Automated melanoma detection: Multispectral imaging and neural network approach for classification

Stefano Tomatis[a)]
Department of Medical Physics, Istituto Nazionale per lo Studio e la Cura dei Tumori, Milan, Italy

Aldo Bono and Cesare Bartoli
Melanoma Unit, Department of Day Surgery, Istituto Nazionale per lo Studio e la Cura dei Tumori, Milan, Italy

Mauro Carrara and Manuela Lualdi
Department of Medical Physics, Istituto Nazionale per lo Studio e la Cura dei Tumori, Milan, Italy

Gabrina Tragni
Department of Pathology and Cytopathology, Istituto Nazionale per lo Studio e la Cura dei Tumori, Milan, Italy

Renato Marchesini
Department of Medical Physics, Istituto Nazionale per lo Studio e la Cura dei Tumori, Milan, Italy

(Received 7 March 2002; accepted for publication 20 November 2002; published 21 January 2003)

Our aim in the present research is to investigate the diagnostic performance of artificial neural networks (ANNs) applied to multispectral images of cutaneous pigmented skin lesions as well as to compare this approach to a standard traditional linear classification method, such as discriminant function analysis. This study involves a series of 534 patients with 573 cutaneous pigmented lesions (132 melanomas and 441 nonmelanoma lesions). Each lesion was analyzed by a telespectrophotometric system (TS) *in vivo*, before surgery. The system is able to acquire a set of 17 images at selected wavelengths from 400 to 1040 nm. For each wavelength, five lesion descriptors were extracted, related to the criteria of the ABCD (for asymmetry, border, color, and dimension) clinical guide for melanoma diagnosis. These variables were first reduced in dimension by the use of factor analysis techniques and then used as input data in an ANN. Multivariate discriminant analysis (MDA) was also performed on the same dataset. The whole dataset was split into two independent groups: i.e., train (the first 400 cases, 95 melanomas) and verification set (last 173 cases, 37 melanomas). Factor analysis was able to summarize the data structure into ten variables, accounting for at least 90% of the original parameters variance. After proper training, the ANN was able to classify the population with 80% sensitivity, 72% specificity, and 78% sensitivity, 76% specificity for the train and validation set, respectively. Following ROC analysis, area under curve (AUC) was 0.852 (train) and 0.847 (verify). Sensitivity and specificity values obtained by the standard discriminant analysis classifier resulted in a figure of 80% sensitivity, 60% specificity and 76% sensitivity, 57% specificity for the train and validation set, respectively. AUC for MDA was 0.810 and 0.764 for the train and verify set, respectively. Classification results were significantly different between the two methods both for diagnostic scores and model stability, which was worse for MDA.
 [DOI: 10.1118/1.1538230]

Key words: melanoma, diagnosis, computer, neural networks

I. INTRODUCTION

The early detection and prompt surgery represent the only effective management of the patient affected by cutaneous melanoma.[1,2] The ABCD (Asymmetry, Border, Color, Dimension) clinical criteria were introduced for assisting the visual recognition of early melanoma[3] and further improvements in diagnostic accuracy have been reported through the use of dermatoscopy.[4,5] As with any clinical diagnosis, the evaluation of a pigmented lesion, either on clinical grounds or with dermatoscopy, is, to a large extent, subjective and the experience acquired by trained physicians is hard to transfer outside specialized centers.

In an attempt to objectively evaluate the involved clinical characteristics, an interest in the automated diagnosis of melanoma have rapidly grown in the last years. Different techniques based on image acquisition of skin lesions has been proposed by the use of different kind of image sources, such as photographic, color-camera,[6–8] and epiluminescence microscopy images.[9–15]

Automated classification is generally based on the evaluation of various shape, color, and texture features extracted by computerized processing of the lesion image. These features are combined to give an output score through a proper function determined by a specific statistical classification tool, after a training process. The score is finally compared against a threshold value to perform a lesion diagnosis.

