Melanoma Computer-Aided Diagnosis: Reliability and Feasibility Study

Marco Burroni,1 Rosamaria Corona,2 Giordana Dell’Eva,3 Francesco Sera,2 Riccardo Bono,2 Pietro Puddu,2 Roberto Perotti,1 Franco Nobile,3 Lucio Andreassi,1 and Pietro Rubegni1

1Department of Dermatology, University of Siena, Italy; 2Istituto Dermopatico dell’Immacolata, Rome, Italy; and 3Italian Cancer League, Siena, Italy

Abstract

Background: Differential diagnosis of melanoma from melanocytic nevi is often not straightforward. Thus, a growing interest has developed in the last decade in the automated analysis of digitized images obtained by epiluminescence microscopy techniques to assist clinicians in differentiating early melanoma from benign skin lesions.

Purpose: The aim of this study was to evaluate diagnostic accuracy provided by different statistical classifiers on a large set of pigmented skin lesions grabbed by four digital analyzers located in two different dermatological units.

Experimental Design: Images of 391 melanomas and 449 melanocytic nevi were included in the study. A linear classifier was built by using the method of receiver operating characteristic curves to identify a threshold value for a fixed sensitivity of 95%. A K-nearest-neighbor classifier, a non-parametric method of pattern recognition, was constructed using all available image features and trained for a sensitivity of 98% on a large exemplar set of lesions.

Results: On independent test sets of lesions, the linear classifier and the K-nearest-neighbor classifier produced a mean sensitivity of 95% and 98% and a mean specificity of 78% and of 79%, respectively.

Conclusions: In conclusion, our study suggests that computer-aided differentiation of melanoma from benign pigmented lesions obtained with DB-Mips is feasible and, above all, reliable. In fact, the same instrumentation used in different units provided similar diagnostic accuracy. Whether this would improve early diagnosis of melanoma and/or reducing unnecessary surgery needs to be demonstrated by a randomized clinical trial.

Introduction

The most effective management of malignant melanoma is early recognition and surgical excision of thin lesions (1), because tumor thickness is universally recognized as the primary determinant of prognosis. Despite the increasing awareness of melanoma, because of the worldwide increase of incidence reported in the last few decades (2), clinical diagnostic accuracy is still disappointing (3–9). Subsequent attempts to develop noninvasive tools to improve early diagnosis resulted in two approaches: epiluminescence microscopy (ELM) and digital image analysis. ELM, first described in 1987 (10), allows the examination of skin lesions with an incident light magnification system with oil at the skin-lens interface, increases to a great extent the lesion morphological detail, and has been reported to improve the accuracy in diagnosing cutaneous lesions, including melanoma (11). Ascierto et al. (12) recently compared data from patient histories and clinical evaluations with ELM-based morphological patterns to characterize skin lesions and minimize interpretation problems. From these comparisons, they proposed new guidelines for the management of pigmented skin lesions (PSL) to provide standard diagnostic and therapeutic approaches and to enhance the early identification of lesions at risk for malignant transformation. However, dermoscopic techniques require formal training and skill in image interpretation through the so-called pattern analysis (13), are highly dependent on subjective judgement, and are scarcely reproducible (14, 15).

Several scoring systems and algorithms such as the ABCD rule for epiluminescence, the seven-point checklist, and the Menzies method (16, 17) have been proposed to improve the diagnostic performance of less experienced clinicians. This simplification has enabled the development of these diagnostic algorithms with good accuracy and reproducibility. However, they showed problems that have not yet been solved. The most important is that the purpose for which they were designed was not achieved, because the within- and between-observer concordance is very low, even for expert observers (18–25).

Digital image analysis has been found to produce objective, reliable descriptions of melanocytic lesions. Hence, a considerable interest has arisen in recent years in the development of computer-assisted, automated analysis of digitized dermoscopic images since the first study by Schindewolf et al. (26–34).

The aim of this study was to evaluate diagnostic accuracy provided by different statistical classifiers on a large set of pigmented skin lesions (melanomas and nevi) grabbed by four digital analyzers (DB-Mips) located in two different dermatological units.

