A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions

Lucia Ballerini, Robert B. Fisher, Ben Aldridge, Jonathan Rees

Abstract This chapter proposes a novel hierarchical classification system based on the K-Nearest Neighbors (K-NN) model and its application to non-melanoma skin lesion classification. Color and texture features are extracted from skin lesion images. The hierarchical structure decomposes the classification task into a set of simpler problems, one at each node of the classification. Feature selection is embedded in the hierarchical framework that chooses the most relevant feature subsets at each node of the hierarchy. The accuracy of the proposed hierarchical scheme is higher than 93% in discriminating cancer and potential at risk lesions from benign lesions, and it reaches an overall classification accuracy of 74% over five common classes of skin lesions, including two non-melanoma cancer types. This is the most extensive known result on non-melanoma skin cancer classification using color and texture information from images acquired by a standard camera (non-dermoscopy).

1 Introduction

Skin cancers are the most common forms of human malignancies in fair skinned populations [18]. Although malignant melanoma is the form of "skin cancer" with the highest mortality, the "non-melanoma skin cancers" (basal cell carcinomas and squamous cell carcinomas, etc.) are far more common. The incidence of both melanoma and non-melanoma skin cancers is increasing, with the number of cases being diagnosed doubling approximately every

L. Ballerini and R. B. Fisher
School of Informatics, University of Edinburgh, Edinburgh, UK,
e-mail: lucia.ballerini@ed.ac.uk, rbf@inf.ed.ac.uk

B. Aldridge and J. Rees
Department of Dermatology, University of Edinburgh, Edinburgh, UK,
e-mail: ben.aldridge@ed.ac.uk, jonathan.rees@ed.ac.uk

15 years [35]. It is widely accepted that early detection is fundamental to reducing the diseases' morbidity and mortality. Automatic detection systems may offer benefit for this key diagnostic task.

There are a considerable number of published studies on classification methods relating to the diagnosis of cutaneous malignancies. The first published work presenting an automatic classification of melanoma could be found in 1987 [11]. A paper describing the first complete system appeared a few years later [29]. The number of published papers has increased every year and the significant progress that has occurred in this field is demonstrated by the recent journal special issue that summarizes the state of the art in computerized analysis of skin cancer images and provides future directions for this exciting subfield of medical image analysis [16].

Different techniques for enhancement, segmentation, feature extraction and classification have been reported by several authors. Enhancement includes color calibration and normalization [32, 54].

Concerning segmentation, Celebi et al. [14] presented a systematic overview of main border detection methods: clustering followed by active contours are the most popular. Improvements in lesion border detection are described in recent papers [39, 54, 26, 61, 66].

Numerous features have been extracted from skin images, including shape, color, texture and border properties [56, 37, 63, 43, 52, 57, 19]. It is common to use features related to the ABCD mnemonic rule [49]. However, our experiments suggested that the use of the ABCD rule in the development of automatic classifiers can be arguably discouraged [64].

Classification methods range from discriminant analysis to neural networks and support vector machines [55, 41, 15]. See Maglogiannis et al. [40] for a review of the state of the art of computer vision system for skin lesion characterization.

These methods have been mainly developed for images acquired by epiluminescence microscopy (ELM or dermoscopy). However, newer technologies, including digital dermoscopy, infrared imaging, multispectral imaging, and confocal microscopy, have recently come to the forefront in providing greater diagnostic accuracy [16].

Moreover published studies mainly focus on differentiating melanocytic naevi (moles) from melanoma. Whilst this is undeniably important (as malignant melanoma is the form of skin cancer with the highest mortality), in the "real-world" the majority of lesions presenting to dermatologists for assessment are not covered by this narrow domain, and such systems ignore other benign lesions and crucially the two most common skin cancers (Squamous Cell Carcinomas and Basal Cell Carcinomas) [27, 10, 60].

The proposed work uses only high resolution color images acquired using standard cameras. To our knowledge only two melanoma pre-screening systems are based on standard camera images [1, 12].

In the current study, color and texture features are used for the classification. We focus on 5 common classes of skin lesions: Actinic Keratosis (AK),

Basal Cell Carcinoma (BCC), Melanocytic Nevus / Mole (ML), Squamous Cell Carcinoma (SCC), Seborrhoeic Keratosis (SK). As far as we can tell there is no research on automatic classification of these lesion types (other than moles) outside our group [39].

Moreover, this paper introduces a new hierarchical framework for skin lesion classification. This framework is comprised of a modified version of the K-Nearest Neighbors (K-NN) classifier, the Hierarchical K-Nearest Neighbors (HKNN) classifier, and a new similarity measure, based on color and texture features, that uses different feature sets for comparing similarity at each node of the hierarchy.

The motivation for using a K-NN classifier can be seen in Fig. 4. It is clear that the clusters overlap greatly, but are distinguishable. No hard boundary could separate them (e.g. as usable by a support vector machine or Bayesian classifier).

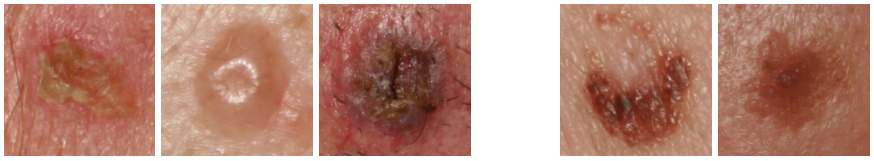
Below we describe how the lesion classes can be organized in a hierarchical scheme (Sect. 2) that suggests the use of the hierarchical classifier (Sect. 3). Then we introduce the feature pool (Sect. 4). Therefore we make 2 claims:

1. The use of a hierarchical K-NN classifier improves classification accuracy from 70% to 74% over a non-hierarchical K-NN, and from 67% and 69% over a flat and a hierarchical Bayes classifier, respectively,
2. This is the most extensive paper to present lesion classification results for non-melanoma skin cancer using color imagery acquired by a standard camera, unlike the dermoscopy method, which requires a specialised sensor.

While 74% is low compared to the 90+% rates achieved by melanoma classification, we argue that 74% is worth publication: a) the melanoma results are from only the 2 class problem of melanoma vs melanocytic naevi (moles), and b) it has taken more than 20 years of research specifically on that problem to reach the 90+% levels, whereas this is the first research on image-based classification of AK, BCC, SCC and SK. We accept that whilst classification rates of this magnitude seem low in the sphere of informatics research, these rates are significantly above what is currently being achieved in non-specialist medical practice [21, 51, 44, 27, 10, 60].

2 Skin class hierarchy

Some images of the five classes are shown in Fig. 1. The hierarchy is fixed *a priori* by grouping our image classes into two main groups. The first group, hence called *Group1*, contains lesion classes: Actinic Keratosis (AK), Basal Cell Carcinoma (BCC) and Squamous Cell Carcinoma (SCC). The second group, hence called *Group2*, contains lesion classes: Melanocytic Nevus/ Mole (ML) and Seborrhoeic Keratosis (SK). We note that AK, BCC, SCC, ML



(a) AK (b) BCC (c) SCC (d) ML (e) SK

Fig. 1 Examples of skin lesion images from the different classes used in this work

and SK are diagnostic classes defined by dermatologists. The two groups were constructed by clustering classes containing images which were visually similar at the first split. However we can give some meaning to two groups observing that the first group comprises BCC and SCC that are the two most common types of skin cancer and AK which is considered a pre-malignant condition that can give rise to SCCs and sometimes can be visually similar to early superficial BCCs. In the second group ML and SK are both benign forms of skin lesions having a similar appearance. The class grouping leads to the hierarchical structure shown in Fig. 2. This structure makes a coarse separation among classes at the upper level while finer decisions are made at a lower level. As a result, this scheme decomposes the original problem into 3 sub-problems.

