

Comput Med Imaging Graph. Author manuscript; available in PMC 2011 October 13.

Published in final edited form as:

Comput Med Imaging Graph. 2007 September; 31(6): 362–373. doi:10.1016/j.compmedimag. 2007.01.003.

# A methodological approach to the classification of dermoscopy images

M. Emre Celebi<sup>a,\*</sup>, Hassan A. Kingravi<sup>a</sup>, Bakhtiyar Uddin<sup>a</sup>, Hitoshi Iyatomi<sup>d</sup>, Y. Alp Aslandogan<sup>a</sup>, William V. Stoecker<sup>b</sup>, and Randy H. Moss<sup>c</sup>

Hassan A. Kingravi: kingravi@cse.uta.edu; Bakhtiyar Uddin: bxu3530@omega.uta.edu; Y. Alp Aslandogan: alp@cse.uta.edu; William V. Stoecker: wvs@umr.edu; Randy H. Moss: rhm@umr.edu

<sup>a</sup>Department of Computer Science and Engineering, The University of Texas at Arlington, Box 19015, 416 Yates Street, Room 300, Nedderman Hall, Arlington, TX 76019-0015, USA

<sup>b</sup>Stoecker & Associates, 1100 W. 10th Street, Rolla, MO 65401-2937, USA

<sup>c</sup>Department of Electrical and Computer Engineering, 226 Emerson Electric Co. Hall, University of Missouri-Rolla, Rolla, MO 65409-0040, USA

dDepartment of Electrical Informatics, Hosei University, Tokyo, Japan

### **Abstract**

In this paper a methodological approach to the classification of pigmented skin lesions in dermoscopy images is presented. First, automatic border detection is performed to separate the lesion from the background skin. Shape features are then extracted from this border. For the extraction of color and texture related features, the image is divided into various clinically significant regions using the Euclidean distance transform. This feature data is fed into an optimization framework, which ranks the features using various feature selection algorithms and determines the optimal feature subset size according to the area under the ROC curve measure obtained from support vector machine classification. The issue of class imbalance is addressed using various sampling strategies, and the classifier generalization error is estimated using Monte Carlo cross validation. Experiments on a set of 564 images yielded a specificity of 92.34% and a sensitivity of 93.33%.

# Keywords

Skin cancer; Dermoscopy; Melanoma; Classification; Support vector machine; Model selection

## 1. Introduction

Malignant melanoma is one of the most rapidly increasing cancers in the world, with an estimated incidence of 59,580 and an estimated total of 7770 deaths in the United States in 2005 alone [1]. Early diagnosis is particularly important since melanoma can be cured with a simple excision if detected early.

Dermoscopy is a non-invasive skin imaging technique that uses optical magnification and either liquid immersion and low angle-of-incidence lighting or cross-polarized lighting to make the contact area translucent, making subsurface structures more easily visible when compared to conventional clinical images [2]. This reduces screening errors, and provides

 $<sup>\</sup>hbox{@ 2007}$  Elsevier Ltd. All rights reserved.

<sup>\*</sup>Corresponding author. Tel.: +1 682 433 6264., celebi@cse.uta.edu (M.E. Celebi).

greater differentiation between difficult lesions such as pigmented Spitz nevi and small, clinically equivocal lesions [3]. However, it has been demonstrated that dermoscopy may actually lower the diagnostic accuracy in the hands of inexperienced dermatologists [4]. Furthermore, for the diagnosis, dermatologists rely on their clinical experience and visual perception. However, diagnosis made by human vision is somewhat subjective, lacking accuracy and reproducibility.

Computerized dermoscopy image analysis systems do not have the limitation of this subjectivity. These systems allow the use of a computer as a second independent diagnostic method, which can potentially be used for the prescreening of patients performed by non-experienced operators and for aiding clinicians [5]. Although computerized analysis techniques cannot provide a definitive diagnosis, they can be used to improve biopsy decision-making, which some observers feel is the most important use for dermoscopy [6]. For example, clinicians can avoid biopsy for such significant classes of lesions as vascular lesions and dysplastic nevi. Finally, automated analysis can serve as an additional tool to improve follow-up, especially for patients with multiple atypical nevi [5].

Studies related to the automated classification of pigmented skin lesion images have appeared in the literature as early as 1987 [7]. Different methods for border detection, feature extraction, and classification have been applied to various image sets mostly obtained from single sources. Table 1 summarizes some key results from 2001 onwards.

In this paper, a methodological approach to the classification of dermoscopy images is presented. Fig. 1 shows an overview of the proposed approach.

The rest of the paper is organized as follows. Section 2 describes the data set collection. Section 3 explains the border detection. Section 4 discusses the feature extraction. Section 5 describes the feature selection. Section 6 presents the support vector machines. Section 7 describes the classification experiments. Finally, Section 8 gives the conclusions and the future work.

# 2. Data set description

The digital dermoscopy images were collected from two dermoscopy atlases [2,17]. The images in [2] were acquired in three university hospitals (University of Graz, Austria, University of Naples, Italy and University of Florence, Italy), while those in [17] were acquired in the Sydney Melanoma Unit, Sydney, Australia. These were true-color images with a typical resolution of  $768 \times 512$  pixels. Since we had no control over the image acquisition and camera calibration, images that satisfied at least one of the following criteria were omitted from the study: (i) the lesion does not fit entirely within the image frame, (ii) presence of too much hair, and (iii) insufficient contrast between the lesion and the background skin. This selectivity was necessary in order to ensure accurate border detection and reliable feature extraction. Fig. 2 shows sample images that were eliminated using these criteria. A total of 596 images free from the above mentioned problems were included in the initial image set.

# 3. Border detection

The first step in the computerized analysis of skin lesion images is the detection of the lesion borders. The importance of the border detection for the analysis is two-fold. First, the border structure provides vital information for accurate diagnosis. Many clinical features such as asymmetry and border irregularity are calculated from the border. Second, the extraction of other important color or texture related clinical features critically depends on the accuracy of the border detection.

For border detection, an automated method that we developed earlier [18] was used. The method is based on the JSEG algorithm [19] and included three main phases: preprocessing, segmentation, and postprocessing. The preprocessing phase included image smoothing using a color median filter with an  $11 \times 11$  kernel, color reduction using the variance-based quantization method [20], and approximate lesion localization based on the Otsu thresholding method [21]. The segmentation phase included calculation of the J-images (measurements of local homogeneities which indicate potential boundary locations), regiongrowing based on the J-images, and region merging based on color similarity in the CIE  $L^*u^*v^*$  color space. Finally, the postprocessing phase included elimination of the regions that belong to the background skin, removal of the isolated regions, and merging of the remaining regions to obtain the final border detection result.

The border detection method was applied to the initial set of 596 images. Fig. 3 shows examples of successful and unsuccessful border detection results. In 32 of the images the results were deemed to be unsatisfactory. Failure mostly occurred in one of two cases: (i) lesions in which there is a very smooth transition between the border and the background skin (see Fig. 3c) and (ii) lesions with regression (scar-like depigmentation) structures (see Fig. 3d). After the exclusion of these images, 564 images remained in the image set.

