NON-MELANOMA SKIN LESION CLASSIFICATION USING COLOUR IMAGE DATA IN A HIERARCHICAL K-NN CLASSIFIER

Lucia Ballerini[⋆] *Robert B. Fisher*[⋆] *Ben Aldridge*[†] *Jonathan Rees*[†]

[⋆]School of Informatics, University of Edinburgh, Edinburgh, UK
[†]Dermatology, University of Edinburgh, Edinburgh, UK

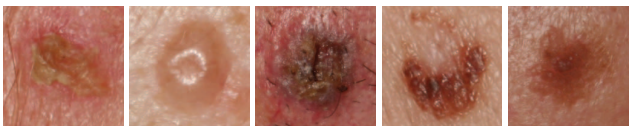
ABSTRACT

This paper presents an algorithm for classification of non-melanoma skin lesions based on a novel hierarchical K-Nearest Neighbors (K-NN) classifier. The K-NN classifier is simple, quick and effective. 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. Colour and texture features are extracted from skin lesions. The accuracy of the proposed hierarchical scheme is higher than 93% in discriminating cancer and pre-malignant 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 published result on non-melanoma skin cancer classification from colour images acquired by a standard camera (non-dermoscopy).

Index Terms— skin lesion images, hierarchical framework, K-NN classifier, skin cancer

1. INTRODUCTION

There are a considerable number of published studies on classification methods related to the diagnosis of cutaneous malignancies. Since 1987 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 summarises the state of the art in Computerized analysis of skin cancer images and provides future directions for this exciting subfield of medical image analysis [1]. Different techniques for segmentation, feature extraction and classification have been reported by several authors. Numerous features have been extracted from skin images, including shape, colour, texture and border properties. Classification methods range from discriminant analysis to neural networks and support vector machines. See Maglogiannis et al. [2] for a review of the state of the art of computer vision system for skin lesion characterisation. These methods are mainly developed for images acquired by epiluminescence microscopy (ELM or dermoscopy) and they focus on differentiating melanocytic nevi (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. Most systems ignore other benign lesions and crucially the two most common type of skin cancer (Squamous Cell Carcinomas and Basal Cell Carcinomas).



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

Fig. 1. Examples of skin lesion images used in this work

Our key contribution is to 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). Some images are shown in Fig. 1. Moreover, we use only high resolution colour images acquired using standard camera (non-dermoscopy).

A large number of classifier combinations have been proposed in the literature [3]. 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 and therefore improving the recognition accuracy [3]. 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. 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 reduction in the feature space avoids many problems related to high dimensional feature spaces, such as the "curse of dimensionality" problem [3]. 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 classification made at the higher levels of the hierarchy). So each classifier uses a different set of features optimised for those images. This forces the individual classifiers to contain potentially independent information. Hierarchical classifiers are well known [4] and commonly used for document and text classification, including a hierarchical K-NN classifier [5]. While we found papers describing applications of hierarchical systems to medical image classification tasks, to the best of our knowledge only a hierarchical neural network model has been applied to skin lesions [6]. They claim over 90% accuracy on 58 images including 4 melanomas. On the other hand, only poor performance was reported relative to the classification of melanoma using the K-NN method [7, 8].

2. FEATURE EXTRACTION

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

Colour features are represented by the mean colours $\mu = (\mu_R, \mu_G, \mu_B)$ of the lesion and their covariance matrices Σ. Let $\mu_X = \frac{1}{N}\sum_{i=1}^{N} X_i$ and $C_{XY} = \frac{1}{N}\left[\sum_{i=1}^{N} X_i Y_i\right] - \mu_X \mu_Y$ where: N is the number of pixels in the lesion, X_i is the colour component of channel X ($X, Y \in \{R, G, B\}$) of pixel i. In the RGB (Red, Green, Blue) colour space, the covariance matrix is: $\Sigma = \begin{bmatrix} C_{RR} & C_{RG} & C_{RB} \\ C_{GR} & C_{GG} & C_{GB} \\ C_{BR} & C_{BG} & C_{BB} \end{bmatrix}$. In this work, RGB, HSV (Hue, Saturation, Value) and CIE_Lab, CIE_Lch and Otha colour spaces were considered. Four normalisation techniques were investigated to reduce the impact of lighting, which were applied before extracting colour features. In the end, we normalised each colour component by dividing each colour 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 normalisation techniques. After experimenting with the 5 different colour spaces, we choose the normalised RGB, because it gave slightly better results than the other colour spaces.