The use of artificial neural network (ANN-) based classifiers, applied to recognition purposes in automated systems for melanoma detection, has been proposed since 1994,[16] and recently applied to dermatoscopic images.[12] The ANN approach is suggested as an alternative classification method to traditional multivariate linear techniques, such as multiple discriminant analysis, because it frequently provides reduced error rates and is a flexible means for mapping a fixed number of inputs into a set of discrete classes, with the advantage of permitting the creation of nonlinear decision boundaries in the input space.[12,16,18]

In the context of a computer-assisted diagnosis of melanoma, a spectrophotometric method based on measurements of lesion reflectance has been first proposed and developed by us.[19–23] This method has been extended and improved by others[17,18,24] using optimized devices. In this study we revise our experience in light of a reelaboration of the data from multispectral reflectance images of cutaneous moles. Our aim in the present research is to investigate the diagnostic performance of ANN applied to multispectral images as well as to compare this approach to a standard traditional linear classification method, such as discriminant function analysis, which has been used by many authors for automated melanoma detection.[9,17,18,22,25]

II. PATIENTS AND METHODS

A. Patients

Between January 1995 and March 2000, 534 patients (319 females and 215 males) with 573 cutaneous pigmented lesions were enrolled in the study at the Istituto Nazionale Tumori of Milan. These patients, bearing lesions that required a surgical biopsy for diagnosis, were seen in our unit for the early diagnosis of a melanoma. Lesions excluded were those appearing as thick and/or large melanomas and awkwardly situated lesions, like those placed at the interdigital spaces, on ears, nose, eyelids, etc. In fact, a lesion can be accurately imaged by our telespectrophotometric system (TS) if surrounded by a portion of almost planar skin 5 mm or more distant from its margins. We also excluded lesions on the scalp due to hair interference on reflectance. The median age of the patients was 36 years (range 10–95). The size of the lesions ranged from 3 to 39 mm in the maximum linear extent with a mean value of 10 mm. Images of the 573 pigmented lesions were acquired *in vivo* before surgery. The slides were evaluated according to widely accepted criteria for the hystopathological diagnoses of the various pigmented lesions.[26] The distribution of the lesions according to the histological diagnosis is represented in Table I. The thickness of the 113 invasive melanomas ranged from 0.16 to 3.24 mm (median 0.68 mm). Of the 132 melanomas, 91 were thin lesions (tumor thickness$<$1 mm or level I).

The 573 cases of our series were analyzed ranked in their original (chronological) order. In addition, the last 173 were from patients consecutively visited at our institution and were used as a verify set for validation and comparison of the spectrophotometry-based classification procedures. This last set was acquired within a 1 year period between April 1999 and March 2000.

TABLE I. Histological diagnosis of 573 pigmented lesions.

Diagnosis	Number (%)
Melanoma	132 (23.0)
Invasive	113 (19.7)
In situ	19 (3.3)
Nonmelanoma	441 (77.0)
Compound nevus	244 (42.6)
Dysplastic nevus	40 (7.0)
Junctional nevus	44 (7.7)
Dermal nevus	26 (4.5)
Lentigo simplex	19 (3.3)
Spindle-cell nevus	15 (2.6)
Spitz nevus	7 (1.2)
Basal cell carcinoma	9 (1.6)
Blue nevus	10 (1.7)
Seborroheic keratosis	9 (1.6)
Other	18 (3.1)
Total	573 (100)

B. Image acquisition

The Telespectrophotometric System consists of a CCD camera, a set of 17 interference filters, a PC, and an illumination system composed by two halogen (2×100 W) and two lamps (2×150 W) with emission in the infrared region. Even illumination of the sample is obtained by the use of diffusing surfaces placed in front of each lamp. The system allows the reflectance imaging of moles at 17 selected wavelengths from 420 to 1040 nm. The filter to filter wavelength interval is about 40 nm.

A frame grabber installed in the computer allows the capturing of multispectral images of mole reflectance within a useful area of 4×5 cm^2 with a spatial resolution of 3.5 pixel/mm and 256 gray levels for each single wavelength. The acquired 17 spectral images are stored in the PC for offline processing. Intensity levels as well as pixel dimension were calibrated according to a set of four reflectance standards and a geometric reference frame, respectively. Details on the system's features and calibration procedures have been reported elsewhere.[21–23]

C. Image analysis

For each spectral image, the system provided five descriptors related to the color and shape of the imaged lesion. These five descriptors, which could represent the clinical features included in the ABCD rule,[3] were defined as follows: (1) *mean reflectance* (MR), i.e., the mean fraction of light reflected or diffused by the lesion; (2) *variegation index* (VI), the standard deviation of the measured reflectance within the lesion; (3) *compactness* (Cmp), defined as (lesion perimeter)^2/($4\pi\cdot$lesion area); (4) *roughness* (Rgh), i.e., the ratio between the lesion perimeter and the perimeter of the convex hull of the lesion; (5) *area* (A), i.e., the lesion area.