Among the 564 cases, 88 were melanoma and 476 were benign. Among the melanoma cases, 18 were melanomas in situ, 47 were thin invasive melanomas, 13 had a thickness between 0.76 and 1.5 mm, 7 had a thickness of more than 1.5 mm, and 3 were metastasized. The distribution of the benign cases was as follows: 309 Clark nevi, 45 Reed/Spitz nevi, 31 seborrheic keratoses, 19 compound nevi, 17 blue nevi, 9 combined nevi, 9 dermal nevi, 9 melanoses, 9 vascular lesions, 7 lentigines, 4 congenital nevi, 3 junctional nevi, 3 dermatofibromas, and 2 hemangiomas. The lesions were biopsied and diagnosed histopathologically in cases where significant risk for melanoma was present; otherwise they were diagnosed by follow-up examination.

# 4. Feature extraction

In this section the features that were used to characterize the skin lesion images are described. A particular problem in the related literature is that a significant number of studies do not report the details of their feature extraction procedure [22]. This is further compounded by the inconsistency in the definitions of some of the features (especially those pertaining to shape) in the computer vision literature. Therefore, in order to enhance the reproducibility of this study, we explain the rationale for each feature and present the algorithmic aspects involved in its computation in as much detail as possible.

### 4.1. Description of the shape features

Shape is an important clinical feature in the diagnosis of pigmented skin lesions. In the following discussion, "object" refers to the binary lesion object obtained as a result of the border detection.

- **a.** Area (*A*): The lesion area can be calculated by counting the number of pixels inside the border. However, this method is not very accurate for objects with rough borders [23]. For this reason, the lesion area was calculated using the method of bit quads [21] which has been shown to be one of the most accurate area estimators in the literature [23].
- **b.** Aspect ratio  $(A_R)$ : Aspect ratio can be defined as the ratio of the length of the major axis  $(L_1)$  to the length of the minor axis  $(L_2)$ . These are given in the first column of Table 2. Here,  $(r_0, c_0)$  denotes the object centroid, and  $m_{pq}$  and  $\mu_{pq}$  denote the (p+q)th order geometric and central moments of the object, respectively.

c. Asymmetry 1 and 2 ( $A_1$  and  $A_2$ ): In order to evaluate the lesion asymmetry, first, the major axis orientation of the object ( $\theta$ ) was calculated (Table 2). Second, the object was rotated  $\theta$  degrees clockwise to align the principal (major and minor) axes with the image (x and y) axes. The object was then hypothetically folded about the x-axis and the area difference ( $A_x$ ) between the overlapping folds was taken as the amount of asymmetry about the x-axis. The same procedure was performed for the y-axis. Two asymmetry measures were calculated from  $A_x$  and  $A_y$  as shown in Table 2.

**d.** Compactness (*C*): Compactness is usually defined as the ratio of the area of the object to the area of a circle with the same perimeter. This measure compares the object with a circle, which is the most compact shape. However, this requires accurate estimation of the object perimeter. Therefore, an alternative version that avoids using the perimeter was calculated as the ratio between the equivalent and maximum diameters.

Other shape features include maximum (lesion) diameter (the maximum distance between two points on the border), eccentricity (a measure of elongation, Table 2), solidity (a measure of border irregularity defined as the ratio between the areas of the object and its convex hull), equivalent diameter (the diameter of a circle that has the same area as the object), and two measures related with the object-oriented bounding box (the smallest rectangle that contains the object and is aligned with the principal axes): rectangularity (the ratio between the areas of the object and object-oriented bounding box) and elongation (ratio between the height and width of the object oriented-bounding box).

Note that features related with the length of the lesion border such as perimeter, circularity, thinness, roundness, form factor, etc. were not considered in this study. This is because these features depend on an accurate estimation of the lesion perimeter. However, the digital perimeter is often considerably different from the actual perimeter for complex shapes such as skin lesions. Furthermore, perimeter estimation depends greatly on the image resolution.

## 4.2. Calculation of the inner and outer peripheral regions

For the calculation of color and texture features, three significant regions in the image were considered: lesion, inner periphery, and outer periphery. The lesion was obtained as a result of the automatic border detection procedure. The inner and outer peripheral regions were determined from the binary border image using a fast Euclidean distance transform algorithm [24]. In order to reduce the effects of peripheral inflammation and errors in automatic border detection, the region inside (outside) the border with an area equal to 10% of the lesion area was omitted and the inner (outer) peripheral region was taken as the adjacent region with an area equal to 20% of the lesion area. This is illustrated in Fig. 4.

# 4.3. Description of the color features

In order to quantify the colors present in a lesion, two statistics (mean and standard deviation) over the channels of six different color spaces and several color asymmetry, histogram distance, and centroidal distance features were calculated. The color spaces considered were RGB, rgb (normalized RGB), HSV, I1/2/3 (Ohta space), I1/2/3 and CIE  $L^*u^*v^*$ . All of these except for the I1/2/3 space are well-known in the literature [21]. The I1/2/3 space is a relatively recent color space model described in [25]. The nonlinear transformation from RGB to I1/2/3 is given by:

$$l1=(R-G)^2/((R-G)^2+(R-B)^2+(G-B)^2)$$

$$l2=(R-B)^2/((R-G)^2+(R-B)^2+(G-B)^2)$$

$$l3=(G-B)^2/((R-G)^2+(R-B)^2+(G-B)^2)$$

Table 3 shows the comparison of these color spaces according to the following criteria: decoupling of chrominance and luminance, invariance to illumination intensity, and perceptual uniformity. The first two criteria are essential for dealing with images that are acquired in uncontrolled imaging conditions such as the ones that are used in this study. The last criterion is necessary for the extraction of one of the color features (histogram distance). As it can be seen from Table 3, none of the six color spaces satisfies all of the criteria. This is the reason why we have considered several color spaces that complement each other.

Now, we describe each color feature in detail.