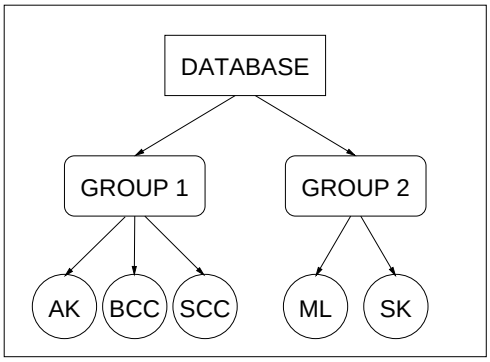


Fig. 2 Block diagram of the hierarchical organization of skin lesion classes

3 Hierarchical K-NN classifier

A large number of classifier combinations have been proposed in the literature [33]. They may have different feature sets, different training sets, different classification methods or different training sessions, all resulting in a set of

classifiers whose output may be combined, with the hope of improving the overall classification accuracy. The schemes for combining multiple classifiers can be grouped into three main categories according to their architecture: 1) parallel, 2) cascading and 3) hierarchical. In the hierarchical architecture, individual classifiers are combined into a structure which is similar to a decision tree classifier. The advantage of this architecture is the high efficiency and flexibility in exploiting the discriminant power of different types of features [33]. A large number of studies have shown that classifier combination can improve recognition accuracy [33]. It has been shown that in many domains an ensemble of classifiers outperforms any of its single components [42]. The approach used in our research falls within the hierarchical model.

Our approach divides the classification task into a set of smaller classification problems corresponding to the splits in the classification hierarchy (see Figure 2). Each of these subtasks is significantly simpler than the original task, since the classifier at a node in the hierarchy need only distinguish between a smaller number of classes. Therefore, it may be possible to separate the smaller number of classes with higher accuracy. Moreover, it may be possible to make this determination based on a smaller set of features.

The proposed approach addresses also the feature selection problem. The reduction in the feature space avoids many problems related to high dimensional feature spaces, such as the "curse of dimensionality" problem [33], where the indexing structures degrade and the significance of each feature decreases, making the process of storing, indexing and classifying extremely time consuming. Moreover, in several situations, many features are correlated, meaning that they bring redundant information about the images that can deteriorate the ability of the system to correctly distinguish them. Dimensionality reduction or feature selection has been an active research area in pattern recognition, statistics and data mining communities. The main idea of feature selection is to choose a subset of input features by eliminating features with little or no predictive information.

It is important to note that the key here is not merely the use of feature selection, but its integration with the hierarchical structure. In practice we build different classifiers using different sets of training images (according to the set of classifications made at the higher levels of the hierarchy). So each classifier uses a different set of features optimized for those images. This forces the individual classifiers to use potentially independent information.

Hierarchical classifiers are well known [45, 28, 58] and commonly used for document and text classification [20, 23, 50, 13], including a hierarchical K-NN classifier [24]. While we found papers describing applications of hierarchical systems to medical image classification and annotation tasks [47, 59, 22], to the best of our knowledge only a hierarchical neural network model has been applied to skin lesions [53]. They claim over 90% accuracy on 58 images including 4 melanomas. Unfortunately many technical details are not described in the paper. On the other hand, only poor performance was reported relative to the classification of melanoma using the K-NN method [7, 31].

Some promising results have been presented very recently by using a K-NN followed by a Decision Tree classifier [12].

3.1 K-NN classifier

K-NN is a well-known classifier. K-NN was first introduced by Fix and Hodges [25] in 1951. It is well explored in the literature and has been shown to have good classification performance on a wide range of real world data sets [17]. Many *lazy learning algorithms* are derivatives of the K-NN. A review of them is presented in the paper of Wetterschereck et al. [62]. A recent application of one of these similarity-based learning algorithms, namely the lazyCL procedure, to melanoma is described by Armengol [5].

To classify an unknown example T, the K-NN classifier finds the K nearest neighbors among the training data and uses the categories of the K neighbors to weight the category candidates. Then majority voting among the categories of data in the neighborhood is used to decide the class label of T. Given M classes $C_1, C_2, \cdots, C_M$ and N training samples $I_1, I_2, \cdots, I_N$, and the classification for I_i with respect to category $C_j (i = 1, \cdots, N; j = 1, \cdots, M)$:

$$y(I_i, C_j) \;=\; \begin{cases} 1 \; I_i \in C_j \\ 0 \; I_i \notin C_j \end{cases} \tag{1}$$

the decision rule in K-NN can be written as:

$$\text{assign } \; T \; \text{ to } \; C_j \; \text{ if } \; score(T, C_j) = \arg\max_{j=1}^{M} \sum_{i=1}^{K} y(I_i, C_j) \tag{2}$$

where the training examples I_i are ranked according to their similarity to the test example T.

The K-NN classifier has only one free parameter K which can be optimized by a leave-one-out cross-validation procedure, given the distance function *Dist* (see eq. 13) which is used to determine the 'nearest' neighbors. Choosing the properties to be used in each classifier is a core issue, and is addressed next. The actual distance metrics are presented in Sect. 4.

3.2 Learning phase

Our Hierarchical K-NN classifier (HKNN) is composed of three distinct K-NN classifier systems, one at the top level, and two at the bottom level. The top level classifier is fed with all the images in the training set. It classifies them into one of the two groups. The other two classifiers are trained using

only the images of the corresponding group (i.e. AK/BCC/SCC or ML/SK) that have been correctly (when in the training stage) classified by the top classifier, and classifies them into one of the 2 or 3 diagnostic classes.

The learning phase consists of the feature selection process for the three distinct K-NN classifiers. A sequential forward selection algorithm [34] (SFS) is used for feature selection. The goal for choosing features is the maximization of the classification accuracy. We used a weighted classification accuracy due to the uneven class distribution of our data set. This is the rate with which the system is able to correctly identify each class. Then we take an average of these rates with respect to the number of classes. Therefore our overall classification accuracy is defined as:

$$\text{Overall accuracy} = \frac{1}{M}\sum_{j=1}^{M}\frac{\text{correctly\_classified}(C_j)}{\text{number\_of\_test\_images}(C_j)} \tag{3}$$

where M is the number of classes.

A leave-one-out cross-validation method is used during feature selection. Each image is used as a test image, all the remaining images in the training set are ranked according to their similarity index to the test image. Finally the test image is classified to the class which is most frequent among the K samples nearest to it using eq. 2. The features that maximize the classification accuracy over all the images in the training set are selected among all the extracted features.

At the end, there will be three sets of features for the three classification tasks, one selected for the top classifier and two selected for the subclassifiers. The feature sets for the two subsystems are also selected using SFS, but only using images from the appropriate classes (i.e. AK/BCC/SCC or ML/SK). Note that, since every subnode in the hierarchy has only a subset of the total classes, and the subnodes each have fewer images, the additional cost of feature selection is not substantially more than that of a flat classification scheme.

3.3 Classification phase

In the classification phase all the test images are classified through the hierarchical structure. Each image is first classified into one of the two groups by the top level classifier that uses the first set of features. Then one of the classifiers of the second level is invoked according to the output group of the top classifier and therefore the image is classified in one of the 5 diagnostic classes using one of the two other subsets of features.

A drawback of the proposed method is that errors on the first classification level can not be corrected in the second level. If an example is incorrectly classified at the top level and assigned to a group that does not contain the

true class, then the classifiers at lower levels have no chance of achieving a correct classification. This is known as the "blocking" problem [58]. An attempt to solve this problem could be to use classifiers on the second level which classify to more than the two or three classes for which they are optimized. Our attempts in this direction show us that not only these classifiers gave much worse results, but also incur additional problems due to the very small number of images wrongly classified in the first level, that makes the classes more unbalanced.

4 Feature description

Here, skin lesions are characterized by their color and texture. In this section we will describe a set of features that can capture such properties.