Texture features are extracted from generalised co-occurrence matrices (GCM), that are the extension of the co-occurrence matrix [9] to multispectral images. Assume an image I having N_x columns, N_y rows and N_g grey 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 quantised grey levels. Let u and v be two colour channels. The generalised co-occurrence matrices are: $P_\delta^{(u,v)}(i,j) = \#\{((k,l),(m,n)) \in (L_y \times L_x) \times (L_y \times L_x) | I_u(k,l) = i, I_v(m,n) = j\}$ i.e. the number of co-occurrences of the pair of grey 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. $(m = k + d, n = l), (m = k + d/\sqrt{2}, n = l + d/\sqrt{2}), (m = k, n = l + d), (m = k - d/\sqrt{2}, n = l + d/\sqrt{2})$. In order to have orientation invariance for our set of GCMs, we averaged the matrices with respect to θ. Quantisation levels $N_G = 64, 128, 256$ are used for the three colour 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 [9], for a total of 3888 texture features (12 features $\times$ 6 inter-pixel distances $\times$ 6 colour pairs $\times$ 3 colour 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. This gives a total of $4 \times 3888 = 15552$ possible texture features, from which we extracted a good subset. All features are normalised to zero mean and unit variance.

The colour and texture features are combined to construct a distance measure between each test image T and a database image I. For colour covariance-based features, the Bhattacharyya distance metric: $BD_{CF}(T,I) = \frac{1}{8}(\mu_T - \mu_I)^T \left[\frac{(\Sigma_T + \Sigma_I)}{2}\right]^{-1} (\mu_T - \mu_I) + \frac{1}{2} \ln \frac{\left|\frac{(\Sigma_T + \Sigma_I)}{2}\right|}{\sqrt{|\Sigma_T||\Sigma_I|}}$ is used, where μ_T and μ_I are the average (over all pixels in the lesion) colour 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 $ED_{TF}(T,I) = \|f^T_{subset} - f^I_{subset}\| = \sqrt{\sum_{i=1}^{S}(f_i^T - f_i^I)^2}$ is used for distances between a subset of S texture features f_{subset}, selected as described later. Other metric distances have been considered, but gave worse results. We aggregated the two distances into a distance matching function as $Dist(T,I) = w \cdot BD_{CF}(T,I) + (1-w) \cdot ED_{TF}(T,I)$ 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.

3. HIERARCHICAL SYSTEM

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 lesions classes: Melanocytic Nevus/ Mole (ML) and Seborrhoeic Keratosis (SK). We note that AK, BCC, SCC, ML and SK are diagnostic classes defined by dermatologists. The two groups were constructed by clustering classes containing images which were

visually similar at this 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 between 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.

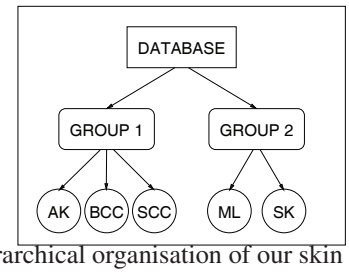


Fig. 2. Hierarchical organisation of our skin lesion classes

We perform feature selection for the 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. A sequential forward selection algorithm (SFS) is used for feature selection. The goal for choosing features is the maximisation 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 $\frac{1}{M}\sum_{j=1}^{M}\frac{\text{correctly_classified}(C_j)}{\text{number_of_test_images}(C_j)}$ 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. The features that maximise 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. 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.

