaving it an open question how difficult the lesions are to diagnose by dermatologists. In the present study, in addition to the histopathological results, we compare the results of the computer system with those of three dermatologists to provide an indication of how challenging our dataset is to either type of analysis.

Korotkov and Garcia [16] recently listed 10 commercial computer-aided diagnosis (CAD) systems for the diagnosis of melanoma based on dermoscopy. As a rule they use powerful and dedicated video cameras. The cost related to the acquisition material and proprietary technologies are likely substantial barriers to the systems gaining widespread popularity among physicians [20]. Perrinaud et al. [21] reported on an independent clinical evaluation of some of these systems, and they found little evidence that such systems benefit dermatologists. Also, current limitations of state-of-the-art CAD systems motivate the development of new algorithms for analysis of skin lesions, and low cost data acquisition tools (e.g., digital cameras and dermatoscopes) are becoming commonly available. Following an approach that should be practical and intuitive to dermatologists, the images considered in this study are acquired by means of a consumer-grade digital camera with a dermatoscope attached. This simple image acquisition setup has been previously discussed, for instance in Gewirtzman and Braun [22], and has been used in the visual comparison system of Baldi et al. [23].

In the following sections we report on our experiments with a simple system we built using off-the-shelf equipment for fully automatic detection of melanomas. Our main contributions include the development of novel image features with the potential to handle morphological structures in dermoscopic images acquired with low-cost, off-the-shelf equipment. Also, we give an indication on how challenging the dataset is by asking three dermatologists to evaluate the same set of images evaluated by the computer system.

The remainder of the paper is organized as follows. Section 2 provides a description of the data, the image segmentation method, feature extraction and selection, and classification. Section 3 describes two experiments and reports on the findings. Section 4 provides a discussion, and conclusions are drawn in Section 5.

2. Materials and methods

2.1. Data

Dermoscopic images of 206 pigmented skin lesions were acquired using a portable dermatoscope (Dermlite Pro II HR, 3Gen LLC, CA, USA) attached to a consumer-grade digital camera (Canon G10, Canon Inc., Tokyo, Japan). Images were acquired at two locations, a private practice clinic in Germany by author H.K., and at the Department of Dermatology at the University Hospital of North Norway, in Tromsø, Norway by author T.R.S. 113 images were obtained consecutively at the private clinic between December 2009 and January 2010 from all patients requiring biopsy or excision of a pigmented skin lesion because of diagnostic uncertainty. In addition, we added 93 images photographed at both sites between December 2009 and December 2010. Of these, 60 images represented benign common lesions not requiring biopsy or excision and 33 images of melanomas. A total number of 206 lesion images were decided on because this number appeared realistic regarding the workloads of the three independent dermatologists participating in the evaluation of this study. All images have the same fixed values for aperture width, ISO value, focus distance, and focal length. The resulting raw images of size 4432 × 3326 pixels were raw-converted to.png images using a fixed white balance. Due to the presence of the dermatoscope, only a circular area with a diameter of 3326 pixels contains image information, which corresponds to about 14 mm on the skin. The typical dark circular fading pattern at the border of the image area due to the geometry of the dermatoscope is discarded. The resulting image is downsampled to 1650 × 1650 pixels to ease computations, resulting in a spatial resolution of approximately 3000 ppi, with a color depth of eight bits per channel.

The set of images was printed on high quality 178 mm × 178 mm paper sheets with a resolution higher than 200 dpi. The printed images were given to three dermatologists familiar with dermoscopy and who were not otherwise involved in the data collection. Two doctors were board certified dermatologists while the third doctor was an experienced resident doctor. They were asked to provide, for each case, (a) the probable diagnostic class (on a visual analog scale indicating benign, suspicious, or malignant), and (b) an indication regarding whether they would recommend excision of the skin lesion. For the purpose of our experiments, we will compare the outcome of the computer method with the doctors' excision recommendations (yes, no). No additional information was provided to the dermatologists (patient's gender, age, etc.). The participants noted their answers on the sheets and had no time constraints. In Table 1 the characteristics of the lesions used in the study are summarized. Notably, the Breslow depth is less than 1 mm in all cases except three, where a Breslow depth of <1 mm indicates early stage melanoma. Pigmented Bowen's disease and

Table 1
Histopathologic findings for lesions used in the study with Breslow tumor depths for nodular and superficial spreading melanomas.

Benign/other‡ lesions	169
Melanocytic nevi	154
Seborrheic keratoses	10
‡Pigmented Bowen's disease	3
Sarcoidal granuloma	1
‡Basal cell carcinoma	1
Malignant melanoma	37
Superficial spreading melanoma	13
In situ melanoma	10
Lentigo maligna	6
Nodular melanoma	2
Melanoma metastasis	2
Undetermined	4
Breslow tumor depth	
Median (mm)	0.61
Interquartile range (mm)	0.33–0.825

basal cell carcinoma are examples of malignant non-melanoma skin cancers. As discussed in Section 3, this set of lesions is very difficult to classify.

The large amount of benign lesions compared to malignant lesions allows us to split the benign class into two subclasses. A certain degree of subjectivity is intrinsic, but to reduce bias, this was done by an independent dermatologist (T.R.S, one of the authors), who was not involved in the accuracy assessment performed by the other three independent dermatologists. Based on the dermoscopic images, out of the 169 benign lesions, 89 were labeled benign "not cut" (i.e., representing a complete benign appearance) and 80 benign "cut" (i.e., displaying an equivocal appearance, where "cut" simply means recommending the lesion to be excised because malignancy cannot be ruled out). Given the low number of malignant lesions available (37), this group was not further split into subclasses. It was therefore viewed as a malignant "cut" class, containing both early stage and also more developed melanomas. The low sensitivity scores achieved by two of the three dermatologists (shown in Section 3) suggest that the malignant class was very challenging.

Fig. 1(a) and (b) shows examples of skin lesions where all three doctors agree and provide the correct excision recommendation according to the histopathological diagnosis. Fig. 1(c) shows a case with agreement between the doctors but with the incorrect diagnosis. The skin lesion is malignant, but all doctors diagnose it as benign and do not recommend excision. Two doctors label the benign lesion in (d) as suspicious and recommend excision. One doctor concludes it is benign and that it should not be excised.

2.2. Image segmentation

The segmentation of a lesion image is performed using a skin lesion segmentation algorithm developed in-house [24]. This fully automatic algorithm selects small seed regions likely to correspond to samples of skin and the lesion. The seed regions are used as initial training samples, and the lesion segmentation problem is treated in an iterative binary classification setting. The features of interest are calculated from the segmented lesion.

2.3. Feature extraction

After automatic segmentation, a set of 53 image-derived features are computed. First, these descriptive features attempt to quantify the asymmetry, the borders, and the colors of the lesions. These are among the anatomical attributes that dermatologists acknowledge are important for diagnosing melanomas. In addition, other generic quantitative attributes, easily computed by image processing algorithms, such as geometry and texture of the lesion, are included. Most of the features described here are developed in-house; for instance, the analysis of variance for the lesion border, geometric features, and our choice of textures. The color and shape related features are inspired by previous studies in the literature (e.g., [25,26,16], among others). The details of the features follow.

2.3.1. Asymmetry of shape (2 features)

Fig. 2(a) is a schematic representation of the binary mask of a lesion with a coordinate system centered on the center of mass. For these two features we work with gray-scale images. Denote $I_{i,j}$ as the gray-scale level of pixel (i, j), where the first index is along the first dimension of Fig. 2(a). Set $I_{i,j} = 0$ if (i, j) is outside the binary mask of Fig. 2(a). We now compare the following regions:

1 $(A_1 \cup A_2)$ versus $(A_3 \cup A_4)$
2 $(A_1 \cup A_4)$ versus $(A_2 \cup A_3)$

For these two combinations, we evaluate

$$\Delta S_1 = \sum_i \sum_{j>0} |I_{i,j} - I_{i,-j}| \quad \text{and} \quad \Delta S_2 = \sum_{i>0} \sum_j |I_{i,j} - I_{-i,j}|. \tag{1}$$

We divide ΔS_1 and ΔS_2 by the area of the binary mask containing the lesion, so that the scores for lesions of different sizes are easily comparable. A large value of ΔS_1 and/or ΔS_2 indicates that there is a strong asymmetry of shape. The symmetry axes are rotated in steps of 10°. We retain the rotation with the lowest average scores of ΔS_1 and ΔS_2. For the retained axes, we sort the scores of the two orthogonal axes, so that these will correspond to the asymmetry of shape features (f_1, f_2). Examples are shown in Fig. 3(a)–(c). Note that Eq. 1 calculates the relative differences of color intensities (more precisely gray scale values) of the lesion. Therefore, this is not strictly an asymmetry description feature, but serves as one.