4.1 Color features

Color features are represented by the mean colors $\mu = (\mu_R, \mu_G, \mu_B)$ of the lesion and their covariance matrices Σ. Let

$$\mu_X = \frac{1}{N}\sum_{i=1}^{N} X_i \quad \text{and} \quad C_{XY} = \frac{1}{N}\left[\sum_{i=1}^{N} X_i Y_i\right] - \mu_X \mu_Y \tag{4}$$

where: N is the number of pixels in the lesion, X_i the color component of channel X ($X, Y \in \{R, G, B\}$) of pixel i. In the RGB (Red, Green, Blue) color space, the covariance matrix is:

$$\Sigma = \begin{bmatrix} C_{RR}\ C_{RG}\ C_{RB} \\ C_{GR}\ C_{GG}\ C_{GB} \\ C_{BR}\ C_{BG}\ C_{BB} \end{bmatrix} \tag{5}$$

In this work, RGB, HSV (Hue, Saturation, Value) and CIE_Lab, CIE_Lch (Munsell color coordinate system [48]) and Otha [46] color spaces were considered. Four normalization techniques were investigated to reduce the impact of lighting, which were applied before extracting color features. In the end, we normalized each color component by dividing each color component by the average of the same component of the healthy skin of the same patient, because it had best performance compared to the other normalization techniques. After experimenting with the 5 different color spaces, we choose the normalized RGB, because it gave slightly better results than the other color spaces (see Sect. 5.4.2)

4.2 Texture features

Texture features are extracted from generalized co-occurrence matrices (GCM). Assume an image I having N_x columns, N_y rows and N_g gray levels. Let $L_x = \{1, 2, \cdots, N_x\}$ be the columns, $L_y = \{1, 2, \cdots, N_y\}$ be the rows, and $G_x = \{0, 1, \cdots, N_g - 1\}$ be the set of quantized gray levels. The co-occurrence matrix P_δ is a matrix of dimension $N_g \times N_g$, where [30]:

$$P_\delta(i,j) = \#\{((k,l),(m,n)) \in (L_y \times L_x) \times (L_y \times L_x) | I(k,l) = i, I(m,n) = j\} \tag{6}$$

i.e. the number of co-occurrences of the pair of gray levels i and j which are a distance $\delta = (d, \theta)$ apart. In our work, the pixel pairs (k,l) and (m,n) have distance $d = 5, 10, 15, 20, 25, 30$ and orientation $\theta = 0°, 45°, 90°, 135°$, i.e. $(m = k + d, n = l), (m = k + d/\sqrt{2}, n = l + d/\sqrt{2}), (m = k, n = l + d), (m = k - d/\sqrt{2}, n = l + d/\sqrt{2})$.

Generalized co-occurrence matrices are the extension of the co-occurrence matrix to multispectral images, i.e. images coded on n color channels [6]. Let u and v be two color channels. The generalized co-occurrence matrices are:

$$P_\delta^{(u,v)}(i,j) = \#\{((k,l),(m,n)) \in (L_y \times L_x) \times (L_y \times L_x) | I_u(k,l) = i, I_v(m,n) = j\} \tag{7}$$

For example, in case of color images, coded on three channels (RGB), we have six cooccurrence matrices: (RR),(GG),(BB) that are the same as gray level co-occurrence matrices computed on one channel and (RG), (RB), (GB) that take into account the correlations between the channels.

In order to have orientation invariance for our set of GCMs, we averaged the matrices with respect to θ. Quantization levels $N_G = 64, 128, 256$ are used for the three color spaces: RGB, HSV and CIE_Lab.

From each GCM we extracted 12 texture features: energy, contrast, correlation, entropy, homogeneity, inverse difference moment, cluster shade, cluster prominence, max probability, autocorrelation, dissimilarity and variance as defined in [30], for a total of 3888 texture features (12 features $\times$ 6 inter-pixel distances $\times$ 6 color pairs $\times$ 3 color spaces $\times$ 3 gray level quantisations). Two sets of texture features are extracted from GCMs calculated over the lesion area of the image, as well as over a patch of healthy skin of the same image. Differences and ratios of each of the lesion and normal skin values are also calculated, giving 2 more sets of features:

$$feature_{l-s} = feature_{lesion} - feature_{healthy\_skin} \tag{8}$$

$$feature_{l/s} = feature_{lesion} / feature_{healthy\_skin} \tag{9}$$

Altogether, for a given feature family we use $\{feature_{lesion}, feature_{healthy\_skin}, feature_{l-s}, feature_{l/s}\}$. This gives a total of $4 \times 3888 = 15552$ possible texture features, from which we extracted a good subset. All features are z-normalized over all training data.

4.3 Ad hoc color ratio features

Color ratio features are designed ad hoc for skin lesions, by observing color variations inside the lesion area. Mean colors μ_A and μ_B are extracted over the areas A and B shown in Figure 3, and their ratios calculated as:

$$ratio = \frac{\mu_A}{\mu_B} \tag{10}$$

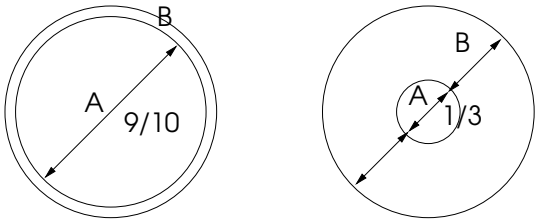


Fig. 3 Areas of lesions where ratio features where calculated.

Two different area sizes are considered. In the first case, the thickness of the border area is 10% of the area of the lesion. In the second case, the diameter of the inner area is 1/3 of the diameter of the whole lesion. Since lesions are not circular, the morphological erosion operator is applied iteratively inward from the border until the desired percentages of lesion area pixels are reached. These features seem particularly useful for BCCs, which present pearly edges.

Ad hoc color ratio features are calculated for the three color spaces: RGB, HSV and CIE_Lab, and all feature set are z-normalized. These properties are included in the texture feature set.

4.4 Distance measure

The color and texture features are combined to construct a distance measure between each test image T and a database image I.

For color covariance-based features, the Bhattacharyya distance metric:

$$BD_{CF}(T,I) = \frac{1}{8}(\mu_T - \mu_I)^T \left[\frac{(\Sigma_T + \Sigma_I)}{2}\right]^{-1} (\mu_T - \mu_I) + \frac{1}{2}\ln \frac{\left|\frac{(\Sigma_T+\Sigma_I)}{2}\right|}{\sqrt{|\Sigma_T||\Sigma_I|}} \tag{11}$$

is used, where μ_T and μ_I are the average (over all pixels in the lesion) color feature vectors, Σ_T and Σ_I are the covariance matrices of the lesion of T and I respectively, and $|\cdot|$ denotes the matrix determinant.

The Euclidean distance:

$$ED_{TF}(T,I) = \|f^T_{subset} - f^I_{subset}\| = \sqrt{\sum_{i=1}^{S}(f^T_i - f^I_i)^2} \quad (12)$$

is used for distances between a subset of S texture features f_{subset}, selected as described later. Other metric distances (mahalanobis, cityblock) have been considered, but gave worse results.

We aggregated the two distances into a distance matching function as:

$$Dist(T,I) = w \cdot BD_{CF}(T,I) + (1-w) \cdot ED_{TF}(T,I) \quad (13)$$

where w is a weighting factor that has been selected experimentally, after trying all the values: $\{0.1, 0.2, \cdots, 0.9\}$. In our case, $w = 0.7$ gave the best results. A low value of $Dist$ indicates a high similarity.

5 Methods

The features described in previous sections were extracted from the lesions in our image database. In this section we will describe in detail the image analysis and the choices of the model parameters.

5.1 Acquisition and preprocessing

Our image database comprises 960 lesions, belonging to 5 classes (45 AK, 239 BCC, 331 ML, 88 SCC, 257 SK). The ground truth used for the experiments is based on the agreed classifications by 2 dermatologists and a pathologist.

Images are acquired using a Canon EOS 350D SLR camera. Lighting was controlled using a ring flash and all images were captured at the same distance (∼50 cm) resulting in a pixel resolution of about 0.03 mm. Lesions are segmented using the region-based active contour approach described in [39]. The segmentation method uses a statistical model based the level-set framework. Morphological opening has been applied to the segmented lesions to be sure to have patches containing only lesions and healthy skin where the features are extracted.

5.2 Highlight removal

Specular highlights appear as small and bright regions in various parts of our skin images. The highlights created by specular reflections are a major obstacle for proper color and texture feature extraction.

