Since advanced melanoma remains practically incurable, early detection is an important step toward a reduction in mortality. High expectations are entertained for a technique known as dermoscopy or epiluminescence light microscopy; however, evaluation of pigmented skin lesions by this method is often extremely complex and subjective. To obviate the problem of qualitative interpretation, methods based on mathematical analysis of pigmented skin lesions, such as digital dermoscopy analysis, have been developed. In the present study, we used a digital dermoscopy analyzer (DB-Dermo-Mips system) to evaluate a series of 588 excised, clinically atypical, flat pigmented skin lesions (371 benign, 217 malignant). The analyzer evaluated 48 parameters grouped into 4 categories (geometries, colors, textures and islands of color), which were used to train an artificial neural network. To evaluate the diagnostic performance of the neural network and to check it during the training process, we used the error area over the receiver operating characteristic curve. The discriminating power of the digital dermoscopy analyzer plus artificial neural network was compared with histologic diagnosis. A feature selection procedure indicated that as few as 13 of the variables were sufficient to discriminate the 2 groups of lesions, and this also ensured high generalization power. The artificial neural network designed with these variables enabled a diagnostic accuracy of about 94%. In conclusion, the good diagnostic performance and high speed in reading and analyzing lesions (real time) of our method constitute an important step in the direction of automated diagnosis of pigmented skin lesions.

Key words: epiluminescence light microscopy; digital dermoscopy analysis; melanoma; nevus; neural network

The incidence of cutaneous malignant melanoma has increased faster than that of all other cancers, except lung cancer in women.1,2 Annual incidence rates have increased on the order of 3–7% in fair-skinned populations in recent decades.3 Because advanced cutaneous melanoma is still incurable, early detection is an important step toward a reduction in mortality.4 Unfortunately, it is often difficult to differentiate early melanoma from other PSLs. High expectations are entertained for a technique known as dermoscopy or ELM.5–9 This noninvasive method enables visual inspection of structures under the skin surface by means of oil immersion; however, evaluation of pigmented skin lesions by this method is often extremely complex and subjective.10 To obviate the problems, certain research groups have developed equipment and methods that enable objective evaluation of the many parameters that have emerged after decades of experience with ELM.12 An example is DDA, which gathers numerical data and enables skin lesion images to be described objectively.13–19 DDA can be combined with sophisticated pattern recognition techniques, such as neural networks, opening promising scenarios in this field. ANNs are a large class of models developed in the cognitive sciences, the structure of which was inspired by that of the nervous system of living beings.20 Certain applications of ANNs in medicine have led to significant improvements in medical decision making.21–24 In a preliminary study, we used a digital dermoscopy analyzer together with a neural network on 147 clinically atypical PSLs (90 nevi and 57 melanomas), to determine its discriminating power with respect to histologic diagnosis. The system evaluated 48 objective parameters used to learn an ANN. Using the ANN with 10 variables selected by a stepwise procedure, we obtained maximum accuracy in distinguishing melanoma from benign lesions of about 93%.

In the present study, we used a digital dermoscopy analyzer (DB-Dermo-Mips system, Biomips Engineering, Siena, Italy) and an optimized ANN to evaluate a wider series of 588 excised, clinically atypical, flat PSLs, both malignant and benign in nature, including pigmented spitz nevi and pigmented basal cell carcinomas.

MATERIAL AND METHODS

Study population

Between 1996 and 2001, 4,200 PSLs were excised in our department, including 670 melanomas and 3,530 benign PSLs. Their images were digitized before excision and saved in a computer archive. From this archive, 588 clinically atypical (asymmetrical with variegated color), flat, impalpable PSLs, 0.4–1 cm in diameter and belonging to different subjects, were selected (40% male, mean age 49 ±15 years). All were difficult to diagnose and therefore suitable for morphologic and parametric evaluation of early melanoma. Blue nevi were excluded, as were lentigo maligna, lentigo maligna melanoma, acral PSL (different and peculiar dermoscopic patterns) and pink skin lesions (amelanotic melanoma and classical Spitz nevi). All PSLs were examined by 3 experienced dermopathologists (C.M., P.L. and P.T.) and identified as nevi (n = 371) and melanomas (n = 217).