The variables related to lesion shape features (i.e., Rgh, Cmp, and A) were determined for wavelengths below 623

nm, where the skin-lesion contrast allowed an optimal recognition of the lesion contour for all cases. The mean reflectance and variegation index were evaluated for all visible and near-infrared wavelengths on the basis of the contour determined by the wavelength of maximum contrast between lesion and skin (498 nm).

D. Statistical analysis

Standard descriptive statistics and univariate significance tests, such as the Student's *t* test, were performed on the dataset.

In order to assess the discriminating capability of both ANN and discriminant analysis the whole set of data was split into a train and a verify set composed by the first 400 cases (95 melanomas) and the remaining 173 cases (37 melanomas), respectively. The train set is required for the instruction of the classifiers, whose diagnostic performances are evaluated against an independent verify set. In the present study the sample size is not small, however, since melanoma is a rare pathology with a variety of different visual characteristics, the considerable fraction of cases devoted to the classifier training was selected to include an adequate number of positives, leaving a significant portion of cases for validation.

Lesion classification is expressed as sensitivity (i.e., the fraction of correctly classified melanomas), specificity (i.e., the fraction of correctly classified nonmelanoma lesions), and accuracy (i.e., the fraction of correctly classified lesions).

Results from both discriminant analysis and neural network classifications were compared by means of receiver operating characteristic (ROC) analysis for both train and verify data. The ROC is a curve showing the various tradeoffs existing between the proportion of true positives and false positive responses, as the decision criterion is systematically varied, i.e., for a given capacity to discriminate between positive and negative cases.[27] From a diagnostic test making a complete discrimination between two distributions, a ROC curve moving toward the left and top boundary of the graph will result; on the contrary, with experimental values unable to make a discrimination, the ROC curve approaches the diagonal line. The area under the ROC curve (AUC) was evaluated as a useful index of the classification ability of the system.

A comparison between results obtained by the two different classification methods was also performed after setting the classification threshold at the 80% diagnostic level for sensitivity in the train set. In this case, differences were tested by means of a contingency table analysis using the McNemar chi square test. The significance of the differences between AUCs was estimated according to Hanley and McNeil.[28,29] All statistical outputs in this study were obtained by the aid of commercially available software (statistica, Stat Soft, Tulsa OK).

E. Factor analysis

Factor analysis was applied to each of the above-described variables, to summarize their wavelength dependence and to obtain a reduced but conceptually meaningful set of new variables, called factors. As a preliminary step, each variable was standardized by the simple linear transformation $V' = (V - \langle V \rangle)/\mathrm{SD}_V$, were V, V' are, respectively, the original and the standardized variable, $\langle V \rangle$ is the corresponding mean value, and SD_V is the variable standard deviation. A further logarithmic transform was applied before standardization to the shape-related variables Cmp, Rnd, and A, to obtain more compact and regular distributions for these parameters.

The factor analysis model used here is a principal-components analysis (PCA). By this technique, the new factors (principal components) used to describe data are determined as mutually uncorrelated linear combinations of the initial variables. Since it is usually difficult to determine whether a factor is more related to one particular interpretable subset of the variables than to another, additional methods aimed to increase the interpretation of data after performing PCA have been developed. All these methods perform a suitable rotation of the reference system in the factor loading space, where factor loadings are defined as the correlation coefficients between factors and initial variables. In this study, the varimax orthogonal rotation algorithm was applied. Factors were retained as new parameters if the ratio between the sum of the selected factors variances over the total variation (the sum of original variables variances) was at least 0.9 (variance explained$=90\%$).

A more complete treatment of the subject can be found described in the literature.[30]

F. Neural networks

An ANN can be considered as essentially composed of a multitude of elemental computing elements (called neurons) organized as a network and reminiscent of the way in which neurons are believed to be interconnected in the brain.