2.3.2. Asymmetry of color intensity, computed in gray-scale (2 features)

The computation of color asymmetry is similar to the computation of shape asymmetry in Eq. (1). For the two combinations of regions, we evaluate

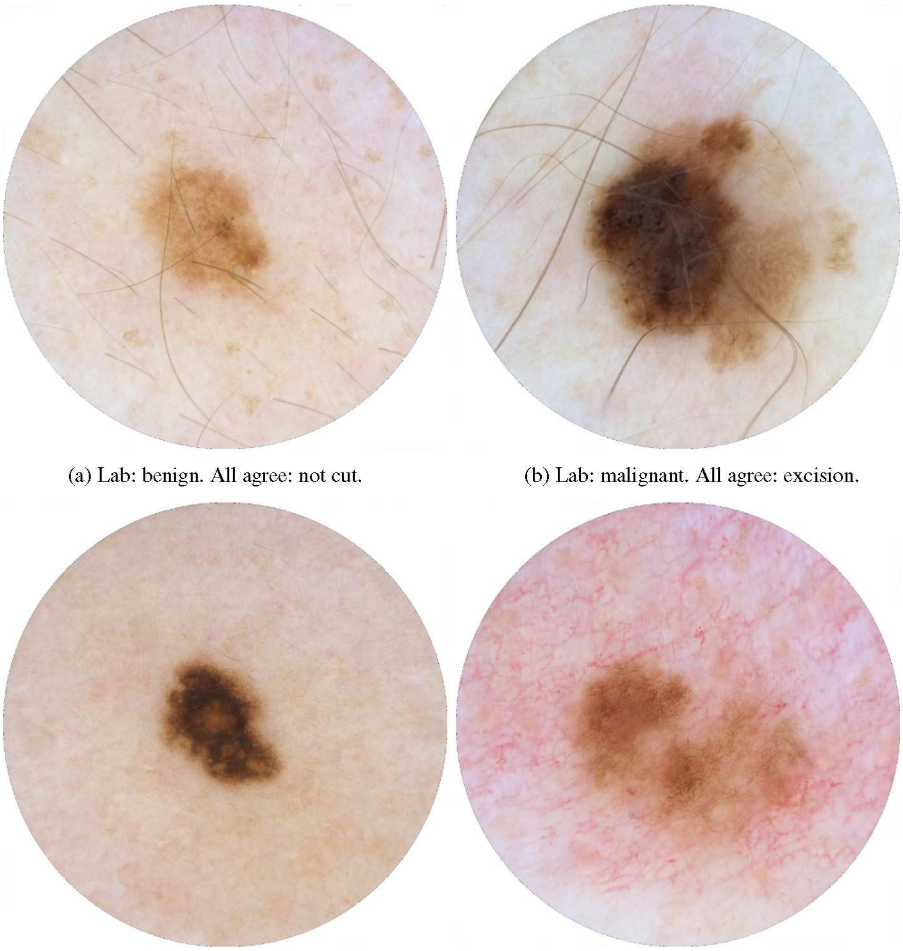
$$\begin{aligned} \Delta C_1 &= \sum_{a=0}^{255} |\widehat{C}_{A_1 \cup A_2}(a) - \widehat{C}_{A_3 \cup A_4}(a)|, \\ \Delta C_2 &= \sum_{a=0}^{255} |\widehat{C}_{A_1 \cup A_4}(a) - \widehat{C}_{A_2 \cup A_3}(a)|, \end{aligned} \tag{2}$$

where, for instance, $\widehat{C}_{A_1 \cup A_2}$ is the estimated distribution of the 256 gray-scales a using pixels belonging to either region A_1 or A_2, and is computed using Gaussian kernel density estimation [27]. Large values of ΔC_1 and/or ΔC_2 indicate that there is a strong asymmetry between the domains compared. The symmetry axes are rotated in steps of 10°. We retain the rotation with the lowest average scores of ΔC_1 and ΔC_2. For the retained axes, we sort the scores for the two orthogonal axes that correspond to our proposed asymmetry of color features (f_3, f_4). Examples are shown in Fig. 3(d)–(f).

2.3.3. Asymmetry of color-shape (2 features)

For these features we first compute the center of mass (x, y) of the binary mask of the lesion. Second, a set of alternative binary masks, with mass centers (x_t, y_t), is generated by applying a threshold to the gray-scale values inside the lesion border at percentiles $t = [0.10, 0.20, \ldots, 0.90]$. Next we compute a vector v whose elements are the Euclidean distances between (x, y) and the different (x_t, y_t). To reduce image scale effects, v is normalized by a lesion-dependent constant given as the radius of a circle of equivalent area as the original binary mask. The features f_5 and f_6 are the mean and

Please cite this article in press as: Zortea M, et al. Performance of a dermoscopy-based computer vision system for the diagnosis of pigmented skin lesions compared with visual evaluation by experienced dermatologists. Artif Intell Med (2013), http://dx.doi.org/10.1016/j.artmed.2013.11.006



(a) Lab: benign. All agree: not cut.

(b) Lab: malignant. All agree: excision.

(c) Lab: malignant. Mistake by all: not cut

(d) Lab: benign. Two recommend excision.

Fig. 1. Examples of feedback from dermatologists, showing cases where all, none, and some of the three doctors give the correct recommendation.

standard deviation of v, respectively. Examples of mass center trajectories are shown in Fig. 4. Large displacements of the center of mass indicate an asymmetric lesion.

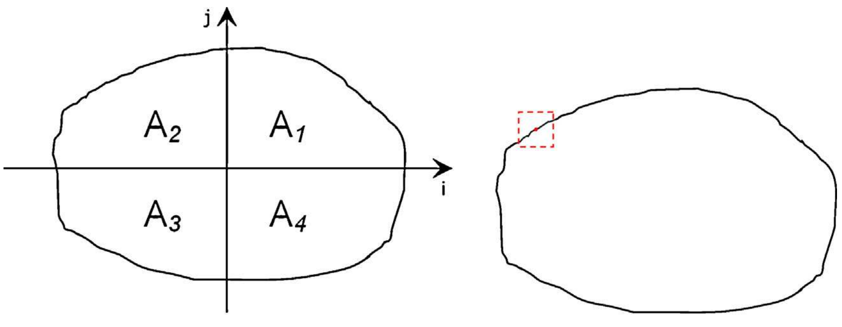
2.3.4. Border, ANOVA-based analysis (3 features)

Suppose we have a segmented lesion such as the one in Fig. 2(b). For a particular region around the border pixel k, we have the pixels $X_{11}, X_{21}, \ldots, X_{n_1 1}$ inside the skin lesion and $X_{12}, X_{22}, \ldots, X_{n_2 2}$ outside the skin lesion, where X_{ij} is the gray-scale observation number i in tissue type $j = 1, 2$. The standard analysis of variance (ANOVA) then yields

$$\sum_{j=1}^{2}\sum_{i=1}^{n_j}(X_{ij}-\overline{X})^2=\sum_{j=1}^{2}\sum_{i=1}^{n_j}(X_{ij}-\overline{X}_j)^2+\sum_{j=1}^{2}n_j(\overline{X}_j-\overline{X})^2, \qquad (3)$$

where

$$\overline{X}=\frac{1}{n_1+n_2}\sum_{j=1}^{2}\sum_{i=1}^{n_j}X_{ij} \quad \textit{and} \quad \overline{X}_j=\frac{1}{n_j}\sum_{i=1}^{n_j}X_{ij}. \qquad (4)$$

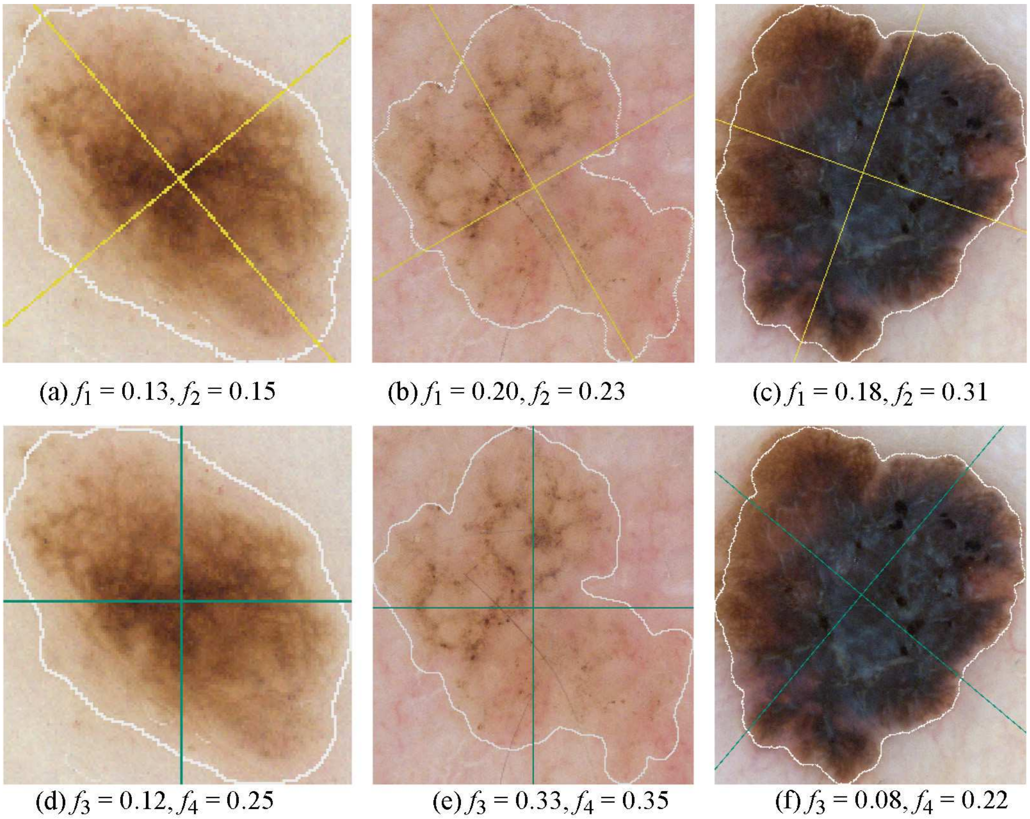


(a) Asymmetry analysis

(b) Border analysis

Fig. 2. (a) Lesion with reference regions A_m, $m = 1, \ldots, 4$. (b) Lesion border with a dotted square region centered on a border pixel.