Specular highlights are often characterized by local coincidence of intense brightness (I) and low color saturation (S). Intensity and saturation are defined as follow:

$$I = \frac{R+G+B}{3} \tag{14}$$

$$S = 1 - \frac{min(R,G,B)}{I} \tag{15}$$

and candidate specular reflection regions can be identified using appropriate threshold values (motivated by [38]):

$$I > I_{thr} \cdot I_{max} \tag{16}$$

$$S < S_{thr} \cdot S_{max} \tag{17}$$

where I_{max} are S_{max} the maximum intensity and saturation in the image respectively.

The most appropriate threshold values experimentally chosen ($I_{thr} = 0.8$) and ($S_{thr} = 0.5$) differ from the values proposed in [38] probably due to the different nature of the images.

We did not apply any subsequent filling procedure on the detected regions, as this may destroy the original texture and therefore have a negative impact of the subsequent feature extraction. Areas identified as "highlight" were simply excluded from the region where the feature extraction process takes place.

5.3 Feature normalization

The features described in previous sections have very different value ranges. To account for this, an objective rescaling of the features is achieved by normalizing to z-scores of each feature set, which is defined as

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \tag{18}$$

where: x_{ij} represents the i^{th} sample measure of feature j, μ_j the mean value of all samples for feature j and σ_j is the standard deviation of the samples for feature j.

In addition, feature values outside the values at 5-95 percentiles have been truncated to the 5^{th} or 95^{th} percentile value, and the normalising μ and σ calculated from the truncated set. The normalising parameters were constant over all experiments.

5.4 Evaluation

To assess performance, training and test sets were created by randomly splitting the data set into 3 equal subsets. The only constraint on otherwise random partitioning was that a class was represented equally in each subset. A 3-fold cross-validation method was used, i.e. 3 sets composed of two-thirds of the data were created and used as training sets for feature selection and the remaining one-third of the data as the test set using the selected features for classification. Thus no training example used for feature selection was used as a test example in the same experiment. Three experiments were conducted independently and performance reported as mean and standard deviation over the three experiments.

In the hierarchical classifiers mentioned in previous sections, the most commonly used performance measures are the classic information retrieval notions of precision and recall, or a combination of the two measures [58]. As we are dealing with a classification task and not a retrieval task, we use the classification accuracy derived from the confusion matrix. In the training stage, confusion matrices are obtained by a leave-one-out scheme, where each image is used as a test image and classified according the known classification of the remaining images in the training set. On the other hand, in the classification stage, confusion matrices are obtained in a slightly different way: each image of the test set is classified according to the known classifications of the K nearest neighbors in the training set.

5.4.1 Influence of the K parameter

Classification results when varying the value of K of the K-NN classifiers have been evaluated. In some experiments we noticed a little improvement by using a smaller value of K for feature selection and a bigger one for classification. Table 1 shows our evaluation. The numbers (mean $\pm$ standard deviation of the accuracy over the three sets) in the first column are obtained in the feature selection stage, *i.e.* using the value of K written on their left. The highest classification accuracy over the test sets for each value of K used during the feature selection are highlighted in boldface.

We chose values of K: 1) to be odd numbers, 2) to be smaller than the training class sizes and 3) to span what seemed like a sensible range. Since performance does not vary too much for the K=11 or K=15 test cases, any value of K in this range is probably approximately equally effective. In the following, the presented results are obtained using the combination of K that gave the best classification accuracy (underlined in the table) on the test set for each subclassifier (top level classifier: train K=15, test K=15; AK/BCC/SCC classifier: train K=15, test K=11; ML/SK classifier: train K=11, test K=15). Recalling that K is the number of nearest samples used to classify the image under examination, it is technically correct to use different

Table 1 Accuracy of the three subclassifiers varying the value of K. Each row shows the value of K used in training, columns show the K used in testing.

(a) Top level

	Training Set	*Test Set*		
		K=7	*K=11*	*K=15*
K=7	95.80 ± 0.53	91.67 ± 0.93	**92.09 ± 1.59**	91.88 ± 1.23
K=11	95.68 ± 0.18	**93.33 ± 0.67**	92.71 ± 1.17	92.61 ± 0.74
K=15	95.73 ± 0.63	93.23 ± 1.42	93.33 ± 0.95	**93.86 ± 0.72**

(b) Group1 (AK,BCC,SCC)

	Training Set	*Test Set*		
		K=7	*K=11*	*K=15*
K=7	79.40 ± 0.75	69.48 ± 0.98	**71.50 ± 1.28**	70.95 ± 2.31
K=11	79.96 ± 3.40	69.07 ± 3.38	70.04 ± 0.88	**70.86 ± 1.14**
K=15	81.87 ± 3.62	70.87 ± 0.91	**72.64 ± 2.41**	71.79 ± 2.06

(c) Group2 (ML,SK)

	Training Set	*Test Set*		
		K=7	*K=11*	*K=15*
K=7	91.97 ± 0.42	85.82 ± 0.88	**86.01 ± 0.86**	85.82 ± 0.39
K=11	91.88 ± 0.54	85.64 ± 0.58	86.00 ± 0.70	**86.19 ± 0.59**
K=15	90.80 ± 1.20	84.55 ± 0.86	**85.84 ± 0.81**	85.67 ± 1.43

values at the classification stage, than those used during the feature selection stage.

5.4.2 Influence of color features

A comparison of the accuracy (mean ± standard deviation over the three sets) of the three subclassifiers using only color features is reported in Table 2, using the K values reported in the previous section. Note that values for RGB are different from Table 1 because texture features are not used here.

The best results are obtained using RGB and Otha color spaces. Actually all accuracies are nearly identical before normalization. After normalization, RGB and Otha color spaces still give best results for the top classifier, while RGB gives much better results for the other two subclassifiers. These data also indicate that color features are more important at the top level of the hierarchy, i.e. in discriminating cancerous vs non cancerous lesions.

Table 2 Accuracy of the three subclassifiers over the three sets using different color spaces, before and after normalization.

(a) Before color normalization

	Top level	*Group1*	*Group2*
RGB	87.82 ± 2.14	64.37 ± 3.80	55.03 ± 1.31
HSV	87.81 ± 2.21	62.61 ± 2.88	55.03 ± 1.31
Lab	87.20 ± 2.68	63.47 ± 2.73	55.95 ± 1.33
Lch	86.46 ± 2.52	63.48 ± 2.91	55.05 ± 0.62
Otha	87.82 ± 0.97	64.37 ± 3.80	55.85 ± 1.37

(b) After color normalization

	Top level	*Group1*	*Group2*
RGB	92.71 ± 0.66	74.38 ± 1.81	84.35 ± 1.19
HSV	89.80 ± 1.95	62.65 ± 4.16	54.45 ± 2.60
Lab	91.04 ± 1.45	62.93 ± 3.68	56.87 ± 1.94
Lch	87.71 ± 1.80	65.23 ± 3.57	53.54 ± 2.40
Otha	92.71 ± 0.66	62.31 ± 2.71	57.79 ± 3.40

5.4.3 Influence of texture features

The texture feature set that best discriminates between the groups at the first level of the hierarchy is different from the feature sets that best discriminate at the second level, and these two sets also differ between each other. Fig. 4 shows a scatter plot of the two top features for each classifier. The list of selected features for each level of the hierarchy is reported in the Appendix (see Table 8). Considering that a potentially very different set of features is selected at each node of the hierarchy, we can say that the hierarchical method, as a whole, actually uses a larger set of features in a more efficient way, without ending up in problems like the "curse of dimensionality". Hence, there is a benefit from the hierarchical scheme.

We can observe that color information is important also in the texture features because texture properties extracted using different color channels are selected.

The plots of the accuracy vs the number of features (from 1 to 10 for each level of the hierarchy) are shown in Fig. 5. We show only the plots for one of the three subsets and for the best K combinations. Keeping in mind that the color features are fixed and feature selection is applied only to texture features, the nearly flat trend of the top level classifier (Fig. 5 top) confirms that color features are more important in discriminating AK/BCC/SC from ML/SK, and adding more texture features does not improve its performance more than 2%. On the other hand, the trend of the AK/BCC/SCC and ML/SK subclassifiers (Fig. 5 bottom) indicates the usefulness of using texture features at this level of the hierarchy.