Materials and Methods

Instrumentation and Image Acquisition

The DB-Mips System consists of a 3CCD PAL video camera with 750 lines of image resolution and 60-decibel signal noise ratio. The camera, operating in the visible spectrum, is connected to a patented hand-held optic system yielding dermo-
scopic images with a magnification power ranging from ×6 to ×40 allowing a horizontal field of view from 40 to 6 mm. The light is provided by a 3200 k source and is homogeneously distributed on the analyzing surface at all magnifications. The three separate components of the Broadcast video signal (768 × 576) are connected to an high-quality frame grabber (set to 768 × 576 resolution and 24 bit/pixel color depth) placed inside a PC. Removable magneto-optical disks (640 MB) are used for image storage.

Image Segmentation and Feature Extraction

The choice of the most useful features to extract from digital images depends on the results of epiluminescence pattern analysis. Although the system saves the microscope magnifications along with the texture analysis, offering an objective evaluation, the different magnifications could confuse clinicians wanting to make subjective comparisons of lesions. The system used a procedure for digital image processing based on the Laplacian filter for segmentation and a zero-crossing algorithm for the border automatic outline (26). It then evaluated 49 parameters for discriminant power. Reproducibility was first tested on digitized images of 100 lesions belonging to 20 subjects (one PSL for each patient recorded five times at 15 min intervals). Absolute differences between single measurements and mean values of a given lesion or parameter never exceeded 5% of the mean value. The parameters, as described previously (19), belonged to four categories: geometries; colors; textures; and islands of color (i.e., color clusters inside the lesion). In brief, the geometric variables were: area; maximum and minimum diameters; radius; variance of contour symmetry; circularity; fractality of borders; and ellipsoidality. Color variables were: mean values of red, green, and blue inside the lesion; mean values of red, green, and blue of healthy skin around the lesion; deciles of red, green, and blue inside the lesion; quartiles of red, green, and blue inside the lesion, mean skin-lesion gradient, variance of border gradient, border homogeneity, and border interruptions. Texture variables were: mean contrast and entropy of lesion; and contrast and entropy fractality. The islands of color variables were: peripheral dark regions; dark area; imbalance of dark region; imbalance green area; red area; dominant green region imbalance; blue-gray area; blue-gray regions; transition area; transition region imbalance; background area; background regions imbalance; red, green, and blue multicomponent; and number of red, green, and blue percentiles inside the lesion.

The system evaluates the above variables and gives the diagnostic probability (in real time during clinical examination) at a rate of 24 checks/s.

Image Databases

The pigmented lesions were selected from the image databases of the Department of Dermatology of the University of Siena, Italy, and the Istituto Dermopatico dell’Immacolata (IDI), a research hospital for skin diseases in Rome, Italy. The selected lesions included all melanomas undergoing ELM examination before surgical excision at the Rome and Siena centers (n = 372) between 1999 and 2003 and a random sample of 449 surgically removed benign melanocytic lesions (with available histological diagnosis), including 85 histologically atypical nevi (architectural disorder and melanocytic atypia). All of the PSL were flat and impalpable. Out of a total of 372 melanomas, 70 (19%) were in situ and 178 (48%) were early melanomas with Breslow thickness ≤0.75 mm.

Linear Discriminant Classifier

Feature Selection. The selection of features for the linear classifier was performed on the entire set of histologically diagnosed images. As a first step in the selection process, the Pearson correlation coefficient was calculated for each possible pair of features. Among groups of highly correlated parameters (r > 0.9) with similar morphological meaning, we selected the one with the best discriminating power. As a second step, we only retained for the final analysis features for which there was a significant (t test) difference between melanomas and non-melanomas and, within these diagnostic classes, no significant difference between centers. This left 10 parameters.

Lesion Classification. For lesion classification, a discriminant analysis approach was used (35). Starting with the two classes of lesions, i.e., melanomas and melanocytic nevi, we calculated a score

\[ z = a_i + \sum_{i=1}^{10} a_i x_i \]

for each lesion, where the weights a were obtained maximizing the distance between the means of melanomas and of nevi in the training set in unidimensional space with standardized variability.

Three classifiers were designed using the training samples of the Siena and Rome centers: (a) training set of IDI, Rome; (b) training set of Siena University; and (c) pool of Rome and Siena training sets. Their accuracies were measured in terms of their performance on the test samples. The discriminant linear function was calculated for each lesion in the test set, and the lesion was assigned to the melanoma group if the value was above the threshold value. This classification was then compared with the histopathological classification (gold standard), and classifier performance was measured as sensitivity and specificity. The method of receiver operating characteristic curves (36) was used to identify the threshold value for a fixed sensitivity of 95%.