Please cite this article in press as: Zortea M, et al. Performance of a dermoscopy-based computer vision system for the diagnosis of pigmented skin lesions compared with visual evaluation by experienced dermatologists. Artif Intell Med (2013), http://dx.doi.org/10.1016/j.artmed.2013.11.006


(a) $f_1 = 0.13, f_2 = 0.15$ (b) $f_1 = 0.20, f_2 = 0.23$ (c) $f_1 = 0.18, f_2 = 0.31$

(d) $f_3 = 0.12, f_4 = 0.25$ (e) $f_3 = 0.33, f_4 = 0.35$ (f) $f_3 = 0.08, f_4 = 0.22$

Fig. 3. Upper row, from left to right: asymmetry of shape axes (in yellow) of one benign and two malignant lesions. Lower row: Asymmetry of color axes (in green) of the same lesions as in the upper row. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

Eq. (3) is frequently written as

$$SS_T(k) = SS_E(k) + SS_R(k) \quad \Rightarrow \quad R_k \equiv \frac{SS_E(k)}{SS_T(k)}. \tag{5}$$

where $SS_T(k)$ is the total sum of squares of the pixels near the border, partitioned into two components related to the effects of the error $SS_E(k)$, and the pixel treatment $SS_R(k)$ (location inside/outside the lesion) in the model.

The above approach is implemented using a sliding window around the border, as illustrated in Fig. 2(b), using the gray-scale version of the image. A square region of size 61 × 61 pixels is centered at each border pixel. This is an empiric choice, and corresponds to about 0.50 mm of the skin surface. The statistics are computed using the pixels inside and outside the lesion border that are contained within the sliding window. In general, for each pixel $k = 1, \ldots, K$ at the border of the lesion, we calculate R_k. Now, by observing the distribution of $\{R_k\}_{k=1}^{K}$, values close to a value of one represent vague differences between the skin lesion area and the skin. For clear differences, values should be close to zero. To distinguish among them we find a set of percentiles for $\{R_k\}$, specifically

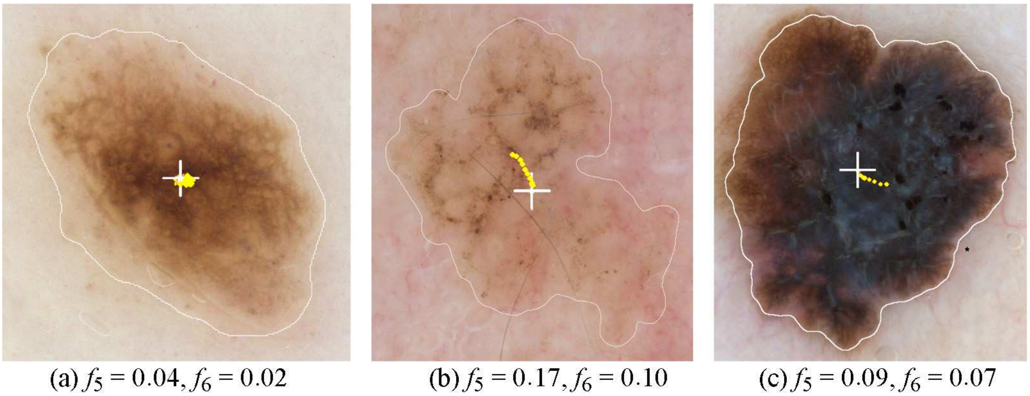
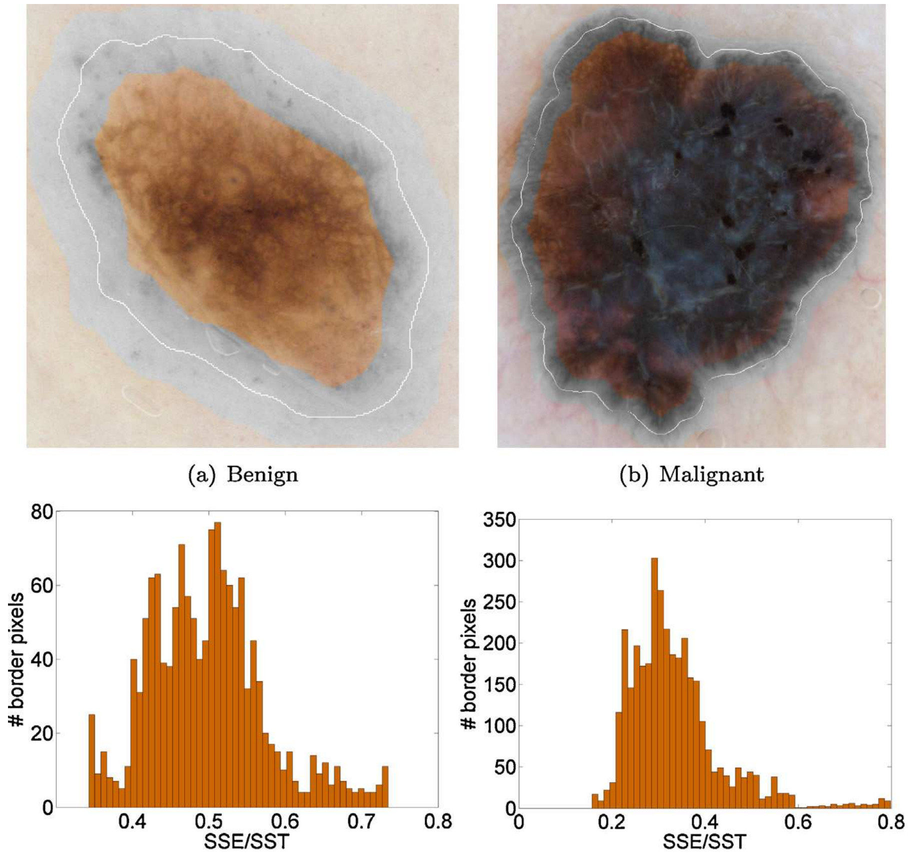

(a) $f_5 = 0.04, f_6 = 0.02$ (b) $f_5 = 0.17, f_6 = 0.10$ (c) $f_5 = 0.09, f_6 = 0.07$

Fig. 4. Output of the asymmetry analysis of color-shape features for the same images as in Fig. 3. For each subfigure, the cross marks the original center of mass, while the dots indicate how the center of mass changes position when thresholds are applied to the gray-scale values of the image, as described for the color-shape features f_5 and f_6.



(a) Benign

(b) Malignant

(c) $\{f_7, f_8, f_9\} = \{0.44, 0.50, 0.54\}$

(d) $\{f_7, f_8, f_9\} = \{0.27, 0.31, 0.37\}$

Fig. 5. ANOVA-based border features. The gray region around the lesion border (white contour) in (a) and (b) indicate the samples used to derive the ANOVA-based border features in (c) and (d), respectively. While the color fading in the benign case is very smooth around the border, the malignant case has a much more abrupt color change across the border. This is summarized by the 25th, 50th, 75th percentiles, corresponding to features f_7, f_8, and f_9, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

Γ_α for $\alpha = \{0.25, 0.5, 0.75\}$ where Γ_α is defined from $P(R > \Gamma_\alpha) = \alpha$. Our proposed features are the 25th (f_7), 50th (f_8), 75th (f_9) percentiles. Fig. 5 shows two examples of the suggested features.

2.3.5. Colors, full lesion (3 features)

The RGB pixel values of the lesion are converted to the CIE L^* a^* b^* color space [28], designed to be perceptually uniform, meaning that long (short) distances within the color space should correspond to large (small) perceived distances between colors. The 3D histogram based on the L, a, and b coordinates is computed using ten thousand randomly selected pixels. The bin size is set to $n = 2$. Only the non-empty bins are considered for the score computation. We use the average number of samples in each bin (f_{10}), the variance (f_{11}), and the percentage of non-empty bins in the color space (f_{12}). The L component values range from 0 to 100, while the a and b components vary between −127 and 127. Fig. 6 shows the distribution of the three L^* a^* b^* components for two example images.

2.3.6. Colors, peripheral versus central (6 features)

The lesion area is divided into the inner and the outer part separated by an "internal border", indicated by the dashed lines in the two upper images of Fig. 6. This border is found by iteratively shrinking the original border. For each iteration, all border points are moved the same number of pixels towards the center of mass. This is done until the outer/inner regions contain 30%/70% of the original pixels, respectively. We compute the mean value of the three L^* a^* b^* components in the inside and outside sets, and take the difference between them. These are the f_{13}, f_{14}, f_{15} features for the L^* a^* b^* channels. Similarly, we compute the probability density estimate of the samples of each L^* a^* b^* channel in the innermost and outermost parts of the regions. The density estimate is based on a normal kernel function, using a window parameter (bandwidth) that is a function of the number of points in the regions [29]. For each channel, we compute the overlapping area of the densities. The resulting features for the L, a, and b channels are referred to as f_{16}, f_{17}, f_{18}, respectively.

2.3.7. Colors, palette approach (2 features)

A palette approach, reported to efficiently estimate the number of colors of pigmented skin lesions [25,26], is used for feature generation. The linear discriminant analysis (LDA) classifier [30] is trained on the colors white, red, light brown, dark brown, blue-gray, and black obtained from a training image, as shown in Fig. 7(a). These are the most relevant colors according to the ABCD rule of dermoscopy. In the ABCD scoring system, the probability of malignancy increases with the number of distinct colors present in the lesion. From these sample colors a statistical classifier is trained to recognize colors in unseen images. We classify the image into different regions of colors and store the number of distinct colors as feature f_{19}. A color must cover at least 1% of the lesion to be counted. In addition, we retain the percentage of the lesion area classified as blue-gray as feature f_{20}. This particular color was found to be a very good indicator of malignancy in the Menzies' scoring method, the seven-point checklist, pattern analysis, and the three-point checklist.