- a. Mean and standard deviation: The mean and standard deviation values calculated over a particular channel quantify the average color and the color variegation in that channel, respectively. One hundred eight color features were calculated as follows: (6 color spaces) × (3 channels in each color space) × (2 statistics: mean and standard deviation) × (3 regions {lesion, inner periphery, outer periphery}). The ratios and differences of the 2 statistics over the 3 regions were also calculated: (outer/inner), (outer/lesion), (inner/lesion), (outer—inner), (outer—lesion), and (inner—lesion). The motivation for calculating the ratio and differences is two-fold. First, the color characteristics of the three regions signify valuable diagnostic information. For example, a sharp transition from the inner periphery to the outer periphery (or vice versa) indicates malignancy. So, in addition to features calculated over the three regions, the differences and ratios might provide additional information about the transition between these regions. The total number of color features in this category was 324.
- **b.** Color asymmetry: This is a measure of the asymmetry in pigment distribution in a particular color channel. It was calculated similarly to the shape asymmetry (which is a measure of the geometric asymmetry) with the exception that pixel values were incorporated in the calculations of the first order geometric moments and the second order central moments as weighting factors. Also, after the hypothetical folding, the absolute brightness difference between the corresponding pixels in the two folds was accumulated as opposed to counting the pixels in one fold that do not have a counterpart in the other fold. The color asymmetry in the *R*, *G*, and *B* channels were calculated using the two asymmetry measures shown in Table 2. The total number of color features in this category was 6.
- c. Centroidal distances: The centroidal distance for a channel is defined as the distance between the geometric centroid (of the binary object) and the brightness centroid of that channel. The brightness centroid was calculated similarly to the geometric centroid except that the moment calculations were weighted by the pixel values. If the pigmentation in a particular channel is homogeneous, the brightness centroid will be close to the geometric centroid and thus the centroidal distance for that channel will be small. In order to achieve invariance to scaling, the distance values were divided by the lesion diameter. The centroidal distance values were calculated for all 3 channels of the 6 color spaces. The total number of color features in this category was 18.
- **d.** LUV histogram distances: In order to determine the color similarity of two regions, the histogram distance in the  $CIE\ L^*u^*v^*$  color space was used. For histogram

computation, the color space was coarsely quantized into  $4 \times 8 \times 8$  bins. The color similarity between the two regions was quantified by the  $L_1$ - and  $L_2$ -norm histogram distances [21]. The use of these norms is justified because the color space is coarsely quantized and there is negligible correlation between adjacent histogram bins.

The histogram distances between pairs of the three regions, i.e. lesion, inner periphery, and outer periphery, using the two distance measures were calculated. The total number of color features in this category was 6.

## 4.4. Description of the texture features

In order to quantify the texture present in a lesion, a set of statistical texture descriptors based on the gray level co-occurrence matrix (GLCM) were employed. GLCM-based texture description is one of the most well-known and widely used methods in the literature [26]. Although many statistics can be derived from the GLCM, eight gray level shift invariant statistics were used in this study in order to obtain a texture characterization that is robust to linear shifts in the illumination intensity. These statistics were *maximum* probability, energy, entropy, dissimilarity, contrast, inverse difference, inverse difference moment, and correlation.

To obtain statistical confidence in the estimation of the joint probability distribution, the normalized GLCM should be reasonably dense. For example, at full dynamic range (G = 256 gray levels for 8-bit images), since very few gray level pairs are repeated, the *entropy* statistic attains similar values for different texture patterns. In a methodological study, Clausi [26] showed that above a certain threshold, for the same eight statistics, with an increasing G, the discrimination power of the statistics remain the same for two of them (*dissimilarity* and *contrast*) while decreasing for the rest. Another problem with high G values is the high computational cost in both the calculation of the GLCM (O(G)) and the statistics (O(G)). Therefore, in order to avoid having a sparse matrix, the images were uniformly quantized to 64 gray levels. Here, the choice of 64 gray levels was arbitrary, though a recent study [26] has demonstrated that this value should not be too low (for example below 24). Another advantage of using a low G value is the reduction of the effects of noise in the image.

In order to obtain rotation invariant features, the normalized GLCM was computed for each of the four orientations ( $\{0^{\circ}, 45^{\circ}, 90^{\circ}, 135^{\circ}\}$ ) and the statistics calculated from these matrices were averaged. These eight statistics were calculated over the three regions, i.e., lesion, inner periphery, and outer periphery, and as in the case of color features, the ratios and differences of the eight statistics over these regions were also calculated. The total number of texture features extracted from each image was 72.

Overall, the number of features extracted from each image was 437 (11 shape, 354 color, and 72 texture features).

### 5. Feature selection

Feature selection is an important preprocessing step in many machine-learning tasks. The purpose is to reduce the dimensionality of the feature space by eliminating redundant, irrelevant or noisy features. From the classification perspective, there are numerous potential benefits associated with feature selection: (i) reduced feature extraction time and storage requirements, (ii) reduced classifier complexity, (iii) increased prediction accuracy, (iv) reduced training and testing times, and (v) enhanced data understanding and visualization.

Feature selection algorithms can be classified into two main categories according to their evaluation criteria: filters and wrappers [27]. Filter methods rely on general characteristics of the data to select a subset of features without involving any learning algorithm. They 'filter' out irrelevant and redundant features before classifier induction begins. On the other hand, wrapper methods use the prediction performance of a predetermined learning algorithm to evaluate the goodness of feature subsets. Although, wrappers are often computationally more expensive, they are better suited to classification tasks in which the classifier is predetermined.

In this study, the filter methodology is adopted for several reasons. First of all, filter methods are usually very fast which allows one to compare several alternative methods within an optimization framework. Secondly, if a wrapper method is to be used on a particular data set, the target-learning algorithm should have at least satisfactory performance on the original data set so that it can provide valuable feedback to the search procedure. However, as demonstrated in the following section, our target learning algorithm, i.e. the SVM algorithm, does not perform well on the original data due to the presence of many irrelevant or redundant features and the unbalanced distribution of the classes. Finally, the only wrapper implementation available to us was the well-known recursive feature elimination (RFE) [28] algorithm which uses a linear SVM classifier to rank the features. However, in order to use the original RFE algorithm as a wrapper, one needs to use a linear SVM classifier for classification. Otherwise, if one uses another learning algorithm or even an SVM kernel other than the linear one, the RFE algorithm turns into a computationally expensive filter method! As will be explained in the next section, we decided to use a radial basis function (RBF) kernel rather than a linear one. In fact, the RFE algorithm can be modified to use an RBF kernel [28]. However, in that case model selection (see Section 7) would become computationally very expensive.

Among the various filter methods proposed in the literature [27] the following three were chosen for their good performance on various data sets:

- ReliefF [29]: In the original Relief algorithm [30], a number of samples are selected at random from the data set and their nearest neighbors are determined. For each selected sample, the values of its features are compared to those of the nearest neighbors and the relevance scores for each feature are updated accordingly. The idea is to estimate the quality of attributes according to how well their values distinguish between samples that are near to each other. The ReliefF algorithm is an extension of the original algorithm that can handle noise and multi-class data sets.
- Mutual information based feature selection (MIFS) [31]: Mutual information
  measures arbitrary dependencies between random variables, and thus is suitable for
  assessing the information content of the features. The MIFS algorithm evaluates the
  mutual information between individual features and the class labels, and selects
  those features that have the maximum mutual information.
- Correlation based feature selection (CFS) [32]: This algorithm tries to find a set of
  features that individually correlate well with the class but have little
  intercorrelation. The correlation between two features is measured by symmetric
  uncertainty [33] which is an improved form of the well-known information gain
  measure [34].

For the *ReliefF* and CFS algorithms the Weka implementations [35] were used. For the MIFS algorithm, the Tanagra [36] implementation was used. Before feature selection, the features were discretized using the technique of Fayyad and Irani [37].