Even though many different network arrangements are possible in principle, the configuration of a three layered network with one hidden layer and one output neuron was considered for our purposes, since, although simple in its architecture, it provides enough flexibility in the description of data and it has been already successfully used to classify pigmented skin lesions.[12,16,18]

The training procedure involves the evaluation of the difference (error) between the network output and the known (target) coded output (histology) of the set used for training. Different procedures have been set up to minimize the error by iterative algorithms, such as back-propagation or a conjugate gradient descent. The training progress can be monitored at each step (epoch) by evaluating at the same time the error of the verify set. A decrease in the train error at the expense of an increase of the verify error often indicates a loss in the generalization capability of the classifier. Different criteria to decide when to stop the learning process can be applied. Iterations can be stopped when the training error function and/or its relative change fall below some specified value, or, according to a method called early stopping, the training process can be terminated when a minimum in the verify error is detected.

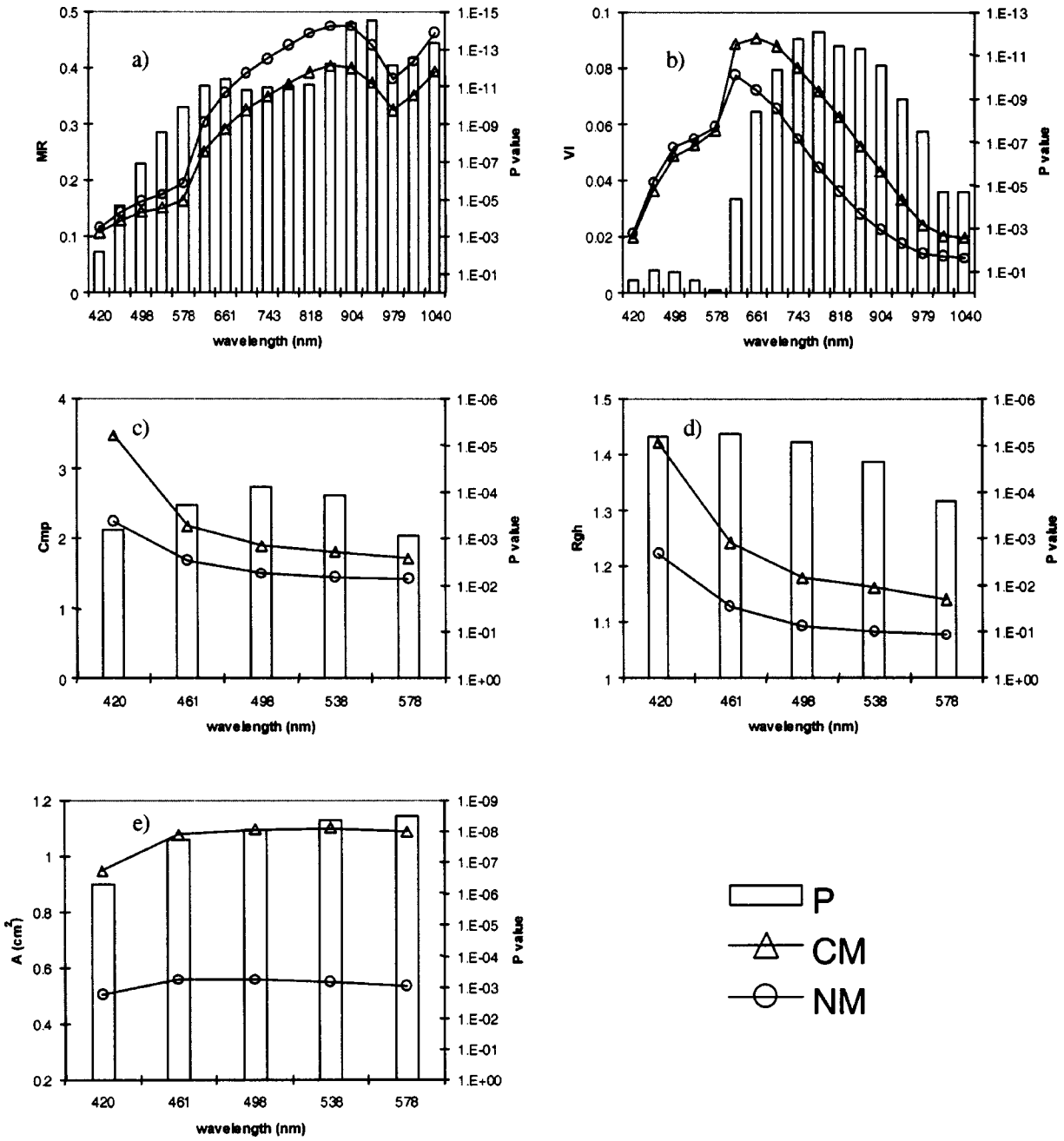

FIG. 1. Lesion descriptors. Mean values of melanomas (△, CM) and nonmelanoma lesions (○, NM). Bars, *t*-student univariate *P* values between the two groups. (a) mean reflectance; (b) variegation index; (c) compactness; (d) roughness; (e) area.