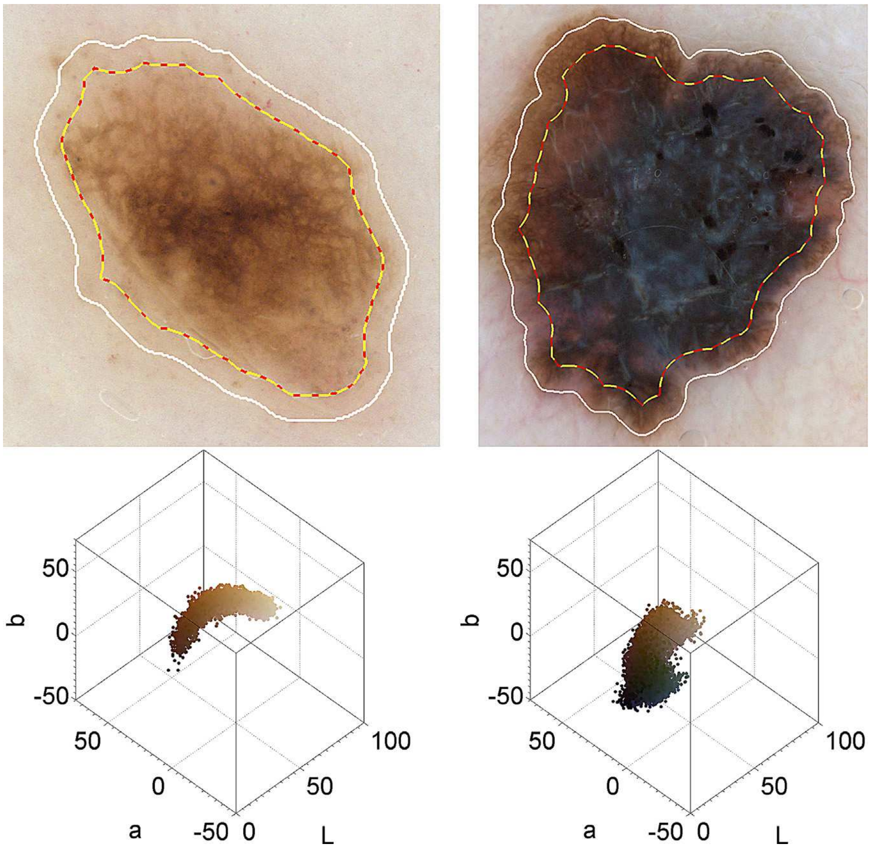
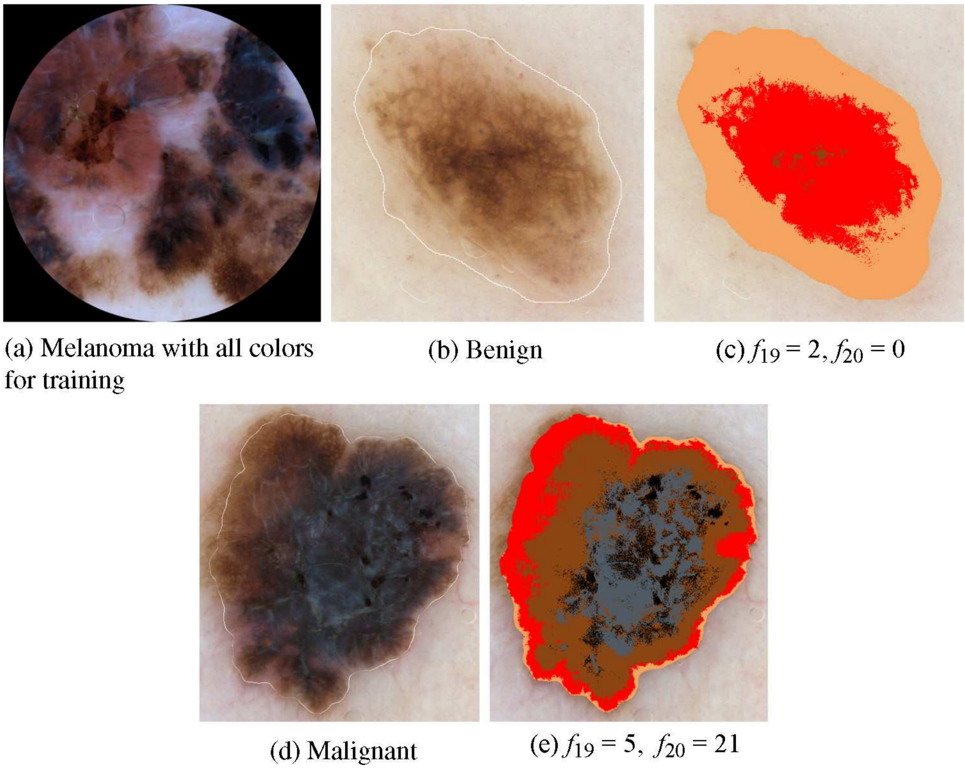


Fig. 6. Top left: benign. Top right: malignant. Lower row: 3D histograms of L^* a^* b^* color space components of top row.



(a) Melanoma with all colors for training

(b) Benign

(c) $f_{19} = 2, f_{20} = 0$

(d) Malignant

(e) $f_{19} = 5,\ f_{20} = 21$

Fig. 7. Output of the color counting palette approach. The number of colors is given as f_{19}, and f_{20} indicates the fraction of blue-gray lesion area. There are small spots classified as dark brown in the center of (c), but they cover less than 1% of the lesion. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

An example of classification is shown in Fig. 7. Arguably light and dark brown would be better classifications than red and light brown in Fig. 7(b) and (c), but the number of colors appears reasonable, and the five colors in Fig. 7(d) and (e) are well in line with our expectations.

2.3.8. Geometric (3 features)

We attempt to capture the lesion disorder by computing what we refer to as the number of “lesion pieces” resulting from applying binary thresholds to the gray-scale version of the lesion. The thresholds are applied at the 25th, 50th, and 75th percentiles of the gray-scale values of the skin lesion. To reduce noise, the number of pieces is computed after morphological opening using a disk element with a radius of five pixels that is applied to the binary masks obtained using each percentile. Call the resulting features f_{21}, f_{22}, f_{23}, respectively. Fig. 8 shows a benign and a malignant case, where the proposed scores are low and high, respectively.

2.3.9. Texture of the lesion (30 features)

Here we attempt to capture local spatial information in the skin lesions. We sample the segmented lesion using boxes of predefined size 41 × 41 pixels that are displaced around the image in a partially overlapping 20 pixels step to reduce computational requirements. For each box, a feature vector containing spatial descriptors that consist of image textures is computed. We use textures in an attempt to discriminate between some of the anatomical structures that dermatologists consider (e.g., the “D” part of the ABCD rule corresponds to the presence of up to five structural features: network, structureless (or homogeneous) areas, branched streaks, dots, and globules). Texture should at a minimum be invariant to rotation, and not very sensitive to acquisition issues. We focus on the use of uniform rotation invariant local binary pattern (LBP) histograms proposed by Ojala et al. [31], computed from the gray-scale version of the images. LBP is among the state-of-the-art methods for describing image textures, a powerful tool for rotation invariant texture analysis and robust in terms of gray-scale variations since the operator is, by definition, invariant against any monotonic transformation of the gray-scale [31]. We compute LBP features using eight sampling points on a circle of radius $R = 2$ pixels (see [31] for additional details). This choice results in a 10-dimensional feature vector, corresponding to the occurrence histogram of unique uniform rotation invariant binary patterns that can occur in the circularly symmetric neighbor set.

The retained feature scores for classification are the 25th ([$f_{24}, \ldots, f_{33}$]), 50th ([$f_{34}, \ldots, f_{43}$]) and 75th ([$f_{44}, \ldots, f_{53}$]) percentiles of each of the 10 texture images (an example is shown in Fig. 9). Note that some of the maps appear to be spatially correlated, while others suggest good potential for discrimination. Despite the very similar colors of the two benign cases, some texture maps are very different. The presence of anatomical structures such as networks in the top and bottom cases are plausible reasons for the differences in the textures. Another promising alternative that could be considered would be focusing on the relative frequency of occurrence of clusters of LBP patterns that could be obtained by automatically clustering all lesions “boxes” into a certain number of classes, rather than taking LBP percentiles from each map. This was investigated earlier in Zortea et al. [32].

2.4. Feature selection

Given the unfavorable ratio between the reduced number of training samples available, especially for the malignant class, and the dimensionality of the input feature vector, feature reduction was considered before training a statistical classifier. In particular, we focused on feature selection. The algorithm used for feature selection was the sequential forward selection (SFS) method [33]. SFS is one of the simplest sub-optimal step-wise search strategies. It iteratively identifies the best feature subset obtained by adding to the current feature subset one feature at a time. The search depth (maximum number of features selected) was empirically set to 10 features in our application.

The optimization score for feature selection is the average of sensitivity and specificity, as it balances both detected and missing lesions. The optimization score is computed on the training set using 5-fold cross validation (CV). CV is used to reduce the risk of overfitting our simple models during the feature selection stage. After training, we choose the subset of features corresponding to the peak of CV accuracy as the best subset for statistical classification.

2.5. Statistical classification

We focus on well-known statistical methods for classification. First, two classical parametric approaches, the LDA and the quadratic discriminant analysis (QDA) [30] are considered. The classification and regression trees algorithm (CART) [34] was also included in the analysis as an example of a non-parametric decision tree learning technique.