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. An attempt to solve this problem could be to use classifiers on the second level which classify in more than the two or three classes for which they are optimised. 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. RESULTS

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 ($\sim$50 cm) resulting in a pixel resolution of about 0.03 mm. Lesions are segmented using the method described in [10]. Specular highlights have been removed [11]. The features described in previous sections have very different value ranges. To account for this, an objective rescaling of the features is achieved by normalising to z-scores of each feature set. 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 selected as constant over all experiments.

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. Classification results when varying the value of K of the K-NN classifiers have been evaluated [11]. Experiments also showed that is reasonable to use 10 features [11].

The overall classification accuracy on the test set is $74.3 \pm 2.5\%$, as shown in the right column of Table 1. The

Table 1. Average percentage 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

Table 2. Comparison of the overall accuracy of the hierarchical and flat classifiers over the three training 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

overall result also includes the $\sim$6% misclassified samples from the first level. The accuracy of the top classifier and the two subclassifiers at the bottom levels are reported in Table 1. The values are the mean $\pm$ standard deviation over the three training and test sets. Recall the top level classifier discriminates between cancer and pre-malignant 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.

In Table 2 we compare our results of our Hierarchical K-NN classifier (HKNN) 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 the same set of 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 set of 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 3 shows the confusion matrix of the whole HKNN system on the test images. This matrix has been obtained from adding the three confusion matrices on 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 3. 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	6	7	269	7	42
SCC	8	34	5	36	5
SK	9	8	33	8	199

5. CONCLUSIONS

We have presented an algorithm (based on a hierarchical K-NN classifier tree) for the classification of 5 common classes of non-melanoma skin lesion. The performance ($\sim$74%) is not yet at the 90% level (achieved after 20+ years of research) for differential diagnosis of moles versus melanoma. However, our work makes 2 key contributions:

1) These results are based only on normal colour images, unlike the dermoscopy method, which requires a specialised sensor.

2) Our multi-class classification method covers the majority of skin lesion types. This differentiates us from most other approaches that concentrate on only two or three class instances of the problem.

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

6. REFERENCES

[1] M. E. Celebi, W. V. Stoecker, and R. H. Moss, "Advances in skin cancer image analysis," *Computerized Medical Imaging and Graphics*, vol. 35, no. 2, pp. 83 – 84, 2011.

[2] I. Maglogiannis and C. N. Doukas, "Overview of advanced computer vision systems for skin lesions characterization," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 5, pp. 721–733, 2009.

[3] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. on PAMI*, vol. 22, no. 1, pp. 4–37, 2000.

[4] A. D. Gordon, "A review of hierarchical classification," *Journal of the Royal Statistical Society. Series A (General)*, vol. 150, no. 2, pp. 119–137, 1987.

[5] R. Duwairi and R. Al-Zubaidi, "A hierarchical K-NN classifier for textual data," *The International Arab Journal of Information Technology*, vol. 8, no. 3, pp. 251–259, July 2011.

[6] B. Salah, M. Alshraideh, R. Beidas, and F. Hayajneh, "Skin cancer recognition by using a neuro-fuzzy system," *Cancer Informatics*, vol. 10, pp. 1–11, 2011.

[7] Y.A. Aslandogan and G.A. Mahajani, "Evidence combination in medical data mining," in *ITCC 2004*, April 2004, vol. 2, pp. 465 – 469 Vol.2.

[8] M. Hintz-madsen et al., "Design and evaluation of neural classifiers application to skin lesion classification," in *Proc. NNSP V*, 1995, pp. 484–493.

[9] R. M. Haralick, K. Shanmungam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. on Systems, Man and Cybernetics*, vol. 3, no. 6, pp. 610–621, 1973.

[10] L. Xiang, B. Aldridge, L. Ballerini, R. Fisher, and J. Rees, "Depth data improves skin lesion segmentation," in *Proc. MICCAI 2009*, 2009, pp. 1100–1107.

[11] L. Ballerini, R. B. Fisher, B. Aldridge, and J. Rees, "A color and texture based hierarchical k-nn approach to the classification of non-melanoma skin lesions," in *Color Medical Image Analysis*. Springer, 2012, (in press).