A full general review on ANNs and some of their applications to image processing can be found in the literature.[31,32]

G. Discriminant analysis

For group classification, MDA is an established statistical method operating by combining input variables into a discriminant function that could be used to classify future cases. The idea underlying this classification method is to develop a linear combination, L, of p variables, say, $L=\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p$ with values for β_1, $\beta_2,\ldots,$ β_p chosen so as to provide a maximum discrimination between the two populations. If the function L is going to discriminate between the two groups, the variations in the values of L *between* the two groups (B) should be much greater than the variation in the values of L *within* the two groups (W). The determination of the β coefficients is accomplished by maximizing the ratio B/W of sum of squares *between* and *within* the two classes. Once the discriminant function is derived using cases belonging to the train set, classification of a new case can be performed by comparing the score, i.e., the value of the discriminant function L for the given lesion, to a proper threshold.

Detailed descriptions of the statistical procedures in-

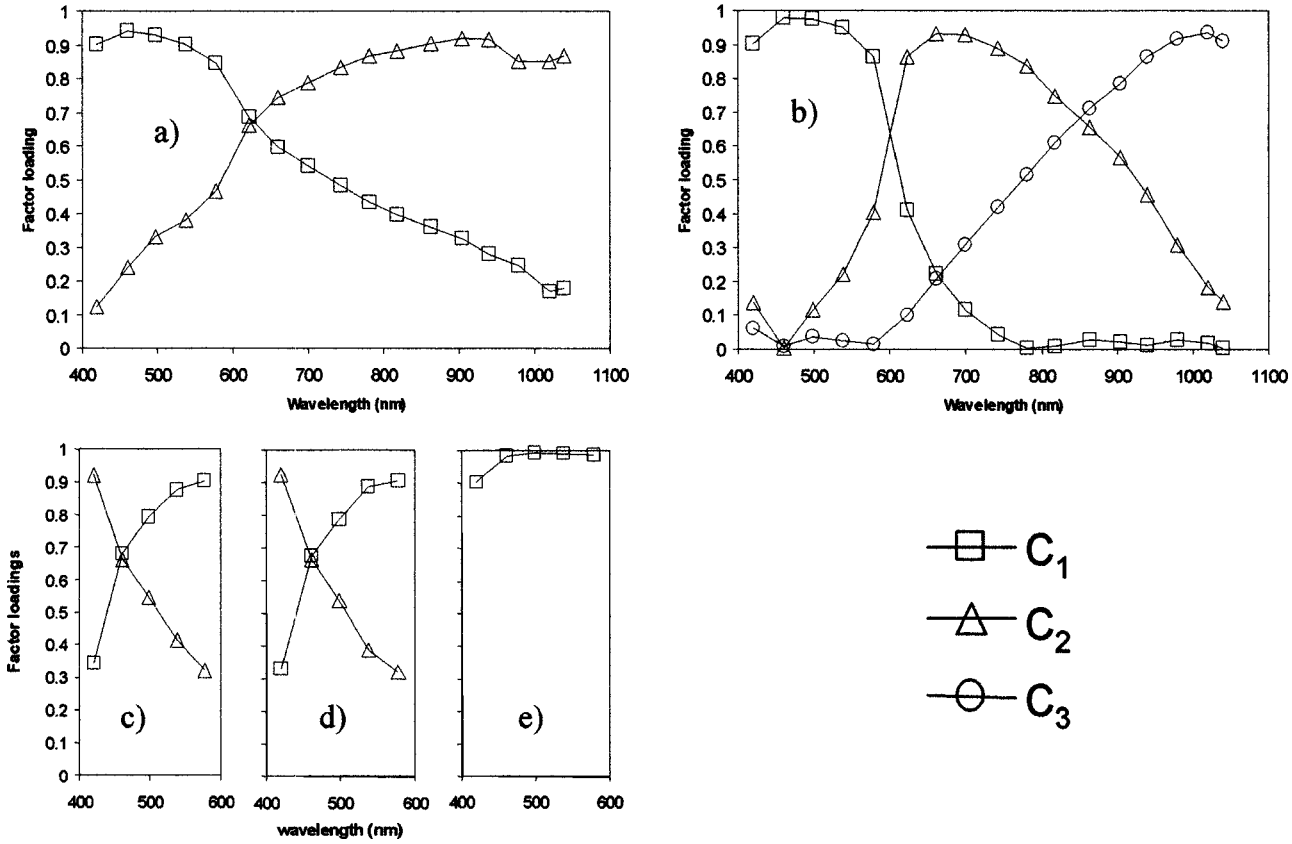

FIG. 2. Factor loadings for MR (a), VI (b), Cmp (c), Rgh (d), and A (e). $C_{1,2,3}$, first, second, and third principal component of variables. The varimax rotation method was applied to all variables except A.

volved in the assessment and use of a linear discriminant analysis model can be found in the literature.[30]

III. RESULTS

A. General statistics and factor analysis

Mean values of each evaluated parameter within the respective wavelength range of definition are plotted in Fig. 1 for all cutaneous melanomas (CM) and the other nonmelanoma lesions (NM). P values from the Student's *t* test are also reported in the figure as bars, to show the significance of the observed differences between the two groups.

Factor analysis was able to greatly reduce the wavelength dependence of the selected variables. After an application of the data reduction procedure to the whole dataset, a total of only 10 parameters out of the initial 49 was obtained.