Measurements

Lesions were imaged by ELM at a magnification of 16 with the DB-Dermo-Mips apparatus. The patient lay on the examination table, and the skin around the lesion was arranged orthogonally to the incident light. If the lesion and/or surrounding skin were hairy, the hairs were carefully removed with scissors or a razor. After 5

Abbreviations: ANN, artificial neural network; CCD, charge coupled device; DDA, digital dermoscopy analysis; ELM, epiluminescence light microscopy; MSE, mean square error; PSL, pigmented skin lesion; ROC, receiver operating characteristic; SE, sensitivity; SLP, single-layer perceptron; SP, specificity.

Grant sponsor: Italian Ministry of University Scientific and Technological Research.

*Correspondence to: Istituto di Scienze Dermatologiche, Università degli Studi di Siena, Policlinico “Le Scotte”, 53100 Siena, Italy.
Fax: +39-577-44238. E-mail: rubegni@unisi.it

Received 10 April 2002; Revised 1 July 2002; Accepted 8 July 2002

DOI 10.1002/jic.10620
min acclimatization at room temperature (25°C), the lesion was recorded as a digital signal and stored. Lesions were then removed surgically and examined histologically. All digital images were analyzed with appropriate algorithms.

**Equipment**

The DBDermo-Mips system consists of a 3CCD PAL Broadcast video camera with 730 lines of image resolution and a 60 db signal/noise ratio. The camera is connected to a patented hand-held 3CCD optic system with 5 magnifications from 6 to 40, allowing fields from 4 mm to 4 cm in size. The camera was calibrated weekly using special paper for white balance. Light was provided by a 150 W source at 3,200 K. Components of the video signal were connected to a frame-grabber interfaced to a Pentium III 500 MHz personal computer with a magneto-optical drive for image storage. The system (DBDermo-Mips) runs under Microsoft Windows, and all software was written in language C/C++ by M.B. and G.D.E.

**Digitization and parameterization**

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 microscopic magnifications along with the texture analysis, offering an objective evaluation, the different magnifications could confuse clinicians wanting to make subjective comparisons of lesions. Here, we only discuss the ×16 images for this reason. The system used a procedure for digital image processing based on a Laplacian filter for segmentation and a zero-crossing algorithm for the border automatic outline. It then evaluated 48 parameters for discriminant power. Reproducibility was first tested on digitized images of 100 lesions belonging to 20 subjects (1 PSL for each patient recorded 5 times at 15 min intervals). Absolute differences between single measurements and to 20 subjects (1 PSL for each patient recorded 5 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 previously described, belong to 4 categories: geometries, colors, textures and islands of color (i.e., color clusters inside the lesion). Briefly, 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 the border gradient, border homogeneity and interruptions of the border. Texture variables were mean contrast and entropy of lesion as well as contrast and entropy fractality. The islands of color variables were peripheral dark regions; dark area; imbalance of dark region; green area; red area; dominant green region imbalance; blue-gray area; blue-gray regions; transition area; transition region imbalance; background area; background region imbalance; red, green and blue multicomponent; and number of red, green and blue percentiles inside the lesion.

**ANN**

We used SLP-ANN (Fig. 1) with a number of input nodes equal to the size n of the input vector u of DBDermo-Mips parameters connected to the single neuron (perceptron) and designed to estimate the probability of melanoma risk. The perceptron had a summing junction and a suitably biased logsig activation function. Logsig is a nonlinear function

\[ y = \frac{1}{1 + e^{-x}} \]  

(1)

mapping the weighted, summed and biased inputs

\[ s = \sum_{i} w_i u_i + b \]  

(2)

in a continuous, monotonous, symmetrical way onto a scalar output value y between 0 and 1. The connection weights, w_i and the perceptron bias, b, are the SLP-ANN parameters to be estimated by iterative learning procedures. Feed-forward ANNs for classification, designed with an output logsig activation function, provide reliable estimates of class-conditional posterior probabilities, \( P(c_i | y) \), where \( c_i \) represents classes of nevi (\( j = 1 \)) or melanomas (\( j = 2 \)). Each input parameter was scaled before presentation to the network to have a 0 mean and unit standard deviation. Scaling increases the efficiency of ANN training. The perceptron output (target) was set at 1 for training examples of melanoma and at 1 for nevi; it therefore expressed the probability of melanoma.