K-Nearest-Neighbor Classifier

The K-nearest-neighbor classifier (37) is a nonparametric method of pattern recognition used to determine the class of an object by its features vector. For a lesion belonging to the test set (query vector), it finds the K vectors closest to the query vector in the training set. The unclassified sample is then assigned to the class represented by the majority of the K closest neighbors. This method uses the nonparametric Bayes decision rule, which does not require prior knowledge of the distribution but instead relies on a training set of objects with known class membership to make decisions on the membership of an unknown object. In other words, it assigns an unknown object to the class with the highest \[ a \text{ posteriori} \] probability, using an Euclidean metric. A \[ a \text{ posteriori} \] probabilities are computed after estimating class-conditional densities. Accurate estimates of class-conditional densities require a large volume of training data. If there are
enough members in the training set, the probability of error for the K-nearest-neighbor classifier is sufficiently close to the Bayes (optimal) probability of error.

Training Set. The training set for the K-nearest-neighbor classifier consisted of 1081 histologically diagnosed skin lesions, including 428 (40%) melanomas and 653 benign pigmented lesions, 58 of which were atypical nevi. These lesions were selected from the image databases of several institutions using the same ELM instrumentation, i.e., IDI-Rome; Siena University Dermatology Clinic; IDI-Capranica; and the Italian Cancer League Clinics of Grosseto, Livorno, Arezzo, Trento, and Siena.

Lesion Classification. K-nearest-neighbor classifier accuracy was estimated on the histologically diagnosed lesions of the Rome and Siena centers by the jackknife procedure. The prevalence of melanomas among the first 100 closest neighbors was determined, and the lesion was assigned to the melanoma group if the prevalence was higher than a threshold value $T_{100}$. The method of receiver operating characteristic curves was used to identify the $T_{100}$ value necessary for a sensitivity of 98%.

Results

Description of the Image Sets

Table 1 shows the distribution of the lesions included in the training and test sets by histological diagnosis and center. The Siena University image database included a higher proportion of in situ melanomas. For each center, the random allocation of lesions to the training and test sets yielded two groups of lesions including melanomas of similar thickness and the same proportion of in situ melanomas and atypical nevi.

Description of the Features Selected for the Linear Discriminant Classifier

The means and SDs of the 10 selected parameters by diagnosis are reported in Table 2. For all parameters, a clear-cut difference between melanomas and nevi was observed. As an additional check of the feature selection procedure, a logistic regression model including all of the selected features was run to assess the association of each feature with melanoma, after adjusting for all of the others. All but two geometric parameters,

Table 1 Description of the image databases of histologically diagnosed pigmented lesions of the skin

<table>
<thead>
<tr>
<th></th>
<th>Melanocytic nevi</th>
<th>Melanomas</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>Common</td>
</tr>
<tr>
<td>IDI, Rome</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Training set</td>
<td>99</td>
<td>83</td>
</tr>
<tr>
<td>Test set</td>
<td>92</td>
<td>75</td>
</tr>
<tr>
<td>Total</td>
<td>191</td>
<td>158</td>
</tr>
<tr>
<td>Siena University</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Training set</td>
<td>121</td>
<td>92</td>
</tr>
<tr>
<td>Test set</td>
<td>137</td>
<td>114</td>
</tr>
<tr>
<td>Total</td>
<td>258</td>
<td>206</td>
</tr>
</tbody>
</table>

* n.a., not available.