Recommending excision of all malignant lesions is of paramount importance, whereas excision of benign ones is tolerable, but should be minimized as much as possible to reduce stress to patients, costs to the health system, etc. It is natural to think that the “cost” to a patient having a benign lesion classified as being malignant (excised) is very different from the opposite case, where a malignant lesion is incorrectly classified as benign (not excised). A high sensitivity is more important than a high specificity in this case. Indeed, it is very important to recognize all melanoma lesions, but it should be noted that one of the main reasons not to recommend existing CAD systems for melanoma screening is their high rate of false positive assessments [35]. Another aspect to be considered is the prior probability of patients with benign and malignant lesions visiting a physician. It is possible to incorporate both the cost and the prior aspects into statistical classifiers. We address these aspects in a classic manner. Let the populations π_1 and π_2 be described by multivariate normal densities with mean vectors and covariance matrices μ_1, Σ_1 and μ_2, Σ_2, respectively. The allocation rule that minimizes the expected cost of misclassification of a skin lesion x_0 to π_1 is given by Johnson and Wichern [30]. In our case, allocate x_0 to π_1 if

$$-\frac{1}{2}x_0'(\Sigma_1^{-1} - \Sigma_2^{-1})x_0 + (\mu_1'\Sigma_1^{-1} - \mu_2'\Sigma_2^{-1})x_0 - k \geq \ln\left[\frac{c(1|2)}{c(2|1)} \cdot \frac{p_2}{p_1}\right], \tag{6}$$

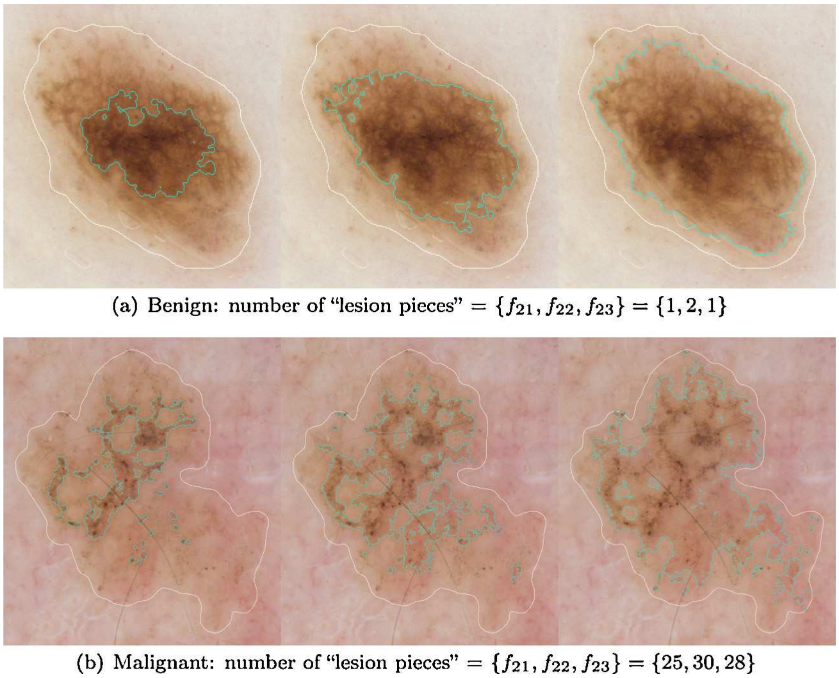
where k is

$$k = \frac{1}{2}\ln\frac{|\Sigma_1|}{|\Sigma_2|} + \frac{1}{2}(\mu_1'\Sigma_1^{-1}\mu_1 - \mu_2'\Sigma_2^{-1}\mu_2),$$

p_i are the prior probabilities for π_i, $c(i|j)$ is the cost of classifying a sample as π_i when, in fact, it is from π_j, for $i, j = 1, 2$, and $'$ is the transpose operator. Allocate x_0 to π_2 otherwise.

Define the right side of Eq. (6) as $\ln(\alpha)$. In our experiments we have empirically chosen α from the set {1.0, 1.5, 2.0, 2.5, . . ., 5.0}. When $\alpha > 1$, we assign more weight to the malignant class. We select the value of α that provides the best average of sensitivity and specificity, measured as the 5-fold cross-validation accuracy in the training. An example of the effect of α in the performance of the LDA classifier is shown in Table 2. In this particular realization, we choose a compromise solution with $\alpha = 2.0$, and later apply the classifier to an unseen test set. Other realizations may require choosing a different value of α.

Please cite this article in press as: Zortea M, et al. Performance of a dermoscopy-based computer vision system for the diagnosis of pigmented skin lesions compared with visual evaluation by experienced dermatologists. Artif Intell Med (2013), http://dx.doi.org/10.1016/j.artmed.2013.11.006


(a) Benign: number of "lesion pieces" = $\{f_{21}, f_{22}, f_{23}\} = \{1, 2, 1\}$

(b) Malignant: number of "lesion pieces" = $\{f_{21}, f_{22}, f_{23}\} = \{25, 30, 28\}$

Fig. 8. Example of geometric features corresponding to the number of "lesion pieces" obtained by applying binary thresholds to the lesion area (delineated by the white contour) at different gray-scale percentiles.

3. Experiments

Two binary classification experiments are considered. In the first experiment we try to separate clearly benign lesions from malignant lesions. In the second experiment we include very challenging benign lesions.

3.1. Experiment 1

In this experiment, we consider the classification task of separating clearly benign "not cut" and malignant "cut" lesions. Out of the 37 malignant lesions available, 27 were randomly selected for training, and the remaining 10 were used for testing. In order to keep the statistical classifiers balanced, the same number of 37 benign lesions (27 plus 10) was used for training and testing, respectively. Here all the benign lesions were randomly sampled from the clearly benign subclass containing 89 lesions. Table 3 shows the classification scores, where each score is computed as the average result based on 20 realizations.

Fig. 10 shows the dispersion of results. QDA achieved the highest sensitivity in the experiment (86.0%). The simpler LDA model achieved a good compromise between sensitivity (80.5%) and specificity (83.5%), with competitive values when compared to the best performing dermatologist (85.0% and 57.5%, respectively). The performance of CART was very similar to LDA. The number of features selected for classification using the step-wise strategy ranged from one to ten for LDA (average of 5.0), depending on the randomly selected training sets, and five to ten for QDA (average of 7.6).

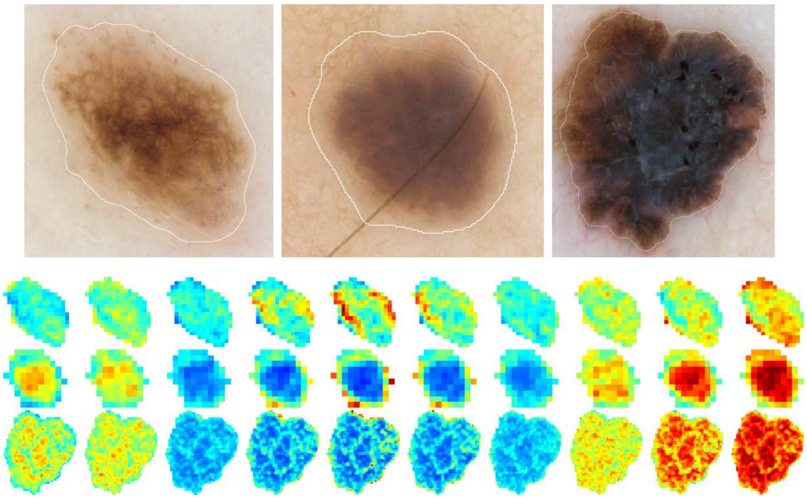
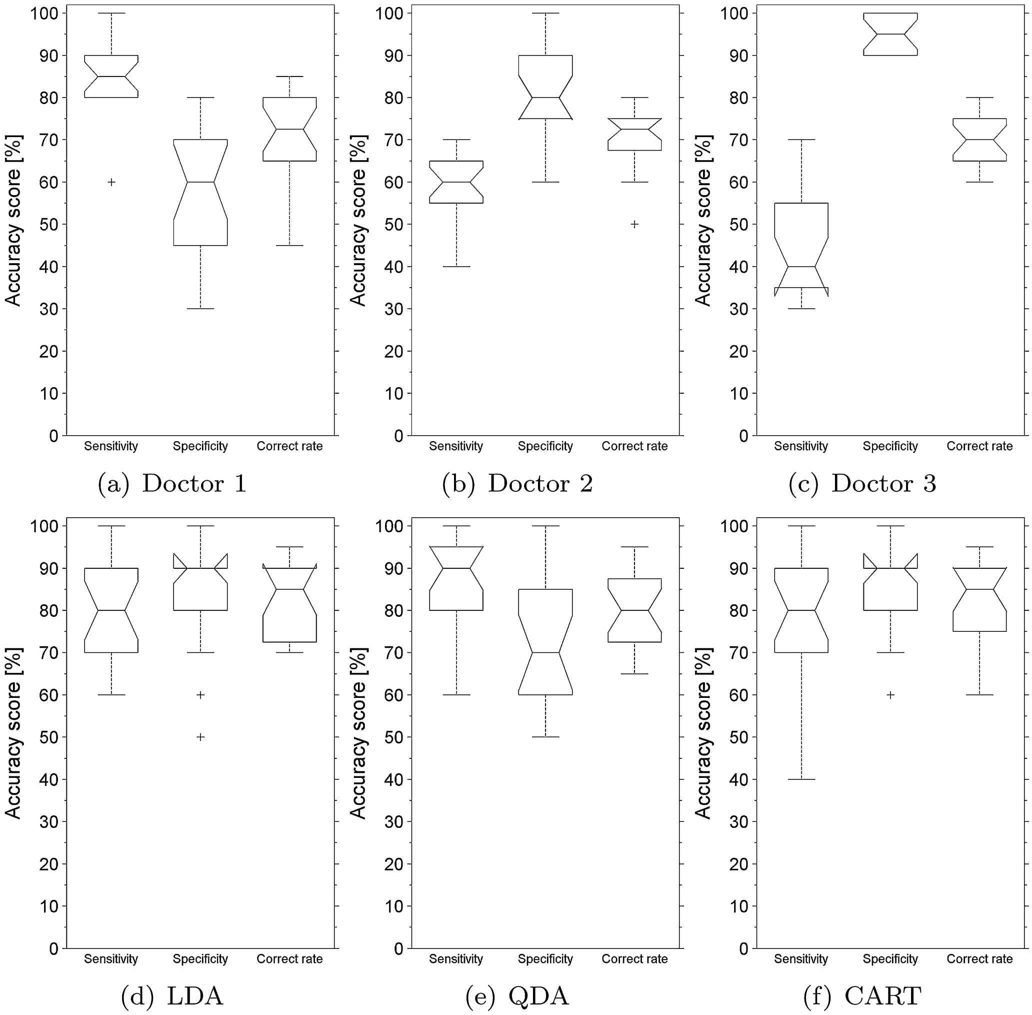