The training process was continuously monitored to improve network generalization power. The simplest possible ANN architecture (only 1 neuron) was used to obtain the best compromise between model compatibility (the ability of the ANN to describe the data in the training set) and robust predictive performance (the capacity to provide constant classification performance even for new cases). Model compatibility can be evaluated by computing the error on the training set. Predictive power must be assessed on testing data by means of a cross-validation procedure. The SLP-ANN was tested to guarantee sufficient flexibility in learning the best discriminant model from examples of data. In particular, it was verified that it could overfit the training data given a sufficiently large number of training iterations (epochs). Overfitting means that the error on the training set can be driven, epoch by epoch, to a value significantly lower than the error on another equally representative set of known cases (testing set). Although the ability to overfit ensures model compatibility, as mentioned above, overfitting must nevertheless be avoided since it causes loss of generalization. We used the leave-one-out method of cross-validation to test and control overfitting directly during the training process. This method is particularly useful in biomedical applications when available data are usually scarce since it allows all of the data to be used efficiently both for training the classification model and for testing its diagnostic performance, albeit at the expense of simplicity and fast processing. For N available input–output data, it is based on constructing all N possible combinations of N-1 cases and using these combinations for N training sessions. The N cases left out (1 per session) were used to calculate the testing error. The commonly used MSE, i.e., the mean of the squared differences between real and network predicted outputs, was chosen. A batch training method that updates neural weights and bias only after all inputs and targets have been presented, i.e., after each epoch, was used. The Levenberg-Marquardt algorithm was used to minimize the MSE.

To train the SLP-ANN, 550 of the 588 available cases were used (30 nevi and 200 melanomas in 550 sessions, each with 2 subsets);
549 cases were used for training, and 1 case at a time was used to check overfitting and stepwise feature selection. A third small, independent subset consisting of the other 17 melanomas and 21 nevi was used to test SLP-ANN diagnostic performance on data not used in the training process.

Another well-known source of generalization loss is the use of too many variables as discriminant factors. Different sets of variables may provide largely overlapping data, and subsets of them therefore have the same discriminant power. The greater the number of discriminant variables, the more parameters have to be estimated by the ANN model. In SLP-ANN, each variable introduces a new connection weight. In general, for a given set of training data, identification of model parameters sharply deteriorates as their number increases. The immediate consequences are significant loss of model robustness. It is therefore advisable to select the best minimum subset of discriminant variables (also called features). We used a computer-aided stepwise technique to choose the number of discriminant features for optimum generalization.29 The SLP-ANN was definitively trained on all N cases for the number of epochs identified as the beginning of overfitting by the final step of the procedure.

To evaluate the diagnostic performance of the SLP-ANN and to control it during the training process, we used the error area over the ROC curve.

RESULTS

General data
Histologic examination of the 588 PSLs revealed 223 common melanocytic nevi, 137 melanocytic nevi with architectural disorder with or without cytologic atypia, 11 pigmented Spitz nevi and 217 melanomas (60 in situ melanomas). The median invasion depth of invasive melanomas was 0.4 mm. Histopathologic diagnosis of melanoma and nevi was made according to the criteria of the NIH Consensus Conference.50 Histopathologic diagnosis showed discordance among pathologists of about 9%. Such cases were classified as melanoma or nevi when at least 2 of 3 dermopathologists agreed on the diagnosis.

SLP-ANN

The feature selection procedure indicated that only 13 variables were sufficient to discriminate melanomas and nevi as well as possible, and this also enabled good generalization power (Table I). In the group of geometric variables, minimum diameter was the only discriminant variable; among color variables, mean red, red tenth and blue quartile were chosen. Contrast was the only discriminant texture variable, and all of the other discriminant variables belonged to the islands of color group.