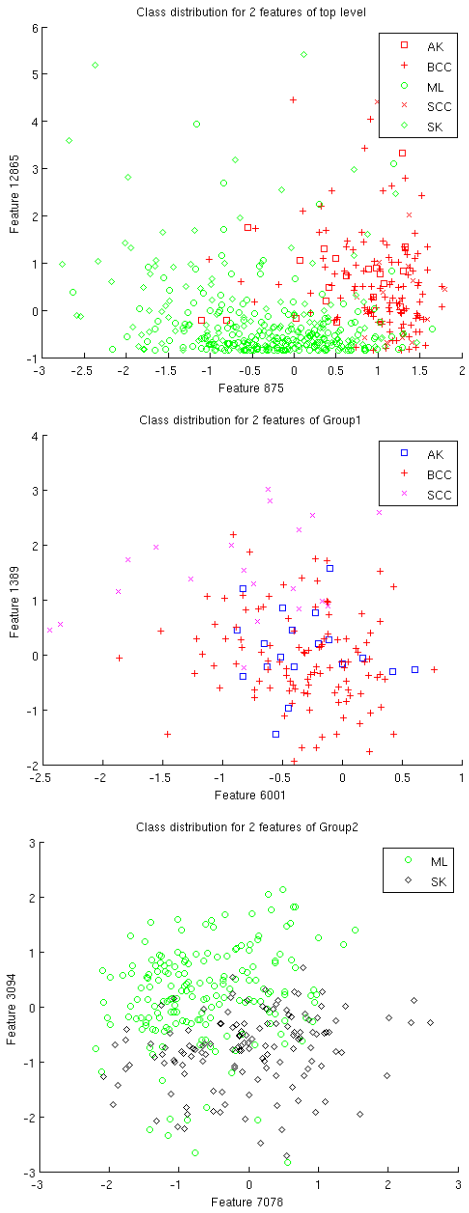


Fig. 4 Scatter plots of the top 2 features for each of the three sets. Top graph shows *Group1* (AK/BCC/SCC) in red and *Group2* (SK/ML) in green.

5.4.4 Influence of feature number and selection algorithm

Referring again to the plots shown in Fig. 5, we can see that is reasonable to stop after adding 10 texture features to color features, as the accuracy on the test set was not significantly improving anymore.

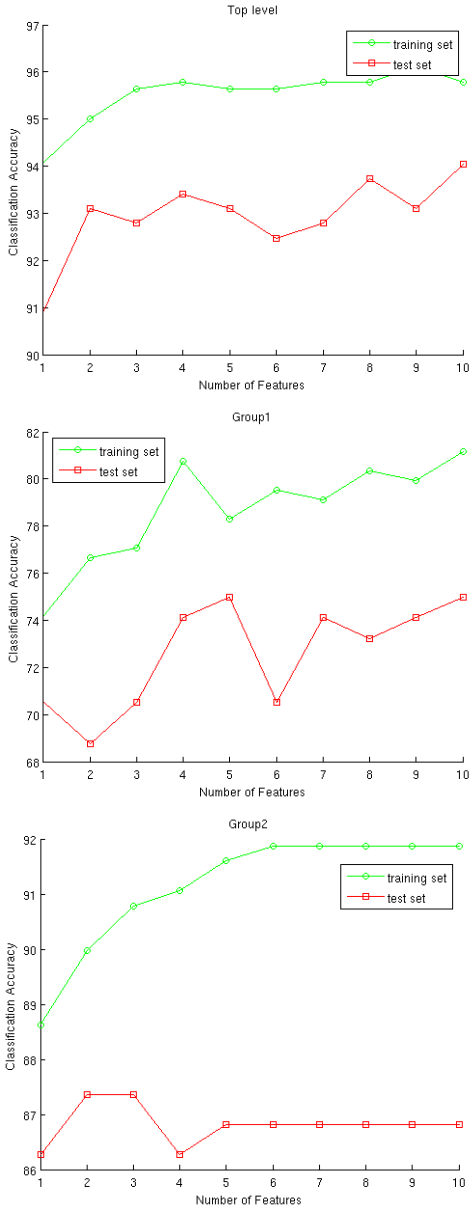


Fig. 5 Plots of accuracy vs number of texure features for one of the 3 subsets, using color feature + 1 to 10 texture features.

A slight overfitting problem evident in some plots suggested us to make experiments using the three sets as train, validation, test sets respectively. Results (mean ± standard deviation of the accuracy over the three subsets)

are reported in Table 3. We stopped selecting additional features when the accuracy on the validation set decreased (once in the top table, twice in the bottom one).

Table 3 Accuracy of the three subclassifiers over the three training sets, validation sets and test sets.

(a) Stop when validation accuracy decrease once

	Top level	*Group1*	*Group2*
Training set	94.06 ± 1.37	71.63 ± 6.20	87.84 ± 1.01
Validation set	91.88 ± 0.81	69.16 ± 0.36	85.99 ± 0.39
Test set	91.56 ± 1.06	71.49 ± 1.91	86.01 ± .33
# Features	5,5,3	6,9,3	3,3,14

(b) Stop when validation accuracy decrease twice

	Top level	*Group1*	*Group2*
Training set	94.06 ± 1.90	70.77 ± 6.17	87.56 ± 1.45
Validation set	92.40 ± 0.77	69.69 ± 1.92	86.55 ± 0.70
Test set	90.83 ± 0.94	71.51 ± 2.70	85.32 ± 3.08
# Features	16,13,2	5,18,2	18,2,12

We did not notice any significant improvement. This is probably due to the smaller training set size that further reduced the size of the smallest class.

The number of features selected for each of the three subsets is in the last row of the tables. The high variation means the number of selected features is not a crucial choice. Indeed, the three subsets are created by randomly splitting the data in such a way that the 5 lesion classes were equally represented in each subset.

The SFS feature selection algorithm is claimed not to be the optimal algorithm, however in our case the use of a sequential forward backward greedy algorithm (see Table 4) did not show any significant improvement. Once again we note a high variation in the number of selected features for the three subsets.

Table 4 Accuracy of the three subclassifiers over the three training sets, validation sets and test sets using a greedy forward backward algorithm

	Top level	*Group1*	*Group2*
Training set	95.35 ± 1.73	80.25 ± 5.09	91.35 ± 0.40
Validation set	91.67 ± 0.45	67.19 ± 1.67	85.45 ± 0.78
Test set	90.52 ± 0.97	72.08 ± 2.92	86.75 ± 2.55
# Features	5,20,10	10,20,4	20,4,20

5.5 Comparison with other methods

In Table 5 we compare our results with the results obtained using a non hierarchical approach, i.e. a flat K-NN classifier and a Bayes classifier that use a single set of features for all the 5 classes. The flat classifiers were trained using features selected using the same SFS algorithm. Results of a hierarchical Bayes classifier, having the same hierarchy as the HKNN classifier and whose subclassifiers were trained using the same features and the same SFS algorithm, are also reported in the table. We see that the use of hierarchy gives an improvement both over the training and test sets.

Table 5 Comparison of the overall percentage accuracy of the hierarchical and flat classifiers over the three training sets and test sets.

	Flat KNN	*HKNN*	*Flat Bayes*	*Hierarc. Bayes*
Training set	77.6 ± 1.4	83.4 ± 1.4	74.3 ± 2.2	81.9 ± 1.5
Test set	69.8 ± 1.6	74.3 ± 2.5	67.7 ± 2.3	69.6 ± 0.4

6 Overall results

The final results are reported in Table 6. The final accuracy of the top classifier and the two subclassifiers at the bottom levels are also reported here. The values are the mean ± standard deviation over the three training and test sets. These results are obtained using best combination of K determined in Sect. 5.4.1, the RGB color features and 10 texture features for each subclassifier. We decided to use a fixed number of features as the train-validation-test scheme did not enhance performance (see considerations in Sect. 5.4.4 about set sizes and number of features). Similarly, the variety of results from the different configurations and numbers of features all have about the same level, given the estimated standard deviations, and so suggest that there is little risk of overtraining.