Table 2 Mean and SD of 10 digital image parameters selected for the linear classifiers, by diagnosis

<table>
<thead>
<tr>
<th>Selected digital image parameters (dermoscopic correlates)</th>
<th>Melanomas ($n = 391$)</th>
<th>Nevi (histological diagnosis) ($n = 449$)</th>
<th>$P$ value ($t$ test)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Geometry</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Area inside the outline (size)</td>
<td>4.097</td>
<td>3.001</td>
<td>0.000</td>
</tr>
<tr>
<td>Variance of contour symmetry with respect to 180° axes</td>
<td>3.952</td>
<td>4.207</td>
<td>0.004</td>
</tr>
<tr>
<td>Fractality of borders (border indenting)</td>
<td>0.786</td>
<td>0.767</td>
<td>0.043</td>
</tr>
<tr>
<td><strong>Colors and Texture</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Skin lesion gradient (mean sharpness of lesion border)</td>
<td>23.785</td>
<td>10.138</td>
<td>0.000</td>
</tr>
<tr>
<td>Variance of skin lesion gradient histogram (variance of sharpness along lesion border)</td>
<td>60.152</td>
<td>77.872</td>
<td>0.000</td>
</tr>
<tr>
<td>Texture entropy (network analysis)</td>
<td>3.367</td>
<td>3.191</td>
<td>0.000</td>
</tr>
<tr>
<td>Isolations (Clusters of colors)</td>
<td>0.324</td>
<td>0.213</td>
<td>0.000</td>
</tr>
<tr>
<td>Imbalance of transition regions between lesion and healthy skin (imbalance with respect to “center of gravity” of skin-lesion transition regions)</td>
<td>0.324</td>
<td>0.213</td>
<td>0.000</td>
</tr>
<tr>
<td>Imbalance of blue-gray areas (imbalance with respect to “center of gravity” of areas of lesion tending to gray-blue color)</td>
<td>0.197</td>
<td>0.055</td>
<td>0.000</td>
</tr>
<tr>
<td>Gradient of the dark areas from lesion center to periphery (peripheral dark areas)</td>
<td>0.330</td>
<td>0.148</td>
<td>0.000</td>
</tr>
<tr>
<td>Number of border abruptions in red band (border abruptions)</td>
<td>6.115</td>
<td>4.866</td>
<td>0.000</td>
</tr>
</tbody>
</table>
i.e., fractality of the border and variance of the contour symmetry, were independently associated with melanoma (data not shown).

Classifier Performance

**Linear Discriminant Classifiers.** Table 3 shows the performance of three linear discriminant classifiers, two of which were trained on separate sets derived from the Rome and Siena centers, whereas the third was built on the pooled Siena and Rome training sets. The three classifiers were then independently tested on the Rome and Siena sets of lesions.

The first linear classifier, constructed on the Rome training set, with a fixed sensitivity of 95% reached a specificity of 83% on the Rome test set. When tested on the whole set of lesions belonging to the Siena Dermatology Department, a substantially stable performance was observed in terms of sensitivity (94%), whereas the specificity was 73% (Table 3). Similar results were obtained with the second classifier, constructed on the Siena University training set, which yielded a sensitivity of 93% and a specificity of 81% on the set of lesions from the Rome center.

**K-Nearest-Neighbor Classifier.** Table 3 shows also the performance of the K-nearest-neighbor classifier on the same sets of lesions. With a fixed sensitivity of 98%, a mean specificity of 79% was obtained on all sets of histologically diagnosed benign lesions, comparable with that obtained by the linear classifiers.

**Discussion**

Since the development of digital ELM, which allowed the acquisition and processing of high-quality images of pigmented skin lesions, there has been a growing interest in developing computerized image analysis (“machine vision”) and proper algorithms to distinguish with high accuracy subtle differences, unperceived by the human eye, between cutaneous melanoma and benign melanocytic lesions (26–34, 38–45). Thus, several research efforts have focused in these last few years on the possibility of introducing into daily clinical practice computer-aided classification or automatic machine vision to increase the accuracy of melanoma diagnosis. In fact, although dermoscopy seems to have a discriminant power significantly higher than clinical examination in classifying pigmented lesions, as documented in a recent meta-analysis (17), the accuracy of dermoscopy is highly variable across different studies and is still far from the desirable levels of 100% sensitivity and high specificity. Sources of variation are likely to arise from differences in sample sizes, proportion of melanomas in the sample, type of instrument used, dermoscopic criteria used, and, last but not least, human variability in feature recognition and coding.