# 6. Support vector machines

Support vector machines (SVMs) have recently drawn considerable attention in the machine learning community due to their solid theoretical foundation and excellent practical performance. They are kernel-based learning algorithms derived from the statistical learning theory [38].

SVMs have several advantages over the more classical classifiers such as decision trees and neural networks. The support vector training mainly involves optimization of a convex cost function. Therefore, there is no risk of getting stuck at local minima as in the case of backpropagation neural networks. Most learning algorithms implement the empirical risk minimization (ERM) principle which minimizes the error on the training data. On the other hand, SVMs are based on the structural risk minimization (SRM) principle which minimizes the upper bound on the generalization error. Therefore, SVMs are less prone to *overfitting* when compared to the algorithms that implement the ERM principle such as backpropagation neural networks. Another advantage of SVMs is that they provide a unified framework in which different learning machine architectures (e.g., RBF networks, feedforward neural networks) can be generated through an appropriate choice of kernel.

## 6.1. General theoretical background

This subsection gives an overview of the SVM theory and is based on [39]. Consider a set of n training data points  $\{x_i \ y_i\} \in \mathbb{R}^d \times \{-1, +1\}i = 1, ..., n$ , where  $x_i$  represents a point in d-dimensional space and  $y_i$  is a two-class label. Suppose we have a hyperplane that separates the positive samples from the negative ones. Then the points x on the hyperplane satisfy  $w \cdot x + b = 0$ , where w is the normal to the hyperplane, and |b|/|w|| is the perpendicular distance from the hyperplane to the origin. If we take two such hyperplane between the positive and negative samples, the support vector algorithm's task is to maximize the distance (margin) between them. In order to maximize the margin,  $||w||^2$  is minimized subject to the following constraints:

$$y_i(x_i \cdot \boldsymbol{w} + b) - 1 \ge 0 \forall i \tag{1}$$

The training samples for which (1) hold are the only ones relevant for the classification. These are called the *support vectors*.

The Lagrangian function for the minimization of  $||w||^2$  is given by:

$$L_{1} = \sum_{i=1}^{n} \alpha_{i} - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_{i} y_{j} \alpha_{i} \alpha_{j} x_{i} x_{j} \quad \text{subject to} \quad \alpha_{i} \ge 0 \forall i$$
and
$$\sum_{i=1}^{n} \alpha_{i} y_{i} = 0$$
(2)

Equation (2) applies only to linearly separable data. In order to handle non-linearly separable data, positive slack variables  $\xi_i$ , i = 1, ..., n are introduced into the constraints:

$$y_i(x_i \cdot w + b) \ge 1 - \xi_i, \quad \xi_i \ge 0 \forall i$$
 (3)

In order to control the trade-off between the model complexity and the empirical risk, a penalty parameter C is introduced into (2):

$$L_{nl} = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j x_i x_j \quad \text{subject} \quad \text{to}$$

$$0 \le \alpha_i \le C \quad \forall i \quad \text{and} \quad \sum_i \alpha_i y_i = 0$$

$$(4)$$

To generalize these equations for non-linear decision functions, the concept of a kernel is introduced. The data seen in the equations so far appears in the form of dot products  $x_i \cdot x_j$ . If we were to map the data to some other (possibly infinite dimensional) Euclidean space H, using a mapping  $\Phi$ , the training algorithm would depend on the data through dot products in H, i.e.  $\Phi(x_i) \cdot \Phi(x_j)$ . Now, if there were a "kernel function" K such that  $K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$ , we would only need to use K in the training algorithm, and would never need to explicitly know what  $\Phi$  is. Substituting the kernel K into the dual SVM gives:

$$L_k = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n y_i y_j \alpha_i \alpha_j K(x_i, x_j) \quad \text{subject} \quad \text{to}$$

$$0 \le \alpha_i \le C \quad \text{and} \quad \sum_i \alpha_i y_i = 0$$
(5)

This formulation allows us to deal with extremely high (theoretically infinite) dimensional mappings without having to do the associated computation.

Some commonly used kernels are:

- Linear:  $K(x_i, x_j) = x_i^T \cdot x_j$
- Polynomial:  $K(\mathbf{x}_i, \mathbf{x}_j) = (\gamma x_i^T \cdot x_j + r)^d, \ \gamma > 0$
- Radial basis function (RBF):  $K(x_i, x_j) = e^{-\gamma ||x_i x_j||^2/2\sigma^2}, \gamma > 0$
- Sigmoid:  $K(x_i, x_j) = \tan h (\gamma x_i^T \cdot x_j + r), \gamma > 0$

In this study, the radial basis function (RBF) kernel was adopted for various reasons. Firstly, the linear kernel cannot handle nonlinearly separable classification tasks, and in any case, is a special case of the RBF kernel [40]. Secondly, the computation of the *RBF* kernel is more stable than that of the polynomial kernel, which introduces values of zero or infinity in certain cases. Thirdly, the sigmoid kernel is only valid (i.e. satisfies Mercer's conditions [39]) for certain parameters. Finally, the RBF kernel has fewer hyperparameters ( $\gamma$ ) which need to be determined when compared to the polynomial ( $\gamma$ , r, d) and sigmoid kernels ( $\gamma$ , r).

#### 6.2. Feature normalization

In classification tasks the features that characterize the samples quite often have different ranges. Many classifiers such as *k*-nearest neighbors and neural networks require that the features be normalized so that their values fall within a specified range. In the case of SVMs, feature normalization is an important preprocessing step that is necessary to prevent features with large ranges from dominating the calculations and also to avoid numerical instabilities in the kernel computations.

One of the most common normalization methods is the z-score transformation [41] given by:

$$z_{ij} = \frac{((x_{ij} - \mu_j)/(3\sigma_j) + 1)}{2}$$

where  $x_{ij}$  represents the value of the *j*th feature of the *i*th sample;  $\mu_j$  and  $\sigma_j$  are the mean and standard deviation of the *j*th feature, respectively.

Assuming that each feature is normally distributed, this transformation guarantees 99% of  $z_{ij}$  be in the [0,1] range. The out-of-range values are truncated to either 0 or 1. The normality of each feature distribution was verified using the moments of skewness and kurtosis (5% significance level). For more information about these tests, the reader is referred to [42].

# 7. Classification experiments

In this section the classification experiments are described. First, the initial experiments with the SVM classifier are presented. Second, the class imbalance problem and the strategies to deal with it are introduced. Third, the model selection procedure and the experimental results are presented.

# 7.1. Initial experiments with the SVM classifier

In order to use the RBF kernel, appropriate values for the kernel parameters C (cost/penalty) and  $\gamma$  (kernel width) need to be determined. The purpose of model selection is to identify the optimal values for these parameters that give the maximum prediction accuracy on future asyet-unseen data. Since there are only two parameters, a grid-search is feasible. Following [43], exponentially growing sequences of values, i.e.,  $C \in \{2^{-5}, 2^{-3}, ..., 2^{15}\}$  and  $\gamma \in \{2^{-15}, 2^{-13}, ..., 2^{3}\}$ , for these parameters were tried. During the grid-search procedure, ten-fold stratified cross-validation was performed to evaluate the goodness of a particular combination of parameter values, i.e.,  $(C_0, \gamma_0)$ . After the grid-search, the *SVM* classifier was trained with the optimal parameters  $(C^*, \gamma^*)$ . Ten times 10-fold stratified Monte Carlo cross-validation was used to estimate the classification error.