Recall the top level classifier discriminates between cancer and premalignant conditions (AK/BCC/SCC) and benign forms of skin lesions (ML/SK). Therefore, its very high accuracy (above 93%) indicates the good performance of our system in identifying cancer and potential at risk conditions. Analysis of the wrongly classified images at the top level pointed out that these were the lesions on which clinical diagnosis of experienced dermatologists was most uncertain.

Table 6 Accuracy of the three subclassifiers and combined classifier over the three training sets and test sets. Note that the Group1/2 results are only over the lesions correctly classified at the top level. On the other hand, the full classifier results report accuracy based on both levels.

	Top level	*Group1*	*Group2*	*Full classifier*
Training set	95.7 ± 0.6	81.9 ± 3.6	91.9 ± 0.5	83.4 ± 1.4
Test set	93.9 ± 0.7	72.6 ± 2.4	86.2 ± 0.6	74.3 ± 2.5

The overall classification accuracy on the test set is $74.3 \pm 2.5\%$, as shown in the right column of Table 6. The overall result also includes the $\sim$6% misclassified samples from the first level.

The overall performance (74%) is not yet at the 90% level (achieved after 20+ years of research) for differential diagnosis of moles versus melanoma, however, our method addresses lesion classes that seem to have no previous automated image analysis (outside of research from our group [39, 9, 8, 36, 65, 4, 2, 3]) and, as highlighted previously, our algorithms' performance is above the diagnostic accuracy currently being achieved by non-specialists.

Table 7 shows the confusion matrix of the whole HKNN system on the test images. This matrix has been obtained by adding the three confusion matrices from the three test sets, as they are disjoint. We note a good percentage of correctly classified BCC, ML and SK. The number of correctly classified AK and SCC at a first glance looks quite low. This is due to the small number of images in each of these two classes. However most of the AKs are misclassified as BCC and we should remember that AK is a pre-malignant lesion. Also many SCC are classified as BCC which is another kind of cancer. Therefore consequences of these mistakes are not as dramatic as if they were diagnosed as benign. An additional split in the hierarchy may improve results.

Table 7 Classification results: confusion matrix on the test images. Rows are true classes, columns are the selected classes.

	AK	**BCC**	**ML**	**SCC**	**SK**
AK	7	27	1	9	1
BCC	2	210	6	14	7
ML	10	10	269	10	42
SCC	8	34	5	36	5
SK	9	8	33	8	199

7 Conclusions

We have presented an algorithm based on a novel hierarchical K-NN classifier, and its application as the first classification of 5 most common classes of non-melanoma skin lesion from color images. Our approach uses a hierarchical combination of three classifiers, utilizing feature selection to tailor the feature set of each classifier to its task. The hierarchical K-NN structure improves the performance of the system over the flat K-NN and a Bayes classifier.

As the accuracy is above 70%, this system could be used in the future as a diagnostic aid for skin lesion images, particularly as the cancerous vs non cancerous results are ∼94%.

These results were produced by optimizing classification accuracy. For medical use future research should include the cost of decisions into the optimization process.

Further studies will include the extraction of other texture related features, the evaluation of other feature selection methods and the use of a weighted K-NN model, where neighbor images are weighted according their distance to the test image. In the future, it would be interesting to extend the hierarchical approach to more than two hierarchical levels, including self-learned hierarchies.

8 Appendix

List of texture features selected for each level of the final tree.

Acknowledgment

We thank the Wellcome Trust for funding this project (Grant No: 083928/Z/07/Z).

References

1. Alcón, J.F., Heinrich, A., Uzunbajakava, N., Krekels, G., Siem, D., de Haan, G., de Haan, G.: Automatic imaging system with decision support for inspection of pigmented skin lesions and melanoma diagnosis. IEEE Journal of Selected Topics in Signal Processing **3**, 14–25 (2009). DOI 10.1109/JSTSP.2008.2011156
2. Aldridge, R.B., Glodzik, D., Ballerini, L., Fisher, R.B., Rees, J.L.: The utility of non-rule-based visual matching as a strategy to allow novices to achieve skin lesion diagnosis. Acta Dermato-Venereologica **91**, 279–283 (2011)
3. Aldridge, R.B., Li, X., Ballerini, L., Fisher, R.B., Rees, J.L.: Teaching dermatology using 3-dimensional virtual reality. Archives of Dermatology **149**(10) (2010)

Table 8 Legend: R=Red, G=Green, B=Blue, H=Hue, S=Saturation, V=Value, L,a,b= Lab color space. Texture features are defined in [30]

(a) Top level

texture feature	interp. distance	quant. levels	color channels	site
Entropy	5	128	HV	lesion
Cluster Shade	10	128	Lb	skin
Inv. Diff. Moment	5	64	SV	lesion
Contrast	15	64	ab	skin
Energy	5	64	HV	lesion
Inv. Diff. Moment	15	64	RG	lesion
Max Probability	5	128	HH	lesion/skin
Cluster Prominence	25	64	GG	skin
Correlation	10	256	HV	lesion
Cluster Prominence	15	64	LL	lesion-skin

(b) Group 1 (AK,BCC,SCC)

texture feature	interp. distance	quant. levels	color channels	site
Variance	30	64	HS	lesion/skin
Energy	25	256	ab	lesion
Inv. Diff. Moment	25	64	BB	lesion
Entropy	15	128	HH	lesion
Max Probability	5	256	RB	lesion/skin
Cluster Prominence	10	64	SS	lesion-skin
Contrast	5	64	SS	lesion
Homogeneity	15	256	RG	lesion
Homogeneity	25	128	SV	lesion/skin
Inv. Diff. Moment	5	64	HS	skin

(c) Group 2 (ML,SK)

texture feature	interp. distance	quant. levels	color channels	site
Correlation	10	256	SS	lesion
Dissimilarity	5	64	bb	skin
Cluster Shade	5	64	HH	lesion
Cluster Shade	5	64	HV	lesion
Cluster Prominence	5	64	HH	lesion
Cluster Prominence	5	64	HS	lesion
Cluster Shade	10	64	HV	lesion
Cluster Shade	5	256	HV	lesion
Correlation	5	64	HV	skin
Contrast	5	64	HH	lesion