To define the ROC curve, 101 pairs of SP and SE were calculated by classifying cases as melanoma when the SLP-ANN output (posterior probability) was greater than a probability threshold ranging from 0 to 1 in steps of 0.01; otherwise, they were classified as nevi. The ROC curve was therefore obtained by plotting the values of 1 SP against SE (Fig. 2). The gray area over the curve was associated with error in the classification. The cost of clinical decision making can be specified by choosing a desired level of SE/SP on the ROC curve. This determines the probability of a skin lesion being classified as melanoma. As an example, Table II lists 5 different choices of the decision probability, each with an SE greater than the corresponding SP. In particular, SE = 99% could be reached with SP = 72.5%, which is still good; and a very high SE (94.3%) was even possible with a similar value of SP (93.8%). Figure 3 shows the diagnostic performance of the SLP-ANN classifier on the independent testing set. With the exception of 1 nevus, the predicted risk probabilities related to nevi (white circles) were distinctly separated from those of melanomas (black circles). For example, if we consider a decision probability of 0.28 (last row of Table II), all 17 melanomas tested were identified correctly, and only 1 of 21 nevi was classified as likely melanoma.

DISCUSSION

We have moved away from completely qualitative diagnoses toward assignment of scores on certain visual characteristics of the lesion, approaching a mechanistic model that can be more reliably implemented using digital computerized analysis. In this new area, the prominent role of clinicians has been to provide the expertise for image interpretation and rationalization of the thought process that leads from image to diagnosis. The role of engineers, physicists and statisticians has been to design new pattern analysis and classification techniques by which the diagnostic process might be replicated with a high degree of precision.

In the present study, we used a digital dermoscopy analyzer and a suitable ANN to achieve high diagnostic accuracy (94%). In particular, neural networks have found many successful applications in medicine.20–25 The flexibility of ANN models, which like humans learn from example, combined with the reliability of automatic digital image processing of the DBDermo-Mips system makes the classification method adaptable to any other environmental/clinical conditions so that a given center/department can use and tune the ANN classifier most suitable not only for the type of lesion to be observed/diagnosed but also for its histopathologic standard. All that is required is a sufficiently large representative set of examples to train the ad hoc ANN. The training set can even be updated with new lesions that the clinician considers the system should learn. The SLP-ANN method is therefore easy to train again for including the characteristics of new lesions.