Initially, the procedure described above was performed on the full data set (564 samples, 437 features). The optimal parameter values  $(C^*, \gamma^*) = (2.0, 0.125)$  yielded 24.7% sensitivity and 97.5% specificity. This unsatisfactory result was expected considering the high number of features (many of which are possibly redundant or irrelevant) and particularly the unbalanced distribution of classes (15.6% melanoma, 84.4% benign). In the next subsection, the class imbalance problem and the strategies to deal with it are described.

# 7.2. Dealing with class imbalance

There has been recent interest in the problem of class imbalance in the machine learning community. This problem typically occurs when one or more classes outnumber the others. In such cases, most classifiers focus on learning the large classes which leads to poor classification accuracy for the small classes. In practice, this may not be catastrophic in domains in which the classes have similar misclassification costs. However, in many domains such as medical diagnosis and fraud detection the misclassification costs are often unequal and classifying the minority (melanoma) samples as majority (benign) implies serious consequences.

The most common classifier performance measure is the *accuracy* defined as the percentage of correctly classified samples. However, accuracy is not an appropriate measure of the classification performance when the data is unbalanced. Consider a data set with a class distribution of 99:1. A classifier that always predicts samples as the majority class will have an accuracy of 99%. This is because the accuracy measure is strongly biased to favor the majority class.

A better performance measure in unbalanced domains is the receiver operating characteristic (ROC) curve. The ROC curve is a plot of the true positive (TP)-rate (percentage of correctly

classified positive samples) versus false positive (FP)-rate (percentage of incorrectly classified negative samples). The points for a ROC curve are obtained by varying a threshold on a classifier's continuous output between its extremes and plotting the (TP-rate, FP-rate) for each threshold value. The curve illustrates the behavior of a classifier without regard to class distributions or error costs, and thus decouples the classification performance from these factors. The area under the ROC curve (AUC) represents the expected predictive performance as a single scalar value [44]. AUC exhibits several desirable properties compared to accuracy. For example, it is independent of the decision threshold and is invariant to apriori class probability distributions. In a recent study Huang and Ling [44] have demonstrated that AUC is a statistically consistent and more discriminatory measure than accuracy. In this study AUC was used to evaluate the goodness of a particular classifier model.

One of the most common and effective techniques for dealing with imbalance is sampling [45]. The motivation for sampling comes from the observation that the natural distribution of the classes might not be optimal from the classification perspective [46]. Several studies have demonstrated that the accuracy degradation on unbalanced data sets is more severe when the classes overlap significantly [47,48]; this is the case in skin lesion classification. For example, early melanomas (melanoma in situ) are often confused with Clark nevi by dermoscopy practitioners.

There are two basic sampling methods: under-sampling (removing majority class samples) and over-sampling (adding minority class samples). In this study, two common sampling methods were compared:

- **a.** *Random under-sampling*, which eliminates randomly chosen majority class samples.
- **b.** Synthetic minority oversampling technique (SMOTE) [49], which over-samples the minority class by taking each minority class sample and introducing synthetic samples along the line segments joining n of the k minority class nearest neighbors. In this study k = 10 was used. The value of n depends on the amount of oversampling. For example, if the amount of over-sampling needed is 200%, only two neighbors from the 10 nearest neighbors are chosen and one synthetic sample is generated in the direction of each.

We decided to add (remove) minority (majority) samples until an approximately balanced class distribution was reached. This was motivated by the results presented in [46], in which the authors demonstrate that when AUC is used as the performance measure, the optimal class distribution for learning tends to be near the balanced distribution. For *random undersampling*, this amounted to randomly removing  $476 \times 80\% = 380$  majority samples resulting in a 96:88 benign to melanoma class ratio. For SMOTE, four synthetic melanoma samples were created from each melanoma sample, resulting in a 476:440 benign to melanoma ratio.

## 7.3. Model selection and experimental results

As demonstrated earlier, the number of features retained by the feature selection algorithm (k) is an important parameter that needs to be considered in order to obtain a good classification performance. Considering the complexity of the problem in our case, a small number of features is not likely to discriminate between the classes well. On the other hand, a large number of features might lead to overfitting as explained in Section 5. With these considerations, the range of k was restricted to [5, 30].

The optimization procedure after the integration of sampling and feature selection was as follows (Fig. 5):

- **a.** Perform feature selection using {CFS, MIFS, ReliefF}
- **b.** Reduce the data dimensionality by keeping only the top k ( $k \in [5,30]$ ) features in the ranking returned by the feature selection algorithm.
- **c.** Perform sampling using { random under-sampling, SMOTE}
- **d.** Perform grid-search.
- **e.** Perform SVM classification (10 times 10-fold Monte Carlo stratified cross validation) using the optimal parameter values returned by the grid-search. Calculate the sensitivity, specificity, and AUC values.

The procedure presented was run on three different processors (one 1.8 GHz Intel Xeon and two 1.5 GHz IBM P5-570) in parallel. The total time taken by the experiments was approximately 12 h.

The optimization results for the two sampling algorithms are shown in Fig. 6a. Each curve is a plot of AUC versus feature subset size (*k*). Given a *k* value, only the highest *AUC* value achieved by any one of the three feature selection algorithms is shown in the plot. It can be seen that *SMOTE* performed better than *random under-sampling* for most *k* values. This was expected since *random under-sampling* discards potentially useful data by removing 80% of the majority (benign) class to make the class distributions approximately balanced. For example, during this removal some of the benign subclasses that have few examples such as congenital nevi and dermatofibroma might be completely eliminated. Also, undersampling severely reduces the data set size which makes learning much more difficult.

Examining Fig. 6a closely, it can be seen that the AUC peaks at k = 18 and the inclusion of features beyond this value does not add much to the classifier performance. This shows that most of the features are in fact either redundant or irrelevant. Therefore, we decided to use the top k = 18 features given by the CFS feature selection algorithm. This gave a specificity of 92.34% and a sensitivity of 93.33%. The ROC curve is shown in Fig. 6b.

### 8. Discussion and conclusions

In this study, a methodological approach to the classification of dermoscopy images is presented. The approach involved border detection, feature extraction, and SVM classification with model selection. The system was tested on a large set of images. Promising results were obtained despite the fact that the images came from different sources and there was no control over their acquisition. The total processing time for the classification of a new image ranges from 5 to 10 s. This can be further reduced by using a faster border detection method such as the one presented in [50].