4. Aldridge, R.B., Zanotto, M., Ballerini, L., Fisher, R.B., Rees, J.L.: Novice identification of melanoma: not quite as straightforward as the ABCDs. Acta Dermato-Venereologica **91**, 125–130 (2011)
5. Armengol, E.: Classification of melanomas in situ using knowledge discovery with explained case-based reasoning. Artif. Intell. Med. **51**, 93–105 (2011). DOI http://dx.doi.org/10.1016/j.artmed.2010.09.001. URL http://dx.doi.org/10.1016/j.artmed.2010.09.001
6. Arvis, V., Debain, C., Berducat, M., Benassi, A.: Generalization of the cooccurence matrix for colour images: Application to colour texture classification. Image Analysis and Stereology **23**(1), 63–72 (2004)
7. Aslandogan, Y., Mahajani, G.: Evidence combination in medical data mining. In: Proc. of International Conference on Information Technology: Coding and Computing, vol. 2, pp. 465 – 469 (2004). DOI 10.1109/ITCC.2004.1286697
8. Ballerini, L., Li, X., Fisher, R.B., Aldridge, B., Rees, J.: Content-based image retrieval of skin lesions by evolutionary feature synthesis. In: C. di Chio et al. (ed.) Application of Evolutionary Computation, no. 6024 in Lectures Notes in Computer Science, pp. 312–319. Istanbul, Turkey (2010)
9. Ballerini, L., Li, X., Fisher, R.B., Rees, J.: A query-by-example content-based image retrieval system of non-melanoma skin lesions. In: B. Caputo (ed.) Proc. MICCAI-09 Workshop MCBR-CDS 2009: Medical Content-based Retrieval for Clinical Decision Support, no. 5853 in LNCS, pp. 31–38. Springer-Verlag Berlin Heidelberg (2010)
10. Basarab, T., Munn, S., Jones, R.R.: Diagnostic accuracy and appropriateness of general practitioner referrals to a dermatology out-patient clinic. British Journal of Dermatology **135**(1), 70–73 (1996). DOI 10.1046/j.1365-2133.1996.d01-935.x. URL http://dx.doi.org/10.1046/j.1365-2133.1996.d01-935.x
11. Cascinelli, N., Ferrario, M., Tonelli, T., Leo, E.: A possible new tool for clinical diagnosis of melanoma: The computer. Journal of the American Academy of Dermatology **16**(2, Part 1), 361 – 367 (1987). DOI 10.1016/S0190-9622(87)70050-4
12. Cavalcanti, P.G., Scharcanski, J.: Automated prescreening of pigmented skin lesions using standard cameras. Computerized Medical Imaging and Graphics **35**(6), 481 – 491 (2011). DOI 10.1016/j.compmedimag.2011.02.007. URL http://www.sciencedirect.com/science/article/pii/S0895611111000395
13. Ceci, M., Malerba, D.: Hierarchical classification of html documents with webclassii. In: In Proc. of the 25th European Conf. on Information Retrieval, pp. 57–72 (2003)
14. Celebi, M.E., Iyatomi, H., Schaefer, G., Stoecker, W.V.: Lesion border detection in dermoscopy images. Computerized Medical Imaging and Graphics **33**(2), 148–153 (2009)
15. Celebi, M.E., Kingravi, H.A., Uddin, B., Iyatomi, H., Aslandogan, Y.A., Stoecker, W.V., Moss, R.H.: A methodological approach to the classification of dermoscopy images. Computerized Medical Imaging and Graphics **31**(6), 362 – 373 (2007). DOI DOI: 10.1016/j.compmedimag.2007.01.003
16. Celebi, M.E., Stoecker, W.V., Moss, R.H.: Advances in skin cancer image analysis. Computerized Medical Imaging and Graphics **35**(2), 83 – 84 (2011). DOI 10.1016/j.compmedimag.2010.11.005
17. Cover, T., Hart, P.: Nearest neighbor pattern classification. Information Theory, IEEE Transactions on **13**(1), 21 – 27 (1967). DOI 10.1109/TIT.1967.1053964
18. Cancer research UK (CRUK). CancerStats, Internet (2011). URL http://info.cancerresearchuk.org/cancerstats. Accessed 03/08/2011
19. Dalal, A., Moss, R.H., Stanley, R.J., Stoecker, W.V., Gupta, K., Calcara, D.A., Xu, J., Shrestha, B., Drugge, R., Malters, J.M., Perry, L.A.: Concentric decile segmentation of white and hypopigmented areas in dermoscopy images of skin lesions allows discrimination of malignant melanoma. Computerized Medical Imaging and Graphics **35**(2), 148 – 154 (2011). DOI 10.1016/j.compmedimag.2010.09.009. URL http://www.sciencedirect.com/science/article/pii/S0895611110001035

20. D'Alessio, S., Murray, K., Schiaffino, R., Kershenbaum, A.: The effect of using hierarchical classifiers in text categorization. In: Proc. of 6th International Conference Recherche d'Information Assistee par Ordinateur, pp. 302–313 (2000)
21. Day, G.R., Barbour, R.H.: Automated melanoma diagnosis: where are we at? Skin Research and Technology **6**, 1–5 (2000)
22. Dimitrovski, I., Kocev, D., Loskovska, S., Deroski, S.: Hierarchical annotation of medical images. Pattern Recognition **44**(10-11), 2436 – 2449 (2011). DOI 10.1016/j.patcog.2011.03.026. URL http://www.sciencedirect.com/science/article/pii/S0031320311001300
23. Dumais, S., Chen, H.: Hierarchical classification of web content. In: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 256–263. ACM, New York, NY, USA (2000). DOI http://doi.acm.org/10.1145/345508.345593. URL http://doi.acm.org/10.1145/345508.345593
24. Duwairi, R., Al-Zubaidi, R.: A hierarchical K-NN classifier for textual data. The International Arab Journal of Information Technology **8**(3), 251–259 (2011)
25. Fix, E., Hodges J. L., J.: Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. International Statistical Review / Revue Internationale de Statistique **57**(3), 238–247 (1989). URL http://www.jstor.org/stable/1403797
26. Garnavi, R., Aldeen, M., Celebi, M.E., Varigos, G., Finch, S.: Border detection in dermoscopy images using hybrid thresholding on optimized color channels. Computerized Medical Imaging and Graphics **35**(2), 105 – 115 (2011). DOI 10.1016/j.compmedimag.2010.08.001. URL http://www.sciencedirect.com/science/article/pii/S0895611110000819
27. Gerbert, B., Maurer, T., Berger, T., Pantilat, S., McPhee, S.J., Wolff, M., Bronstone, A., Caspers, N.: Primary care physicians as gatekeepers in managed care: Primary care physicians' and dermatologists' skills at secondary prevention of skin cancer. Arch Dermatol **132**(9), 1030–1038 (1996). DOI 10.1001/archderm.1996.03890330044008
28. Gordon, A.D.: A review of hierarchical classification. Journal of the Royal Statistical Society. Series A (General) **150**(2), 119–137 (1987). URL http://www.jstor.org/stable/2981629
29. Green, A., Martin, N., McKenzie, G., Pfitzner, J., Quintarelli, F., Thomas, B.W., O'Rourke, M., Knight, N.: Computer image analysis of pigmented skin lesions. Melanoma Research **1**, 231–236 (1991)
30. Haralick, R.M., Shanmungam, K., Dinstein, I.: Textural features for image classification. IEEE Transactions on Systems, Man and Cybernetics **3**(6), 610–621 (1973)
31. Hintz-madsen, M., Hansen, L.K., Larsen, J., E., O., Drzewiecki, K.T.: Design and evaluation of neural classifiers application to skin lesion classification. In: Proceedings of the 1995 IEEE Workshop on Neural Networks for Signal Processing V, pp. 484–493 (1995)
32. Iyatomi, H., Celebi, M.E., Schaefer, G., Tanaka, M.: Automated color calibration method for dermoscopy images. Computerized Medical Imaging and Graphics **35**(2), 89 – 98 (2011). DOI 10.1016/j.compmedimag.2010.08.003. URL http://www.sciencedirect.com/science/article/pii/S0895611110000832
33. Jain, A.K., Duin, R.P.W., Mao, J.: Statistical pattern recognition: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(1), 4–37 (2000)
34. Jain, A.K., Zongker, D.: Feature-selection: Evaluation, application, and small sample performance. IEEE Transactions on Pattern Analysis and Machine Intelligence **19**(2), 153–158 (1997)
35. Ko, C.B., Walton, S., Keczkes, K., Bury, H.P.R., Nicholson, C.: The emerging epidemic of skin cancer. British Journal of Dermatology **130**, 269–272 (1994)
36. Laskaris, N., Ballerini, L., Fisher, R.B., Aldridge, B., Rees, J.: Fuzzy description of skin lesions. In: D.J. Manning, C.K. Abbey (eds.) Medical Imaging 2010: Image Perception, Observer Performance, and Technology Assessment, *Proceedings of the SPIE*, vol. 7627, pp. 762,717–1 to 762,717–10 (2010)