Factor loadings for color- (MR and VI) and shape-related (Cmp, Rgh, and A) variables are plotted in Fig. 2, where each factor is indicated by marking the variable name with a progressive numerical label. The percentage of the total variation in the data explained by the different factors was 90.19 and 93.5 for MR and VI, whereas values of 93.08, 92.75, and 94.68 were obtained for Cmp, Rgh, and A, respectively.

For each variable, plots in Fig. 2 allow an interpretation of associated factors: a factor with high loadings in a determined wavelength interval describes the main behavior of the corresponding variable in the same spectral region.

The reduced variable set was also subjected to a univariate Student's *t* test: all ten factors resulted highly significant ($P<0.01$) except VI_1 (not significant, $P=0.06$) and MR_1 (significant at the 5% level, $P=0.02$). T-test results are shown in Table II. The correlation coefficient, *r*, between factors resulted to be consistently high (i.e., $r>0.5$) between roundness- and roughness-related components ($r=0.96$ between Cmp_1 and Rgh_1; $r=0.77$ between Cmp_2 and Rgh_2), and between MR_2 and VI_2 ($r=-0.63$).

B. Classification of lesions

In this work, we found that optimal results could be obtained even with a relatively simple architecture of a network composed by a single hidden layer with five neurons and one output neuron to classify lesions into the two different groups of CM and NM. This network architecture is complex enough to permit a significant nonlinear setting of the decision boundary arrangement, although simple enough to reasonably reduce the risk to overfit the data by following their noise distribution. The network final structure was attained on a trial and error basis, after various attempts, performed by changing both neuron configuration and input components. Not all the extracted factors were selected as inputs for the network. Factors not chosen were scarcely effective for

TABLE II. *T* test results on the ten components resulting from factor analysis.[a]

Factor	NM (mean value)	CM (mean value)	*t* value	P^b
MR_1	0.05	−0.18	2.41	0.02
MR_2	0.19	−0.63	8.77	0.00
VI_1	0.04	−0.14	1.90	0.06
VI_2	−0.11	0.38	−5.09	0.00
VI_3	−0.14	0.45	−6.09	0.00
A_1	−0.19	0.63	−8.73	0.00
Rgh_1	−0.13	0.44	−5.88	0.00
Rgh_2	−0.09	0.31	−4.17	0.00
Cmp_1	−0.11	0.36	−4.79	0.00
Cmp_2	−0.11	0.36	−4.78	0.00

[a]Variables scaled to 0 mean and 1 standard deviation.
[b]$P=0.00$ means $P<0.000\,05$.

discrimination when inserted in the model with the other potential input components. A final selection of input factors was MR_1, MR_2, VI_2, A_1, and Cmp_2.