Our study gives an important contribution to this research area for several reasons. First, it is, to our knowledge, the second study on computer classification of pigmented lesions that compares the performance of different automatic classifiers on independent test sets of lesions (46). Second, this is the only study that assessed the performance of the classifiers on distinct test sets of lesions taken by different instruments in different times and locations belonging to patients from two different population groups. Third, our study highlights the importance of factors such as classifier design and feature selection in computer-aided diagnosis that are generally overlooked in the previously published studies. We adopted a very conservative procedure of feature selection for the linear classifier to obtain a relatively small set of robust parameters to discriminate melanoma from benign melanocytic lesions. This strategy and the use of a hold-out (separate training and test sets) design allowed performance estimates that were likely conservatively biased (47–49). Although conservatively biased, the performances of the linear classifiers were remarkably accurate, with a mean sensitivity of 95% and a mean specificity of 77% and highly stable on sets of lesions derived from different dermatology centers, where the referral criteria for patients with pigmented lesions and the operating conditions of the instruments could have been different.

The most critical requirement of the K-nearest-neighbor classifier is to have a training set including enough examples of each class of pigmented lesions to adequately represent the full range of measurements that can be expected from each class. The use of a training set of 1081 lesions in our study allowed accurate computations of *a posteriori* probabilities, after estimating class-conditional densities, and an estimate of the Bayes error rate. With a misclassification rate for the K-nearest-neigh-

---

**Table 3** Performance of four classifiers on different sets of lesions, expressed as sensitivity (percentage of melanomas correctly classified) and specificity (percentage of melanocytic nevi correctly classified)

<table>
<thead>
<tr>
<th>Center</th>
<th>Sets of lesions</th>
<th>Linear classifier 1</th>
<th>Linear classifier 2</th>
<th>Linear classifier 3</th>
<th>K-NN classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDI Rome</td>
<td>Training</td>
<td>77</td>
<td>95%</td>
<td>95%</td>
<td>98%</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>78</td>
<td>95%</td>
<td>95%</td>
<td>98%</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>155</td>
<td>93%</td>
<td>95%</td>
<td>98%</td>
</tr>
<tr>
<td>Siena University</td>
<td>Training</td>
<td>107</td>
<td>95%</td>
<td>95%</td>
<td>98%</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>110</td>
<td>96%</td>
<td>95%</td>
<td>98%</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>217</td>
<td>94%</td>
<td>95%</td>
<td>98%</td>
</tr>
<tr>
<td>IDI Rome + Siena University</td>
<td>Training</td>
<td>184</td>
<td>95%</td>
<td>95%</td>
<td>98%</td>
</tr>
</tbody>
</table>

* Linear classifier 1, constructed on the Rome IDI training set; linear classifier 2, constructed on the Siena University training set; linear classifier 3, constructed on the pooled Rome and Siena University training sets.

b K-NN, K-nearest-neighbor classifier.
bor classifier of 12.5%, the Bayes error rate was greater than 6.25% and below 12.5%.

Comparing the performances of the two classifiers, the Bayes nonparametric approach, yielding a sensitivity of 98% and a specificity of 79%, seemed to give results similar to the geometrical linear discriminant approach (sensitivity of 95% and specificity of 77%). Optimizing the procedures of feature selection and weight definition could additionally improve the performance of the K-nearest-neighbor classifier.

In conclusion, our study suggests that computer-aided differentiation of melanoma from benign pigmented lesions obtained with DB-Mips is feasible and, above all, reliable. In fact, the same instrumentation used in different units on different data sets provided similar diagnostic accuracy. Although the bottom line in the diagnosis of melanoma is likely to continue to depend on the clinical insight of the physician and on the expertise of the pathologist, computer-aided diagnosis could provide clinicians an objective second opinion, at expert level, based on consistently extracting and analyzing image features. To what extent the combination of human and machine-based diagnoses would affect the decision-making process in the management of patients with pigmented lesions by improving the detection of early melanoma and/or decreasing unnecessary surgery remains to be evaluated by well-designed, randomized clinical trials in the field.

Acknowledgments

We thank Drs. M. Biagioli and P. Taddeucci from the Department of Dermatology, University of Siena, Siena, Italy; Drs. S. Pallotta, and G. Ferranti from the Istituto Dermopatico dell’Immacolata, Rome, Italy; and Drs. M. Nudo, G. Rossi, F. Carlesimo, and F. Lechiancole from the Istituto Dermopatico dell’Immacolata, Caprana, Italy.

References

Melanoma Computer-Aided Diagnosis: Reliability and Feasibility Study

Marco Burroni, Rosamaria Corona, Giordana Dell'Eva, et al.