Fig. 9. Top row, example images from left to right: Benign "not cut", benign "cut", and malignant. The following rows of artificially colored images are texture images derived using the LBP algorithm and correspond to the three top images (row 1/2/3 with top left/middle/right image). Blue and red corresponds to lower and higher values of texture, respectively. Maps 1–9 from left to right were linearly scaled in the range {0–0.18}, whereas 10 is in {0–0.36}. This is kept fixed for the three sets shown above, so the values are therefore directly comparable by visual inspection. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

Please cite this article in press as: Zortea M, et al. Performance of a dermoscopy-based computer vision system for the diagnosis of pigmented skin lesions compared with visual evaluation by experienced dermatologists. Artif Intell Med (2013), http://dx.doi.org/10.1016/j.artmed.2013.11.006



(a) Doctor 1 (b) Doctor 2 (c) Doctor 3

(d) LDA (e) QDA (f) CART

Fig. 10. Experiment 1: accuracy scores across 20 random test sets both for doctors (a–c) and the computer methods (d–f). On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points that are considered not to be outliers, and the outliers are plotted individually. Notches indicates comparison intervals. Two medians are significantly different at the 5% level if their intervals do not overlap [36]. The interval endpoints are the extremes of the notches. When the sample size is small, some notches may extend beyond the end of the box, as shown in (c, d, f).

Visual analysis of the box notches in Fig. 10(a) and (d) suggests that (median) sensitivity differences between computer (LDA) and the best doctor (#1) are not significant, but the computer has a significantly higher specificity. The substantial difference in specificity also significantly affects the correct prediction rate (82.0% versus 71.3%).

Notice that the differences in the correct rate among the three doctors are not significant. However, although their correct rates are similar, the sensitivity and specificity outcomes are surprisingly divergent. Similarly, the differences in the correct rate among the computational approaches are not significant, but the results outperform the doctors' correct rate.

Table 2
Example of 5-fold cross-validation accuracies in the training set for LDA as a function of cost α in Experiment 1. Sensitivity (SE), specificity (SP) and correct rate (CR) scores are given in percent. The corresponding number of features at the peak of accuracy is included. In bold is the retained solution.

α	SE	SP	$CR = (SE + SP)/2$	# features
1.0	88.88	92.59	90.74	1
1.5	92.59	92.59	92.59	2
2.0	96.29	92.59	**94.44**	7
2.5	92.59	85.18	88.88	2
3.0	92.59	88.88	90.74	3
3.5	92.59	88.88	90.74	7
4.0	92.59	81.48	87.03	1
4.5	92.59	85.18	88.88	7
5.0	96.29	81.48	88.88	8

3.2. Experiment 2

In this experiment, we add atypical benign lesions that resemble melanomas (dysplastic nevi) to the test set, those lesions in the set labeled benign "cut". Given the increased classification difficulty, we preferred to keep the same binary classification strategy adopted in the first experiment. Therefore, the binary classifiers are

Table 3
Average accuracy scores for a test set including clearly benign ("not cut") and malignant ("cut") lesions, in 20 realizations. Sensitivity, specificity, and correct rate scores are in %. The average number of excisions is also included (ideally it should be 10 in this case).

	SE	SP	CR	# excisions
Doctor 1	85.0	57.5	71.3	12.8
Doctor 2	58.5	83.0	70.8	7.6
Doctor 3	44.5	95.0	69.8	5.0
LDA	80.5	83.5	82.0	9.7
QDA	86.0	73.0	79.5	11.3
CART	79.0	85.0	82.0	9.4

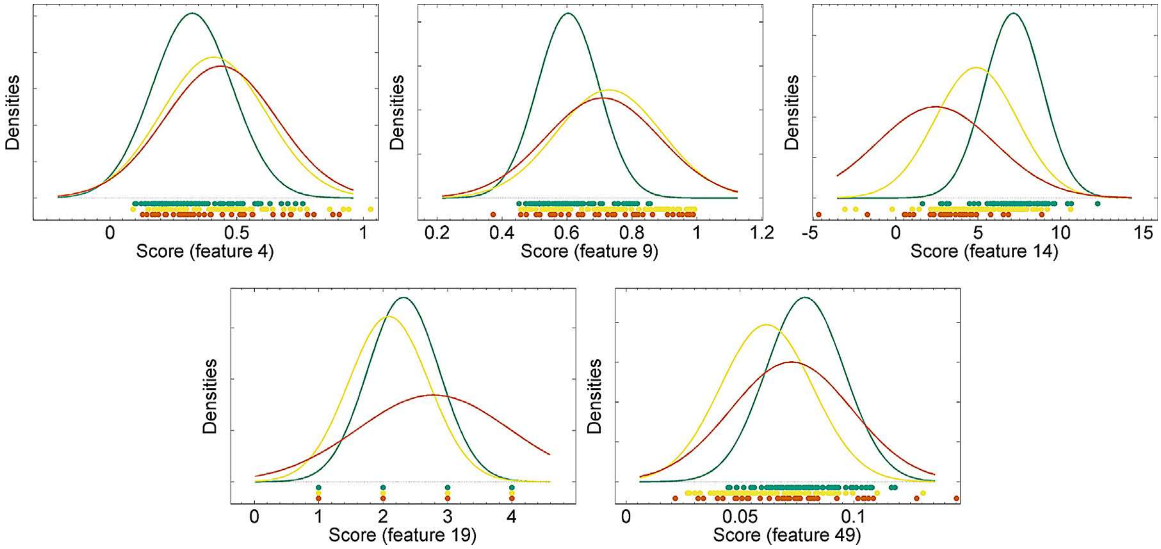


Fig. 11. Examples of different feature scores. The colored dots mark feature scores connected to cases of benign "not cut" (green), benign "cut" (yellow), and malignant (red). The solid lines are fitted Gaussian density estimates of the respective class of cases. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

exactly the same as in the previous experiment, trained to distinguish between clearly benign "not cut" and malignant "cut" lesions. Notice that in this classification setting, the addition of benign lesions from the subclass containing the more difficult benign cases will affect the specificity score only. The sensitivity in this experiment remains identical as in Experiment 1 because no additional malignant samples are available, nor are the classifiers changed.

The reason for training the classifier using only clearly benign "not cut" and malignant "cut" is illustrated in Fig. 11. In short, given our proposed set of explanatory features, the statistical distributions of the subclass benign "cut" often seems to be closer to the malignant class than to the benign "not cut." This is the main reason for our experimental choice, where the important subclass of difficult benign was not considered during the training phase, but only in the testing of the classifier.

With the above reasoning in mind, an important question is how many of those difficult benign cases should be included in the test set? In our experiment, we have empirically chosen a benign test set consisting of 50% of clearly benign "not cut" and 50% of more challenging benign "cut" cases. To summarize, the definition of the classes and the partition of the training and test sets in this experiment is: The training set contains a total of 27 malignant ("cuts") and 27 benign ("not cut") cases, the same training size used in Experiment 1, while the test set contains 10 malignant ("cut") and 10 benign ("not cut") cases, plus 10 benign ("cut") cases.

The classification scores, computed as the average over 20 realizations, are shown in Table 4. The respective box plots showing the dispersion of results are shown in Fig. 12. Again, QDA was preferable to LDA and CART, with higher sensitivity (86.0%), but LDA and CART have higher specificities (62.3% and 63.0%, respectively). These are competitive figures when compared to the best performing dermatologist (85.0% and 47.8%, respectively, as shown in Table 4). As discussed in the first experiment, sensitivity and specificity differences among the three doctors are very large, and in this second experiment the correct rate score is influenced to a greater extent by the specificity (due to the higher number of benign lesions in the test set). Also in this experiment, results suggest that the higher (median) specificity of the computer methods (LDA and CART) is statistically significant when compared to the best doctor.

Table 4
Average accuracy scores for a test set including clearly benign ("not cut"), suspicious benign ("cut"), and malignant ("cut") lesions, for 20 realizations. Sensitivity, specificity, and correct rate scores are in %. The average number of excisions is also included (ideally it should be 10 in this case).