This study differs from earlier studies in several aspects. First, the images used in this study came from different sources. For this reason, care was taken to extract features that are invariant to changes in the imaging conditions. Second, starting from the border detection until the classification, the whole procedure was fully automated. Third, for border detection a published method was used. Fourth, certain diagnostic classes that occur frequently in the clinical setting such as Clark nevi and seborrheic keratoses were not excluded from the image set. In addition, difficult lesions such as melanomas in situ and Spitz nevi were not excluded. Finally, the issue of feature selection was addressed in an optimization framework without using arbitrary cutoffs.

Despite the high accuracy that can be achieved by computer-aided diagnostic systems employing statistics obtained from low-level features such as the one presented, at least two issues need to be addressed before these systems can gain greater clinical acceptance. First,

higher level features based on a particular diagnostic scheme such as pattern analysis [2] should be integrated with the existing low level features. Second, the image set should be expanded to provide better training and testing for the developed algorithms.

The elimination of some of the images because of the problems noted in Section 2 might be considered a limitation of the present study. However, this is in line with the limitations imposed in previous studies. Studies vary in the detail given about which lesions are included in the training and test sets. In some studies, incompletely imaged lesions were omitted as we have done [14,9]. In some others, some of the lesions that were eliminated here were made acceptable by other means, such as shaving hairs prior to image acquisition [11] or by obtaining manual borders for the lesions [16]. An alternative approach to overcome the hair problem could be using a software razor such as Dullrazor [51] that can, with some modifications, eliminate hairs without altering the pigment network. It is likely that further research, by extraction of critical features such as atypical pigment networks, globules, and blue-white areas [52], can increase the diagnostic accuracy of computerized dermoscopy image analysis systems.

# **Acknowledgments**

This work was supported by grants from NIH (SBIR grant 1R43 CA-101639-01), NSF (#0216500-EIA), Texas Work-force Commission (#3204600182), and James A. Schlipmann Melanoma Cancer Foundation. The permission to use the images from the EDRA Interactive Atlas of Dermoscopy is gratefully acknowledged.

### References

- 1. Jemal A, Murray T, Ward E, Samuels A, Tiwari RC, Ghafoor A, et al. Cancer statistics 2005. CA: Cancer J Clin. 2005; 55(1):10–30. [PubMed: 15661684]
- Argenziano, G.; Soyer, HP.; De Giorgi, V.; Piccolo, D.; Carli, P.; Delfino, M., et al. Dermoscopy: A Tutorial. Milan: EDRA Medical Publishing & New Media; 2002.
- Steiner K, Binder M, Schemper M, Wolff K, Pehamberger H. Statistical evaluation of epiluminescence dermoscopy criteria for melanocytic pigmented lesions. J Am Acad Dermatol. 1993; 29(4):581–8. [PubMed: 8408794]
- Binder M, Schwarz M, Winkler A, Steiner A, Kaider A, Wolff K, et al. Epiluminescence microscopy. a useful tool for the diagnosis of pigmented skin lesions for formally trained dermatologists. Arch Dermatol. 1995; 131(3):286–91. [PubMed: 7887657]
- Hoffmann K, Gambichler T, Rick A, Kreutz M, Anschuetz M, Grunen-dick T, et al. Diagnostic and neural analysis of skin cancer (DANAOS). A multicentre study for collection and computer-aided analysis of data from pigmented skin lesions using digital dermoscopy. Br J Dermatol. 2003; 149(4):801–9. [PubMed: 14616373]
- Bystryn JC. Epiluminescence microscopy: A re-evaluation of its purpose. Arch Dermatol. 2001; 137(3):377–8. [PubMed: 11255355]
- 7. Cascinelli N, Ferrario M, Tonelli T, Leo E. A possible new tool for clinical diagnosis of melanoma: The computer. J Am Acad Dermatol. 1987; 16(2):361–7. [PubMed: 3819073]
- 8. Elbaum M, Kopf AW, Rabinovitz HS, Langley RG, Kamino H, Mihm MC Jr, et al. Automatic differentiation of melanoma from melanocytic nevi with multispectral digital dermoscopy: A feasibility study. J Am Acad Dermatol. 2001; 44(2):207–18. [PubMed: 11174377]
- 9. Ganster H, Pinz A, Rohrer R, Wildling E, Binder M, Kittler H. Automated melanoma recognition. IEEE Trans Med Imaging. 2001; 20(3):233–9. [PubMed: 11341712]
- 10. Hintz-Madsen, M.; Hansen, LK.; Larsen, J.; Drzewiecki, KT. A probabilistic neural network framework for detection of malignant melanoma. In: Naguib, RNG.; Sherbet, GV., editors. Artificial neural networks in cancer diagnosis, prognosis, and patient management. Boca Raton: CRC Press; 2001. p. 141-83.

11. Rubegni P, Burroni M, Cevenini G, Perotti R, Dell'Eva G, Barbini P, et al. Digital dermoscopy. Analysis and artificial neural network for the differentiation of clinically atypical pigmented skin lesions: A retrospective study. J Invest Dermatol. 2002; 119(2):471–4. [PubMed: 12190872]