This network was trained by applying first back propagation (BP) and then the conjugate gradient descent (CGD) error minimization algorithms. The best model found had the training stopped after 325 epochs (200 with BP and then 125 with CGD). The corresponding sum of squares error in the train and verify set was of 0.339 and 0.336, respectively. In Fig. 4 the error plot is shown at each epoch for train and validation sets: iteration lasted when the minimum in the verify error was found. Figure 3(a) shows ANN ROC curves for both the train and verify set. The same variables were used as input of the linear discriminant analysis model, and the corresponding ROC curves for the train and verify set are reported in Fig. 3(b). Classification results for both ANN and MDA are reported in Table III.

Sensitivity and specificity pairs corresponding to classifications derived after selecting the 80% threshold in sensitivity for the train set are represented as specific points on the ROC curves in Fig. 3 (a center of circles, where closed symbols indicate the verify set).

IV. DISCUSSION

The results reported in Table III and Fig. 3 show that ANN diagnostic performance is significantly better than that from a MDA linear technique. This finding agrees with the conclusions of another study comparing the two classification methods, even if using data based on punctual (not imaging) spectrophotometric measurements in the wavelength range between 320 and 1100 nm.[18] In that study accuracy values for the two methods of 86.7% (ANN) and 72.7% (MDA) are reported over a set of 74 lesions (26 melanomas), with sensitivity and specificity values, respectively, of 83.2% and 88.9% (ANN); 75.2% and 71.1% (MDA). The involved ANN is based on a set of 7 input features from reflectance spectra, 1 hidden layer (10 units), and a dichotomous output. Furthermore, differences in the collected cases do not allow a significant comparison between our classifications and those of the previously cited study. In fact, cases in Ref. 18 differ from ours both with respect to sample size and the histologic type of lesions.

A recent paper[24] reports results by an instrumentation similar to that used by us,[21] although with enhanced acquisition features, especially in spatial resolution, to allow spectrophotometric imaging at dermatoscopic magnifications. In that paper a linear and a nonlinear classifier were compared in the attempt to separate between melanoma and melanocytic nevi on a series of 246 lesions (63 melanomas). However, the nonlinear classifier was not of the ANN type and its working mechanism not fully explained by the authors. In addition, the performance achieved by the nonlinear classifier (95% sensitivity, 68% specificity) was worse than the one reached by the linear one (100% sensitivity, 84% specificity), which was trained with a different set of input vari-

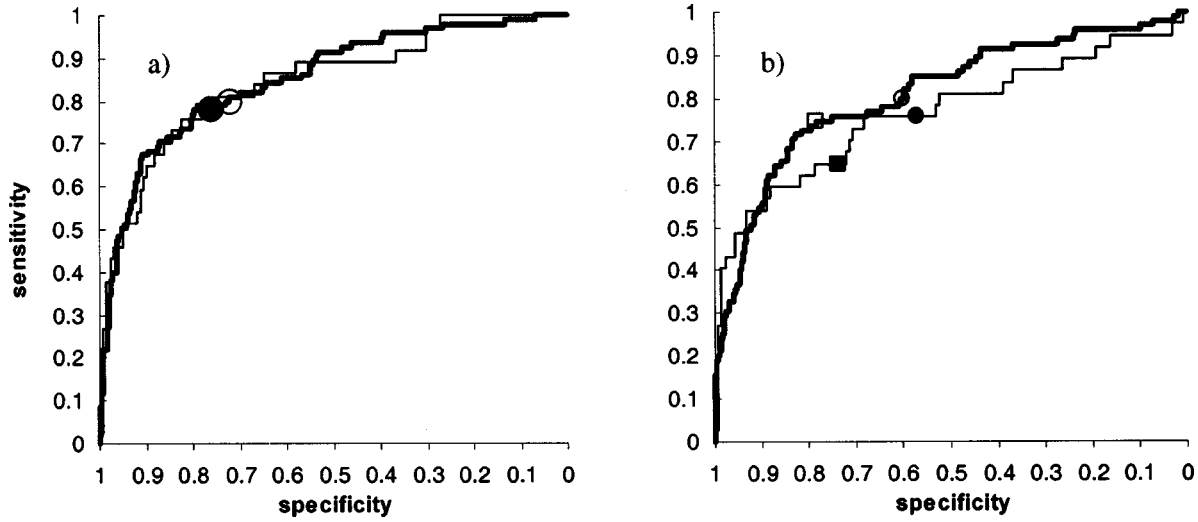

FIG. 3. ROC curves for ANN (a) and MDA (b) for the train (—) and verify set (—). Symbols, specific points (center of symbols) used for the diagnostic rates estimation after setting a threshold of 80% (circles), or 75% (squares) sensitivity in the train set. Closed symbols, verify set.