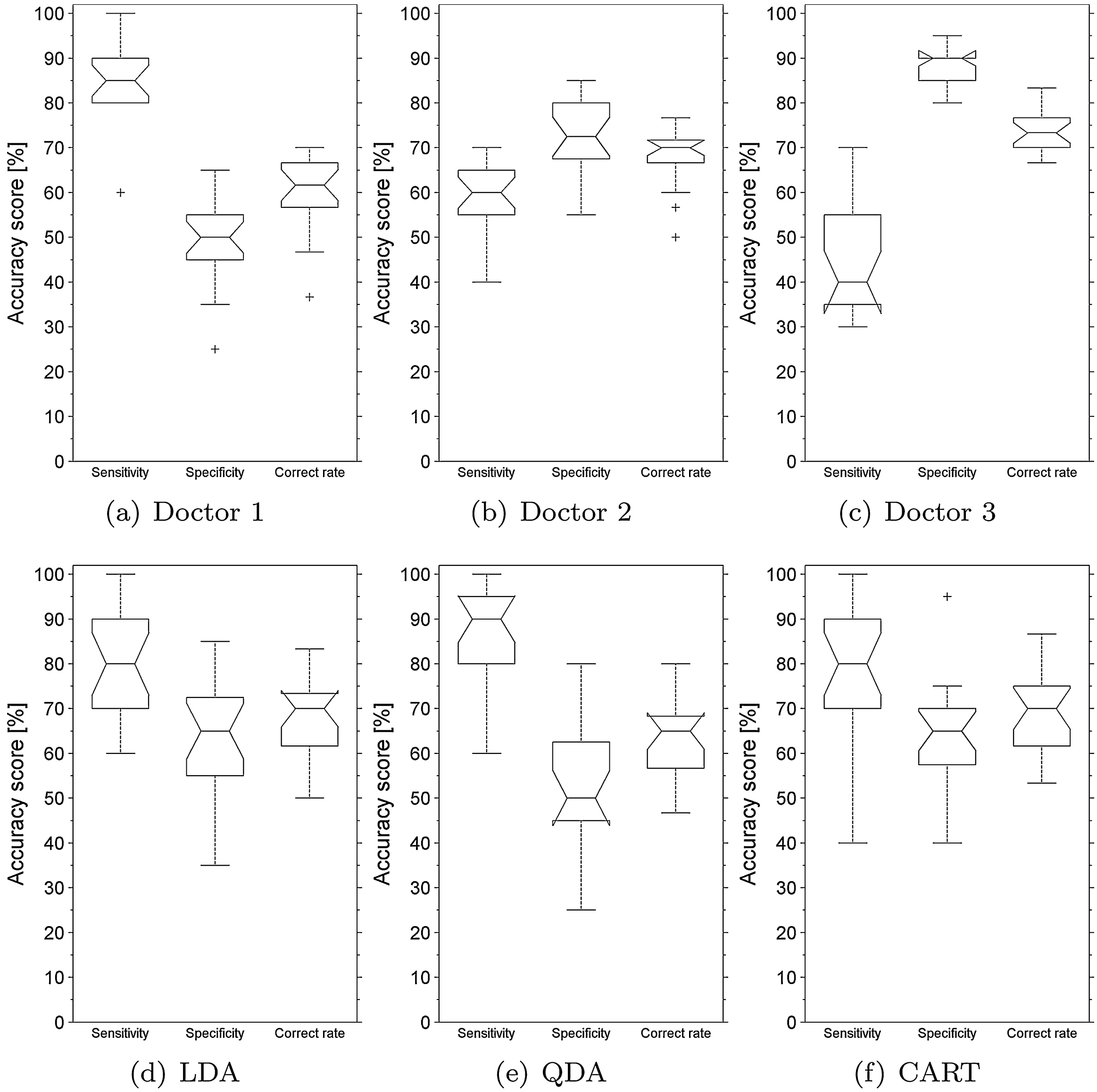
	SE	SP	CR	# excisions
Doctor 1	85.0	47.8	60.2	19.0
Doctor 2	58.5	72.5	67.8	11.4
Doctor 3	44.5	87.8	73.3	6.9
LDA	80.5	62.3	68.3	15.6
QDA	86.0	52.0	63.3	18.2
CART	79.0	63.0	68.3	15.3

Table 3 and 4 also show the number of excisions for the test set containing 10 malignant cases in each random realization. Ideally the 10 malignant lesions should be excised. Note that the addition of the difficult benign cases in Experiment 2 significantly increased the average number of excisions recommended by Doctor 1 (from 12.8 to 19.0). The increment was slightly lower for the computer methods (from 11.3 to 18.2).

We conclude the analysis by showing examples of scatter plots of features selected during the experiments in Fig. 13. Different subsets of training lesions often lead to the selection of different combinations of features, and the high variability of the classification scores shown in Figs. 10 and 12 are mainly due to the small training sets used in the experiments. Fig. 14 shows how many times each feature was selected in the 20 realizations by LDA and QDA. It provides an indication of the importance of each feature in the final classification.

4. Discussion

The goal of this study was to investigate feature extraction and classification in PSL. The data set contained all common types of PSL including the most prevalent subclasses of melanoma. The findings were evaluated on the basis of histopathology reports as the gold standard. In addition, we compared the results of the classifiers with the performance of three physicians practicing dermatology. The dermatologists made their assessments on the basis of dermoscopic images only, i.e., no clinical data were provided. This was done in order to evaluate how the doctors would assess the dermoscopic features. We have not investigated a complete CAD



(a) Doctor 1 (b) Doctor 2 (c) Doctor 3

(d) LDA (e) QDA (f) CART

Fig. 12. Experiment 2: accuracy scores across 20 random test sets for doctors (a–c) and the computer methods (d–f). See caption of Fig. 10 for further details.

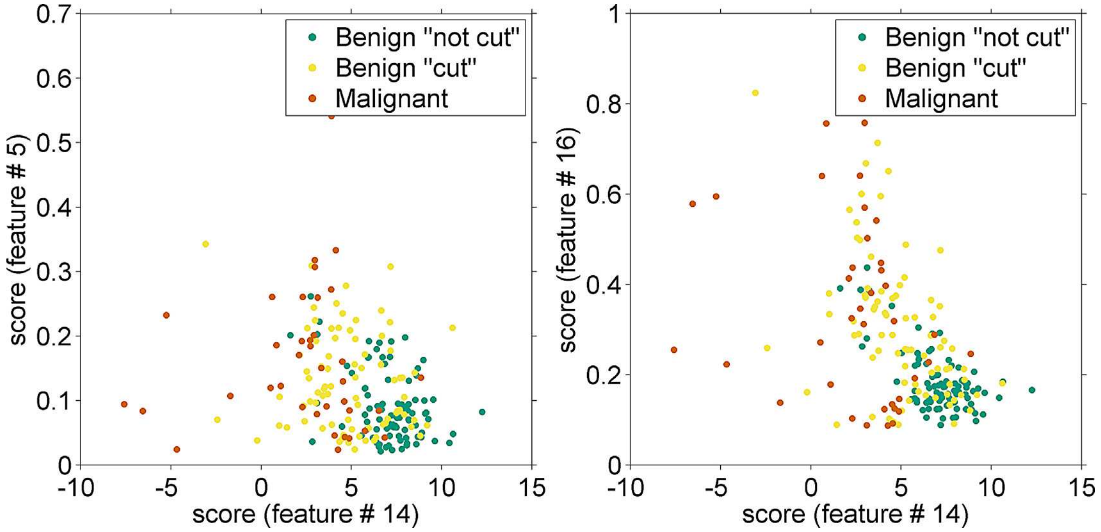


Fig. 13. Examples of the first two features selected by LDA (left) and CART (right). All 206 samples are shown for visualization purposes (but not for feature selection).

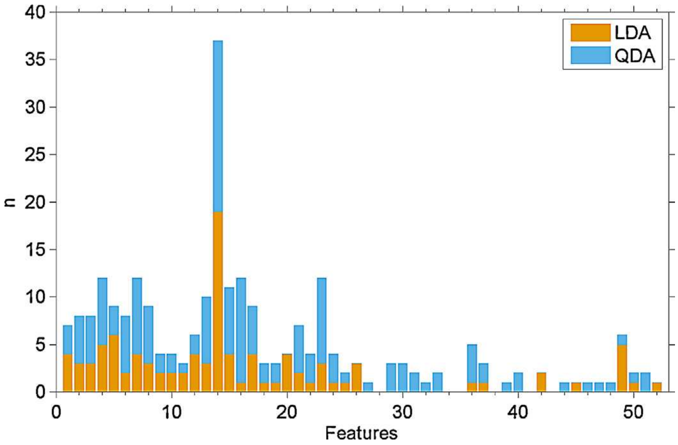


Fig. 14. Number of times (n) each feature was included in the set of selected features in the 20 random data sets. The numbers for the two classifiers are stacked.

system in a real clinical setting, and therefore the interpretation of clinical data was not considered to be within the scope of this study. In order to develop useful features and classifiers, we believe that the classifiers should be compared to doctors' assessment of dermoscopic features in both simulated and real clinical environments.

In a real clinical setting, the performance of the dermatologists presented in Table 3 would likely improve. The suggestions made by the dermatologists in the experiments should not be interpreted as the final decision that would be made in clinical practice where additional information would be taken into account. We have performed a controlled, paired computer versus doctors experiment. Doctors, like computers, should benefit from additional information about the patient, such as change history of the lesion [37] and family history, or use of the "ugly ducklings" approach [38].

There are several open questions related to classification experiments for diagnosis of skin lesions. An important one is whether the data set used to train and test the performance of the automated diagnosis method is sufficiently representative. The data set used in this study contains a large number of early stage melanomas as indicated by the histopathology reports, thus indicating that the data set is challenging. This is reflected in the feedback from the dermatologists, where even dermatologists with experience in dermoscopy had difficulties in determining the correct diagnosis. We found that even with this challenging dataset, the features we developed were relevant to the classification problem.

The features described are neither exhaustive nor optimized, and the possibilities for adding or improving upon features are limitless. In particular, palette-based color features are not robust to changes in white balance or exposure, but designing robust color features is itself a challenging problem.

For all lesions used in this study hairs are not occluding the images to a significant degree. It is likely that the presence of hairs will decrease the performance of the classifier, and in a final system this should be addressed. We have developed several algorithms for the detection of hair pixels, see Thon et al. [39] and Møllersen et al. [40], and others have studied imputation of hair pixels, e.g., Abbas et al. [41]. The features used in this study are designed so that imputation is not necessary, and merely the identification and removal of hair pixels is required, although the segmentation step is inevitably sensitive to hair occlusion.

4.1. Future work

The system described here could form the basis for a CAD tool for use in primary care. We intend to design this tool, and perform pilot testing both in specialist and primary care settings to fine-tune both feature extraction and classification, and the user interface before commencing clinical trials. Thus, we should obtain an understanding of how the system performs in realistic settings, and pertinent information should be presented so that the system is effective in clinical decision support.

Future trials could contribute to a better understanding of the classification performance of the features when alternative segmentation algorithms are used. Since no segmentation algorithm will perform well for all types of skin lesions, especially the challenging ones, a combination of features derived from different segmentation algorithms (see e.g., [42]) may lead to a more robust final classification.