<table>
<thead>
<tr>
<th>Variables (description) (measure)</th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peripheral dark (gradient of dark regions from the center to periphery of lesions) (%)</td>
<td>0.241</td>
<td>0.195</td>
</tr>
<tr>
<td>Contrast (evaluated by co-occurrence matrices) (0–4)</td>
<td>1.380</td>
<td>0.295</td>
</tr>
<tr>
<td>Minimum diameter (mm)</td>
<td>5.521</td>
<td>3.003</td>
</tr>
<tr>
<td>Background region imbalance (%)</td>
<td>0.072</td>
<td>0.062</td>
</tr>
<tr>
<td>Transition area (between lesion and healthy skin) (%)</td>
<td>0.228</td>
<td>0.055</td>
</tr>
<tr>
<td>Background area (area of the region between mean color of the lesion and dark regions) (%)</td>
<td>0.086</td>
<td>0.056</td>
</tr>
<tr>
<td>Red lesion decile (decile of histogram of red in lesion) (0–31)</td>
<td>9.154</td>
<td>4.247</td>
</tr>
<tr>
<td>Transition region imbalance (between lesion and healthy skin) (%)</td>
<td>0.170</td>
<td>0.094</td>
</tr>
<tr>
<td>Red lesion value (mean red value of pixels inside lesion) (0–31)</td>
<td>12.822</td>
<td>4.026</td>
</tr>
<tr>
<td>Skin-lesion gradient (mean sharpness of lesion border) (0–100)</td>
<td>15.096</td>
<td>11.102</td>
</tr>
<tr>
<td>Blue lesion quartile (quartile of histogram of blue in lesion) (0–31)</td>
<td>3.841</td>
<td>1.898</td>
</tr>
<tr>
<td>Circularity (difference between lesion and circle of equal area) (%)</td>
<td>0.783</td>
<td>0.098</td>
</tr>
<tr>
<td>Green skin value (mean green value of pixels in skin around lesion) (0–31)</td>
<td>17.753</td>
<td>2.078</td>
</tr>
</tbody>
</table>
Dermoscopy improves diagnostic accuracy by 5–35% with respect to simple clinical observation depending on the type of skin lesion and physician experience. Indeed, the type of lesion evaluated in the various studies may give completely different classification results. Since research in this field is mainly orientated toward early diagnosis of cutaneous melanoma, like other research groups, we tested lesions difficult to diagnose and included melanoma simulators (Reed nevi) with the aim of developing increasingly sophisticated and specific diagnostic algorithms. This is why we excluded advanced melanomas, blue nevi and seborrheic keratoses, most of which can be diagnosed correctly on the basis of clinical or ELM criteria. Their inclusion would have slightly worsened the diagnostic accuracy of the lesions that concern us most. In other words, we do not want to misclassify a single melanoma to gain a correct diagnosis of seborrheic warts and/or blue nevi. Amelanotic melanomas were also excluded but for a different reason, namely, that they are clinically different from the lesions considered here and their borders cannot be well defined. Besides the type of skin lesion, another factor that can significantly affect the results obtained with classical ELM is physician experience. Binder et al. demonstrated that ELM pattern analysis increases the diagnostic performance of ELM experts but decreases that of clinicians not specially trained for ELM with respect to clinical observation. Dermoscopy is based essentially on the definition and recognition of fine morphologic and chromatic features of stuctures that are difficult to describe unequivocally. This is confirmed by a recent trend toward maximum simplicity in the description of ELM patterns of PLS. This simplification has enabled the development of diagnostic algorithms with good accuracy and reproducibility. Indeed, if we consider studies similar to the present one with regard to the type of lesion investigated, Kittler et al. obtained a diagnostic accuracy of 87% using the so-called ABCD score. Argenziano et al. compared different diagnostic algorithms for clinically doubtful melanocytic skin lesions and obtained better results with their method, a 7-point checklist (95% SE and 75% SP). However, these diagnostic algorithms have problems that have not yet been solved. The most important is that the purpose for which they were designed was not achieved since the within- and between-observer concordance rates are very low, even for expert observers. Our system proved to have similar diagnostic accuracy to the above studies but the considerable advantage of being independent of the examiner. In particular, when we chose a point of equal SE and SP on the ROC curve, we obtained accuracies around 94%, though when the model was used for new lesions it was necessary to consider the positive and negative predictive values, i.e., the percentages of classified cases that are really melanomas and nevi, respectively. It is well known that prevalence significantly affects predictive values. Caution is needed when using values of prevalence estimated on the whole population because the prior probability of melanoma in persons undergoing DDA may be significantly higher than the known prevalence. The more the melanoma prevention clinic is specialized, the higher the percentage of melanomas among observed lesions (prevalence in that particular clinical environment). In our study, the posterior probability of melanoma (positive predictive value) was estimated by SLP-ANN using a prior probability of about 36% (200 melanomas in 550 cases), which is undoubtedly higher than the real clinical probability. Our patients were referred by dermatologists. For correct practical use, posterior probability should be combined with decision probability, which defines classification performance objectively in terms of SE and SP chosen on the ROC curve. This makes it possible to adapt the classification method to clinical aims. To use the method for screening, minimizing false-negatives, we can use a decision threshold with a high SE. In this case, a lower prevalence will only improve negative predictive value; i.e., it will decrease the risk of not recognizing malignant lesions.

In conclusion, in the diagnosis of cutaneous malignant melanoma, instrument data should obviously not be the only basis on which the need for surgical excision of a suspect lesion is assessed. Like other methods, whether clinical or instrumental, our system is not sufficient alone and should be supplemented with medical history, clinical evaluation and, above all, common sense.

**TABLE II** - SP and SE calculated by classifying cases as melanomas when the SLP-ANN output (posterior probability) was greater than a chosen probability threshold (decision probability), otherwise as nevi.

<table>
<thead>
<tr>
<th>Decision probability</th>
<th>SE</th>
<th>SP</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>1.000</td>
<td>0.469</td>
</tr>
<tr>
<td>0.16</td>
<td>0.991</td>
<td>0.725</td>
</tr>
<tr>
<td>0.22</td>
<td>0.973</td>
<td>0.858</td>
</tr>
<tr>
<td>0.25</td>
<td>0.964</td>
<td>0.929</td>
</tr>
<tr>
<td>0.28</td>
<td>0.943</td>
<td>0.938</td>
</tr>
</tbody>
</table>

**FIGURE 2** - ROC curve describing the SLP-ANN classification performance. Gray area represents classification error.

**FIGURE 3** - Risk probabilities predicted by the SLP-ANN classifier for testing data set (21 nevi and 17 melanomas).
REFERENCES