37. Lee, T.K., Claridge, E.: Predictive power of irregular border shapes for malignant melanomas. Skin Research and Technology **11**(1), 1–8 (2005)
38. Lehmann, T.M., Palm, C.: Color line search for illuminant estimation in real-world scenes. J. Opt. Soc. Am. A **18**(11), 2679–2691 (2001). DOI 10.1364/JOSAA.18.002679. URL http://josaa.osa.org/abstract.cfm?URI=josaa-18-11-2679
39. Li, X., Aldridge, B., Ballerini, L., Fisher, R., Rees, J.: Depth data improves skin lesion segmentation. In: G.Z.Y. et al. (ed.) Proc. 12th Int. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI), London, pp. 1100–1107 (2009)
40. Maglogiannis, I., Doukas, C.N.: Overview of advanced computer vision systems for skin lesions characterization. IEEE Transactions on Information Technology in Biomedicine **13**(5), 721–733 (2009). DOI http://dx.doi.org/10.1109/TITB.2009.2017529
41. Maglogiannis, I., Pavlopoulos, S., Koutsouris, D.: An integrated computer supported acquisition, handling, and characterization system for pigmented skin lesions in dermatological images. IEEE Transactions on Information Technology in Biomedicine **9**(1), 86–98 (2005)
42. Martínez-Otzeta, J.M., Sierra, B., Lazkano, E., Astigarraga, A.: Classifier hierarchy learning by means of genetic algorithms. Pattern Recognition Letters **27**(16), 1998–2004 (2006)
43. Mete, M., Kockara, S., Aydin, K.: Fast density-based lesion detection in dermoscopy images. Computerized Medical Imaging and Graphics **35**(2), 128 – 136 (2011). DOI 10.1016/j.compmedimag.2010.07.007. URL http://www.sciencedirect.com/science/article/pii/S0895611110000789
44. Morrison, A., O'Loughlin, S., Powell, F.C.: Suspected skin malignancy: a comparison of diagnoses of family practitioners and dermatologists in 493 patients. International Journal of Dermatology **40**(2), 104–107 (2001). DOI 10.1046/j.1365-4362.2001.01159.x. URL http://dx.doi.org/10.1046/j.1365-4362.2001.01159.x
45. Murtagh, F.: A survey of recent advances in hierarchical clustering algorithms. The Computer Journal **26**(4), 354–359 (1983)
46. Ohta, Y.I., Kanade, T., Sakai, T.: Color information for region segmentation. Computer Graphics and Image Processing **13**(1), 222 – 241 (1980)
47. Pourghassem, H., Ghassemian, H.: Content-based medical image classification using a new hierarchical merging scheme. Computerized Medical Imaging and Graphics **32**(8), 651 – 661 (2008). DOI DOI: 10.1016/j.compmedimag.2008.07.006
48. Rahman, M.M., Desai, B.C., Bhattacharya, P.: Image retrieval-based decision support system for dermatoscopic images. In: IEEE Symposium on Computer-Based Medical Systems, pp. 285–290. IEEE Computer Society, Los Alamitos, CA, USA (2006). DOI http://doi.ieeecomputersociety.org/10.1109/CBMS.2006.98
49. Rigel, D.S., Russak, J., Friedman, R.: The evolution of melanoma diagnosis: 25 years beyond the ABCDs. CA: A Cancer Journal for Clinicians **60**(5), 301–316 (2010). DOI 10.3322/caac.20074. URL http://dx.doi.org/10.3322/caac.20074
50. Rodriguez, C., Boto, F., Soraluze, I., Pérez, A.: An incremental and hierarchical K-NN classifier for handwritten characters. In: Proceedings of the 16th International Conference on Pattern Recognition (ICPR'02) Volume 3, pp. 98–101. IEEE Computer Society, Washington, DC, USA (2002)
51. Rosado, B., Menzies, S., Harbauer, A., Pehamberger, H., Wolff, K., Binder, M., Kittler, H.: Accuracy of computer diagnosis of melanoma: A quantitative meta-analysis. Arch Dermatol **139**(3), 361–367 (2003). DOI 10.1001/archderm.139.3.361. URL http://archderm.ama-assn.org/cgi/content/abstract/139/3/361
52. Sadeghi, M., Razmara, M., Lee, T.K., Atkins, M.: A novel method for detection of pigment network in dermoscopic images using graphs. Computerized Medical Imaging and Graphics **35**(2), 137 – 143 (2011). DOI 10.1016/j.compmedimag.2010.07.002. URL http://www.sciencedirect.com/science/article/pii/S0895611110000674
53. Salah, B., Alshraideh, M., Beidas, R., Hayajneh, F.: Skin cancer recognition by using a neuro-fuzzy system. Cancer Informatics **10**, 1–11 (2011)

54. Schaefer, G., Rajab, M.I., Celebi, M.E., Iyatomi, H.: Colour and contrast enhancement for improved skin lesion segmentation. Computerized Medical Imaging and Graphics **35**(2), 99 – 104 (2011). DOI 10.1016/j.compmedimag.2010.08.004. URL http://www.sciencedirect.com/science/article/pii/S0895611110000844
55. Schmid-Saugeons, P., Guillod, J., Thiran, J.P.: Towards a computer-aided diagnosis system for pigmented skin lesions. Computerized Medical Imaging and Graphics **27**, 65–78 (2003)
56. Seidenari, S., Pellacani, G., Pepe, P.: Digital videomicroscopy improves diagnostic accuracy for melanoma. Journal of the American Academy of Dermatology **39**(2), 175–181 (1998)
57. Stoecker, W.V., Wronkiewiecz, M., Chowdhury, R., Stanley, R.J., Xu, J., Bangert, A., Shrestha, B., Calcara, D.A., Rabinovitz, H.S., Oliviero, M., Ahmed, F., Perry, L.A., Drugge, R.: Detection of granularity in dermoscopy images of malignant melanoma using color and texture features. Computerized Medical Imaging and Graphics **35**(2), 144 – 147 (2011). DOI 10.1016/j.compmedimag.2010.09.005. URL http://www.sciencedirect.com/science/article/pii/S0895611110000996
58. Sun, A., Lim, E.P., Ng, W.K.: Performance measurement framework for hierarchical text classification. Journal of the American Society for Information Science and Technology **54**, 1014–1028 (2003)
59. Tommasi, T., Dedelaers, T.: The Medical Image Classification Task, vol. 32, chap. 12, pp. 221–238. ImageCLEF: The Information Retrieval Series (2010)
60. Viola, K.V., Tolpinrud, W.L., Gross, C.P., Kirsner, R.S., Imaeda, S., Federman, D.G.: Outcomes of referral to dermatology for suspicious lesions: Implications for teledermatology. Arch Dermatol **147**(5), 556–560 (2011). DOI 10.1001/archdermatol.2011.108
61. Wang, H., Moss, R.H., Chen, X., Stanley, R.J., Stoecker, W.V., Celebi, M.E., Malters, J.M., Grichnik, J.M., Marghoob, A.A., Rabinovitz, H.S., Menzies, S.W., Szalapski, T.M.: Modified watershed technique and post-processing for segmentation of skin lesions in dermoscopy images. Computerized Medical Imaging and Graphics **35**(2), 116 – 120 (2011). DOI 10.1016/j.compmedimag.2010.09.006. URL http://www.sciencedirect.com/science/article/pii/S089561111000100X
62. Wettschereck, D., Aha, D.W., Mohri, T.: A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. Artificial Intelligence Review **11**, 273–314 (1997)
63. Wollina, U., Burroni, M., Torricelli, R., Gilardi, S., Dell'Eva, G., Helm, C., Bardey, W.: Digital dermoscopy in clinical practise: a three-centre analysis. Skin Research and Technology **13**, 133–142(10) (May 2007). DOI 10.1111/j.1600-0846.2007.00219.x
64. Zanotto, M.: Visual description of skin lesions. Master's thesis, School of Informatics – University of Edinburgh (2010)
65. Zanotto, M., Ballerini, L., Aldridge, B., Fisher, R.B., Rees, J.: Visual cues do not improve skin lesion ABC(D) grading. In: D.J. Manning, C.K. Abbey (eds.) Medical Imaging 2011: Image Perception, Observer Performance, and Technology Assessment, *Proceedings of the SPIE*, vol. 7966, pp. 79,660U–1 – 79,660U–10 (2011)
66. Zhou, H., Schaefer, G., Celebi, M.E., Lin, F., Liu, T.: Gradient vector flow with mean shift for skin lesion segmentation. Computerized Medical Imaging and Graphics **35**(2), 121 – 127 (2011). DOI 10.1016/j.compmedimag.2010.08.002. URL http://www.sciencedirect.com/science/article/pii/S0895611110000820