5. Conclusions

The proposed system performs as well as or better than dermatologists on both sensitivity and specificity when the only information available is the dermoscopic image. In a clinical setting, additional information will be available both for the medical doctor and the CAD system, and will increase the sensitivity and specificity scores of both. To what extent is yet to be investigated.

Our main contributions include the design of novel image features for dermoscopic images that prove useful for computer-based classification of skin lesions. Also, we give an indication on how challenging the dataset used to validate such features is by having dermatologists evaluate the same set of images evaluated by a computer system.

The specificity scores of the dermatologists are highly dependent on the type of PSL included in the dataset, as is also the case for the proposed computer system. This, along with the lack of publicly available image databases, makes it difficult to compare the performances reported in other similar studies.

Our image acquisition equipment is off-the-shelf and the feature reduction and classification methods are well-known. The proposed features, however, are new, and have been demonstrated as useful in discriminating melanomas from non-melanomas. Due to the limited amount of melanomas in our database, the classifiers may have suffered from inadequate training. A larger database will likely improve the classifiers, but not beyond the discriminatory power inherent in the chosen individual features.

Conflict of interest statement

The authors have no conflict of interest to disclose.

Acknowledgments

This research was performed as a part of Tromsø Telemedicine Laboratory, a Centre for Research-based Innovation funded by the Research Council of Norway, grant no. 174934. We gratefully acknowledge the dermatologists Ingeborg Bachmann (Haukeland University Hospital), Lars Kåre Fagerheim (Stavanger University Hospital), and Katja Bremnes (University Hospital of North Norway), who voluntarily participated in the experiments. We are very grateful to the anonymous reviewers for their valuable suggestions.

References

[1] Garbe C, Leiter U. Melanoma epidemiology and trends. Clinics in Dermatology 2009;27(1):3–9.
[2] Holterhues C, de Vries E, Louwman MW, Koljenović S, Nijsten T. Incidence and trends of cutaneous malignancies in the Netherlands, 1989–2005. Journal of Investigative Dermatology 2010;130(7):1807–12.
[3] Kittler H, Pehamberger H, Wolff K, Binder M. Diagnostic accuracy of dermoscopy. Lancet Oncology 2002;3(3):159–65.

Please cite this article in press as: Zortea M, et al. Performance of a dermoscopy-based computer vision system for the diagnosis of pigmented skin lesions compared with visual evaluation by experienced dermatologists. Artif Intell Med (2013), http://dx.doi.org/10.1016/j.artmed.2013.11.006

[4] Vestergaard ME, Macaskill P, Holt PE, Menzies SW. Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting. British Journal of Dermatology 2008;159(3):669–76.

[5] Rosendahl C, Tschandl P, Cameron A, Kittler H. Diagnostic accuracy of dermatoscopy for melanocytic and nonmelanocytic pigmented lesions. Journal of the American Academy of Dermatology 2011;64(6):1068–73.

[6] Pehamberger H, Steiner A, Wolff K. In vivo epiluminescence microscopy of pigmented skin lesions. I. Pattern analysis of pigmented skin lesions. Journal of the American Academy of Dermatology 1987;17(4):571–83.

[7] Stolz W, Riemann A, Cognetta A, Pillet L, Abmayr W, Holzel D, et al. ABCD rule of dermatoscopy: a new practical method for early recognition of malignant-melanoma. European Journal of Dermatology 1994;4(7):521–7.

[8] Menzies S, Ingvar C, McCarthy W. A sensitivity and specificity analysis of the surface microscopy features of invasive melanoma. Melanoma Research 1996;6(1):55–62.

[9] Soyer H, Argenziano G, Zalaudek I, Corona R, Sera F, Talamini R, et al. Three-point checklist of dermoscopy: a new screening method for early detection of melanoma. Dermatology 2004;208(1):27–31.

[10] 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.

[11] Henning J, Dusza S, Wang S, Marghoob A, Rabinovitz H, Polsky D, et al. The CASH (color, architecture, symmetry, and homogeneity) algorithm for dermoscopy. Journal of the American Academy of Dermatology 2007;56(1):45–52.

[12] Rosendahl C, Cameron A, McColl I, Wilkinson D. Dermatoscopy in routine practice 'Chaos and Clues'. Australian Family Physician 2012;41(7):482–7.

[13] Bourne P, Rosendahl C, Keir J, Cameron A. BLINCK – a diagnostic algorithm for skin cancer diagnosis combining clinical features with dermatoscopy findings. Dermatology Practical and Conceptual 2012;2(2):55–61.

[14] Marghoob AA, Malvehy J, Braun R, editors. An atlas of dermoscopy. London, UK: Taylor & Francis; 2012.

[15] Skvara H, Teban L, Fiebiger M, Binder MHK. Limitations of dermoscopy in the recognition of melanoma. Archives of Dermatology 2005;141(2):155–60.

[16] Korotkov K, Garcia R. Computerized analysis of pigmented skin lesions: a review. Artificial Intelligence in Medicine 2012;56(2):69–90.

[17] 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.

[18] Dolianitis C, Kelly J, Wolfe R, Simpson P. Comparative performance of 4 dermoscopic algorithms by nonexperts for the diagnosis of melanocytic lesions. Archives of Dermatology 2005;141(8):1008–14.

[19] Day G, Barbour R. Automated skin lesion screening – a new approach. Melanoma Research 2001;11(1):31–5.

[20] Tenenhaus A, Nkengne A, Horn J, Serruys C, Giron A, Fertil B. Detection of melanoma from dermoscopic images of naevi acquired under uncontrolled conditions. Skin Research and Technology 2010;16(1):85–97.

[21] 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.

[22] Gewirtzman A, Braun R. Computerized digital dermoscopy. Journal of Cosmetic Dermatology 2003;2(1):14–20.

[23] Baldi A, Murace R, Dragonetti E, Manganaro M, Guerra O, Bizzi S, et al. Definition of an automated content-based image retrieval (CBIR) system for the comparison of dermoscopic images of pigmented skin lesions. BioMedical Engineering OnLine 2009;8:1–10.

[24] Zortea M, Skrøvseth SO, Schopf TR, Kirchesch HM, Godtliebsen F. Automatic segmentation of dermoscopic images by iterative classification. Journal of Biomedical Imaging 2011;2011:1–19.

[25] Seidenari S, Pellacani G, Grana C. Computer description of colours in dermoscopic melanocytic lesion images reproducing clinical assessment. British Journal of Dermatology 2003;149(3):523–9.

[26] Pellacani G, Grana C, Seidenari S. Automated description of colours in polarized-light surface microscopy images of melanocytic lesions. Melanoma Research 2004;14(2):125–30.

[27] Wand M, Jones M. Kernel smoothing. Boca Raton, FL: Chapman & Hall/CRC; 1995.

[28] Carter E, Ohno Y, Pointer M, Robertson A, Seve R, Schanda J, et al. Colorimetry. Publication 15. Tech. Rep. Vienna, AU: CIE Central Bureau; 2004.

[29] Bowman A, Azzalini A. Applied smoothing techniques for data analysis: the kernel approach with S-plus illustrations. New York, NY: Oxford University Press; 1997.

[30] Johnson R, Wichern D. Applied multivariate statistical analysis, vol. 4. Englewood Cliffs, NJ: Prentice Hall; 1992.

[31] Ojala T, Pietikä inen M, Mäenpää T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence 2002;24(7): 971–87.

[32] Zortea M, Skrøvseth SO, Godtliebsen F. Automatic learning of spatial patterns for diagnosis of skin lesions. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), vol. 2010. New York, NY: IEEE; 2010. p. 5601–4.

[33] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York, NY: Springer; 2009.

[34] Breiman L, Friedman J, Stone C, Olshen R. Classification and regression trees. Boca Raton, FL: Chapman & Hall/CRC; 1984.

[35] Frühauf J, Leinweber B, Fink-Puches R, Ahlgrimm-Siess V, Richtig E, Wolf I, et al. Patient acceptance and diagnostic utility of automated digital image analysis of pigmented skin lesions. Journal of the European Academy of Dermatology and Venereology 2012;26(3):368–72.

[36] Schenker N, Gentleman JF. On judging the significance of differences by examining the overlap between confidence intervals. The American Statistician 2001;55(3):182–6.

[37] Geilhufe M, Skrøvseth SO, Godtliebsen F. Digital monitoring of changes in skin lesions. In: Macedo M, editor. IADIS International Conference e-Health. Freiburg, Germany: International Association for Development of the Information Society; 2010. p. 229–33.

[38] Grob J, Bonerandi J. The 'ugly duckling' sign: identification of the common characteristics of nevi in an individual as a basis for melanoma screening. Archives of Dermatology 1998;134(1):103–4.

[39] Thon K, Rue H, Skrøvseth SO, Godtliebsen F. Bayesian multiscale analysis of images modeled as Gaussian Markov random fields. Computational Statistics and Data Analysis 2012;56(1):49–61.

[40] Møllersen K, Kirchesch HM, Schopf TG, Godtliebsen F. Unsupervised segmentation for digital dermoscopic images. Skin Research and Technology 2010;16(4):401–7.

[41] Abbas Q, Garcia IF, Emre Celebi M, Ahmad W. A feature-preserving hair removal algorithm for dermoscopy images. Skin Research and Technology 2013;19(1):27–36.

[42] Skrøvseth SO, Schopf TR, Thon K, Zortea M, Geilhufe M, Møllersen K, et al. A computer aided diagnostic system for malignant melanomas. In: 3rd International Symposium on Applied Sciences in Biomedical and Communication Technologies (ISABEL), 2010. Rome, Italy: IEEE; 2010.

Please cite this article in press as: Zortea M, et al. Performance of a dermoscopy-based computer vision system for the diagnosis of pigmented skin lesions compared with visual evaluation by experienced dermatologists. Artif Intell Med (2013), http://dx.doi.org/10.1016/j.artmed.2013.11.006