- 12. Kahofer P, Hofmann-Wellenhof R, Smolle J. Tissue counter analysis of dermatoscopic images of melanocytic skin tumors: Preliminary findings. Melanoma Res. 2002; 12(1):71–5. [PubMed: 11828260]
- Sboner A, Eccher C, Blanzieri E, Bauer P, Cristofolini M, Zumiani G, et al. A multiple classifier system for early melanoma diagnosis. Artif Intell Med. 2003; 27(1):29–44. [PubMed: 12473390]
- 14. Iyatomi H, Oka H, Saito M, et al. Quantitative assessment of tumour extraction from dermoscopy images and evaluation of computer-based extraction methods for an automatic melanoma diagnostic system. Melanoma Res. 2006; 16(2):183–90. [PubMed: 16567974]
- 15. Blum A, Luedtke H, Ellwanger U, Schwabe R, Rassner G, Garbe C. Digital image analysis for diagnosis of cutaneous melanoma. development of a highly effective computer algorithm based on analysis of 837 melanocytic lesions. Br J Dermatol. 2004; 151(5):1029–38. [PubMed: 15541081]
- Menzies SW, Bischof L, Talbot H, Gutenev A, Avramidis M, Wong L. The performance of SolarScan: An automated dermoscopy image analysis instrument for the diagnosis of primary melanoma. Arch Dermatol. 2005; 141(11):1388–96. [PubMed: 16301386]
- 17. Menzies, SW.; Crotty, KA.; Ingwar, C.; McCarthy, WH. An atlas of surface microscopy of pigmented skin lesions: dermoscopy. 2. Sydney: McGraw-Hill; 2003.
- 18. Celebi ME, Aslandogan YA, Stoecker WV, Iyatomi H, Oka H, Chen X. Unsupervised border detection in dermoscopy images. Skin Res Technol. 2007; 13:1–9.
- 19. Deng Y, Manjunath BS. Unsupervised segmentation of color-texture regions in images and video. IEEE Trans Pattern Anal Machine Intell. 2001; 23(8):800–10.
- 20. Wan SJ, Prusinkewicz P, Wong SKM. Variance-based color image quantization for frame buffer display. Color Res Appl. 1990; 15(1):52–8.
- 21. Pratt, WK. Digital image processing: PIKS inside. 3. New York: Wiley; 2001.
- 22. Day GR, Barbour RH. Automated melanoma diagnosis: where are we at? Skin Res Technol. 2000; 6(1):1–5. [PubMed: 11428935]
- 23. Yang, L.; Albregtsen, F.; Lonnestad, T.; Grottum, P. Methods to estimate areas and perimeters of blob-like objects: a comparison. Proceedings of the IAPR workshop on machine vision applications; 1994. p. 272-6.
- 24. Felzenszwalb, PF.; Huttenlocher, DP. Distance Transforms of Sampled Functions. Cornell Computing and Information Science TR2004-1963;2004. Available at: http://people.cs.uchicago.edu/~pff/papers/dt.pdf
- 25. Gevers T, Smeulders AWM. Color-based object recognition. Pattern Recog. 1999; 32(3):453-64.
- 26. Clausi DA. An analysis of co-occurrence texture statistics as a function of gray level quantization. Canad J Remote Sens. 2002; 28(1):45–62.
- Liu H, Yu L. Toward integrating feature selection algorithms for classification and clustering. IEEE Trans Knowledge Data Eng. 2005; 17(4):491–502.
- 28. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine Learn. 2002; 46(1–3):389–422.
- 29. Kononenko, I.; Šimec, E. Induction of decision trees using RELIEFF. In: Riccia, GD.; Kruse, R.; Viertl, R., editors. CISM courses and lectures. Heidelberg: Springer Verlag; 1994.
- 30. Kira, K.; Rendell, IA. The feature selection problem: traditional methods and a new algorithm. In: Swartout, WR., editor. Proc 10th Nat Conf Artifi Intell. AAAI Press; 1992. p. 129-34.
- 31. Battiti R. Using mutual information for selecting features in supervised neural net learning. IEEE Trans Neural Netw. 1994; 5(4):537–50. [PubMed: 18267827]
- 32. Hall, MA. Correlation-based feature selection for discrete and numeric class machine learning. In: Langley, P., editor. Proc 17th Int Conf Machine Learn. San Francisco: Morgan Kaufmann; 2000. p. 359-66.
- 33. Press, WH.; Flannery, BP.; Teukolsky, SA.; Vetterling, WT. Numerical recipes in C: The art of scientific computing. 2. New York: Cambridge University Press; 1992.
- 34. Quinlan, JR. C4.5: Programs for machine learning. San Francisco: Morgan Kaufmann; 1993.

35. Witten, IH.; Frank, E. Data mining: Practical machine learning tools and techniques. 2. San Francisco: Morgan Kaufmann; 2005.

- 36. Rakotomalala R. TANAGRA: A Free software for research and academic purposes. Proc EGC'2005 Conf. 2005; 2:697–702.
- 37. Fayyad, U.; Irani, K. Multi-interval discretization of continuous attributes as pre-processing for classification learning. In: Bajcsy, R., editor. Proceedings of the 13th international joint conference on artificial intelligence. Vol. 2. San Francisco: Morgan Kaufmann; 1993. p. 1022-7.
- 38. Vapnik, V. Statistical learning theory. New York: Wiley; 1998.
- 39. Burges CJC. A tutorial on support vector machines for pattern recognition. Data Mining Knowledge Discov. 1998; 2(2):121–67.
- 40. Keerthi SS, Lin C-J. Asymptotic behaviors of support vector machines with Gaussian kernel. Neural Comput. 2003; 15(7):1667–89. [PubMed: 12816571]
- 41. Aksoy S, Haralick RM. Feature normalization and likelihood-based similarity measures for image retrieval. Pattern Recog Lett. 2001; 22(5):563–82.
- 42. Thode, HC. Testing for normality. New York: CRC Press; 2002.
- 43. Hsu, C-W.; Chang, C-C.; Lin, C-J. A practical guide to support vector classification. Available at http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf
- 44. Huang J, Ling CX. Using AUC and accuracy in evaluating learning algorithms. IEEE Trans Knowledge Data Eng. 2005; 17(3):299–310.
- 45. Weiss GM. Mining with rarity: a unifying framework. SIGKDD Explor. 2004; 6(1):7–19.
- 46. Weiss GM, Provost FJ. Learning when training data are costly: the effect of class distribution on tree induction. J Artif Intell Res. 2003; 19:315–54.
- 47. Japkowicz, N. Learning from imbalanced data sets: a comparison of various strategies. In: Japkowicz, N., editor. AAAI Workshop on Learning from Imbalanced Data Sets. AAAI Press; 2000. p. 10-5.
- 48. Prati, RC.; Batista, GEAPA.; Monard, MC. Class imbalances versus class overlapping: an analysis of a learning system behavior. In: Monroy, R.; Arroyo-Figueroa, G.; Sucar, LE.; Sossa, H., editors. Proceedings of the third Mexican international conference on artificial intelligence, lecture notes in computer science. Vol. 2972. 2004. p. 312-21.
- 49. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002; 16:321–57.
- 50. Celebi, ME.; Kingravi, HA.; Stoecker, WV.; Moss, RH.; Aslandogan, YA. Fast and accurate border detection in dermoscopy images using statistical region merging to appear in the Proceedings of the SPIE Medical Imaging 2007 Conference; San Diego, CA. February 2007;
- 51. Lee TK, Ng V, Gallagher R, Coldman A, McLean D. Dullrazor—A software approach to hair removal from images. Comput Biol Med. 1997; 27(6):533–43. [PubMed: 9437554]
- 52. Celebi, ME.; Kingravi, HA.; Aslandogan, YA.; Stoecker, WV. Detection of blue-white veil areas in dermoscopy images using machine learning techniques. Proceedings of the SPIE Medical Imaging 2006 Conference; 2006. p. 1861-8.

# **Biographies**

**Dr. M. Emre Celebi** received his BSc degree in computer engineering from Middle East Technical University (Ankara, Turkey) in 2002. He received his MSc and PhD degrees in computer science and engineering from the University of Texas at Arlington (Arlington, TX, USA) in 2003 and 2006, respectively. He is currently a visiting assistant professor in the Department of Computer Science and Engineering at the University of Bridgeport (Bridgeport, CT, USA). His research interests include medical image analysis, color image processing, content-based image retrieval, and open-source software development.

**Hassan A. Kingravi** received his Bachelor of Science degree in computer science and engineering from the University of Texas at Arlington in 2006. He worked in the INFOLAB at UTA from 2005 to 2006. He is currently pursuing a Masters degree in computer

engineering at Texas A&M University. His research interests include pattern recognition and machine learning.

**Bakhtiyar Uddin** received his BSc in computer science and engineering from the University of Texas, Arlington in 2005. He worked on the DESCARTES project focusing on medical image analysis from June 2005 to June 2006. He is currently pursuing his Masters in University of Texas, Austin. His research interests include machine learning and data mining.

**Dr. Hitoshi Iyatomi** is a research associate in Hosei University, Tokyo, Japan. He received his BE, ME degrees in electrical engineering and PhD degree in science for open and environmental systems from Keio University in 1998, 2000 and 2004, respectively. During 2000–2004, he was employed by Hewlett Packard Japan. His research interests include intelligent image processing and development of practical computer-aided diagnosis systems.

**Dr. Y. Alp Aslandogan** received his BSc degrees in computer science and mathematics from Bogazici University in Istanbul, Turkey. He received his Masters degree in computer science from Case Western Reserve University in Cleve-land, Ohio and his PhD degree in electrical engineering and computer science from the University of Illinois in Chicago. He has served on the technical program committees of IEEE International Conference on Information Technology, International Conference on Information Reuse and Integration and Emerging Technologies Conference. Dr. Aslandogan currently serves on the faculty of the Department of Computer Science at Prairie View A&M University. His research interests include biomedical informatics, databases and data mining, multimedia information retrieval, and information visualization.

William V. Stoecker, MD received the BS degree in mathematics in 1968 from the California Institute of Technology, the MS in systems science in 1971 from the University of California, Los Angeles, and the MD in 1977 from the University of Missouri, Columbia. He is adjunct assistant professor of computer science at the University of Missouri-Rolla and clinical assistant professor of Internal Medicine-Dermatology at the University of Missouri-Columbia. His interests include computer-aided diagnosis and applications of computer vision in dermatology and development of handheld dermatology databases.

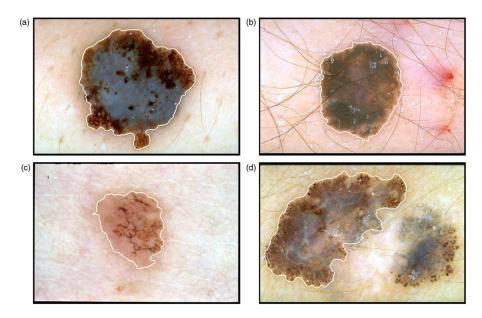
**Dr. Randy H. Moss** received his PhD in electrical engineering from the University of Illinois, and his BS and MS degrees from the University of Arkansas in the same field. He is now professor of electrical and computer engineering at the University of Missouri-Rolla. He is an associate editor of both Pattern Recognition and Computerized Medical Imaging and Graphics. His research interests emphasize medical applications, but also include industrial applications of machine vision and image processing. He is a senior member of IEEE and a member of the Pattern Recognition Society and Sigma Xi. He is a past recipient of the Society of Automotive Engineers Ralph R. Teetor Educational Award and the Society of Manufacturing Engineers Young Manufacturing Engineer Award. He is a past National Science Foundation Graduate Fellow and National Merit Scholar.



**Fig. 1.** Schematic of the system.



**Fig. 2.** Sample images omitted from the study: (a) incompletely imaged lesion (diameter = 35 mm), (b) too much hair, and (c) insufficient contrast between the lesion and the background skin.

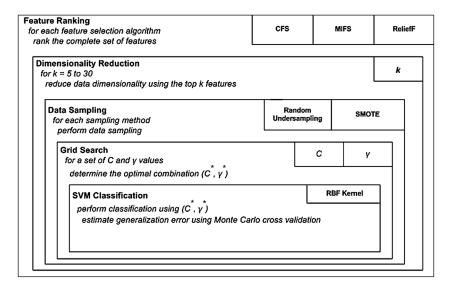


**Fig. 3.** Examples of successful (a, b) and unsuccessful (c, d) border detection results.

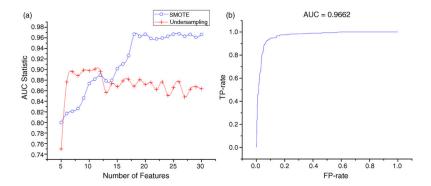


Fig. 4.

Inner and outer peripheral regions of a sample lesion image.



**Fig. 5.** Overview of the optimization framework.



**Fig. 6.**(a) Optimization results for the two sampling methods, (b) ROC curve for the optimal parameter set.

Celebi et al.

Table 1

Summary of recent studies using dermoscopy images

Source	Year	Source Year Segmentation method	Classifier	Total images Mel. (%) Dys. (%) Sens. Spec.	Mel. (%)	Dys. (%)	Sens.	Spec.
[8]	2001	Thresholding	NR	246	26	45	100	84
[6]	2001	Thresholding + color clustering	kNN	5363	2	19	73	68
[10]	2001	Thresholding	ANN	58	38	19	77	75
[11]	2002	Edge detection	ANN	147	39	29	93	92.75
[12]	2002	None	CART	40	50	30	100	91
[13]	2003	NR	Multiple classifiers	152	28	NR	81	74
[14]	2004	Thresholding + region growing	ANN	319	24	59	9.98	90.2
[15]	2004	NR	Logistic regression	837	10	11	80.0	82.4 82.7
[16]	2005	2005 Semi-automatic + manual	Logistic regression	2430	16	25	91	92

NR: not reported, kNN: k nearest-neighbor, ANN: artificial neural network, CART: classification and regression trees, Mel.: melanoma, Dys.: dysplastic, Sens.: sensitivity, Spec.: specificity.

Page 23

Table 2

Formulae for some of the shape features

Aspect ratio Eccentricity Asymmetry

$$m_{pq} = \sum_{i=0}^{\text{rows}} \sum_{j=0}^{\text{cols}} {}_{i}^{p} {}_{j}^{q}$$

 $(r_0, c_0) = (m_{10}/m_{00}, m_{01}/m_{00})$ 

$$\theta = \frac{1}{2} \tan^{-1} \left( \frac{2\mu_{11}}{\mu_{20} - \mu_{02}} \right)$$

$$\mu_{pq} = \sum_{i=0}^{\text{rows}} \sum_{j=0}^{\text{cols}} (i - r_0) P \cdot (j - c_0) q \qquad \varepsilon = \frac{(\mu_{02} - \mu_{20})^2 + 4\mu_{11}}{(\mu_{02} + \mu_{02})^2} \qquad A_1 = \frac{\min(A_X, A_Y)}{A} \times 100 \%$$

$$L_{1,\,2} = (8(\mu_{02} + \mu_{20} \pm ((\mu_{02} - \mu_{20})^2 + 4\mu_{11})^{1/2}))^{1/2}$$

$$A_2 = \frac{A_X + A_y}{A} \times 100 \%$$

 $A_R = L_1/L_2$ 

Celebi et al.

Table 3

Comparison of several color spaces used in the study

Criterion	RGB	rgb	HSV	$RGB \ rgb \ HSV \ II/2/3 \ II/2/3 \ L^*u^*v^*$	11/2/3	$L^*u^*v^*$
Decoupling of chrominance and luminance	No	No	No Yes Yes	Yes	No	Yes
invariance to illumination intensity	No	Yes	H, S No	No	Yes	No
Perceptual uniformity	No	No	No No	No	No	Yes

Page 25