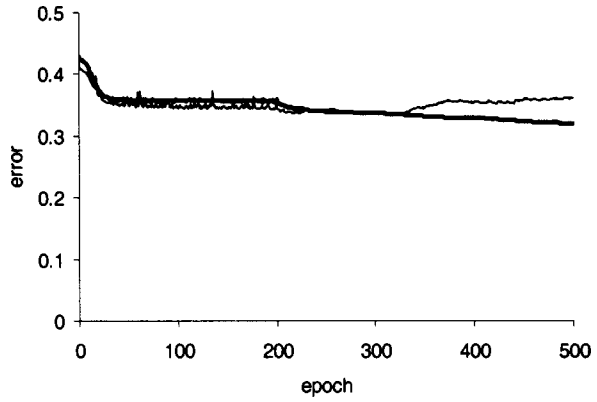

FIG. 4. Error graph for train (—) and verify set (—). Train was stopped after 325 epochs, corresponding to the verify error minimum.

ables, thus deserving a more advanced study to stress the comparison results. The sample population histologies in Ref. 24 were different from those in our series, making difficult a direct comparison between results of the two systems.

As regards validation of classification models, the size of our series (573 cases) was estimated to be large enough to allow the splitting of data into two independent train and validation samples, while the leave one out (LOO) method was used in both the two previously cited works. By this validation strategy, classifiers are developed by training on all but one of the lesions in the set, and testing is done on the lesion left out. The lesion ''left out'' is circulated over the entire set, and the score on the LOO tests is used to calculate the diagnostic rates. Since a reinstruction of the classifier is required for each lesion, the use of the LOO method on large samples can make the testing procedure very computationally intensive, especially when ANNs are involved, therefore, this method is generally used to estimate the performance of a classifier when only a sample of small size is available.

Our lesions were ranked according to their original (chronological) order, and cases in the validation set were consecutive, i.e., consecutively seen in a same day of the working week at the melanoma unit of our institution. We consider this partition-ordering criterion better than other arrangements, since instruction is needed only once, based on a previously acquired training sample, and validated over a consecutive case series, as it is more likely to occur in a real situation.

It could be objected that the choice of the early stopping technique in the network instruction is not suitable, since this method induces an indirect dependence between train and validation data. To overcome this problem, one additional dataset, called a test set is usually involved for final confirmation of the obtained results. Actually, our data did not appear to be significantly affected by the training strategies that could be adopted, even though it was based on the training error alone. For this purpose, network learning was also stopped, for example, when the train error function fell below 0.34 (a threshold error where an ''acceptable'' classification of at least 80% sensitivity and 70% specificity could be obtained) and/or when a steady slow error decreasing behavior (e.g., expressed by a minimum error improvement of 0.001 in a 35 epochs window) occurred. However, when allowed by the number of cases available, a three sets data partition is always preferable.[31]

As previously stated, a 400 cases training set was considered representative of the lesion population i.e., able to enclose the wide variability of our sample. In an attempt to further confirm our results, we also split this 400 cases training set into a new train and validation set, but with a relatively small size (50 lesions) assigned to the last, in order to have an as low as possible perturbation of the training set size and the same last 173 consecutive cases now used as a test set. This setup allowed us to train, with early stopping, a network whose output did not appear to significantly differ from the one described before, once train and verify original sets were compared to the new train and test sets, respectively. Nevertheless, owing to the reduced size of the verify set in this case, a risk of error in estimating the network performance still exists, and results from a larger study are necessary to confirm our data.

Based on our two-sets sample splitting, a comparison can be made between classifications obtained for both train and validation sets, to estimate how well a model is able to generalize, i.e., to correctly classify, new data. For the ANN, plots in Fig. 3(a) show good generalization characteristics, since ROC plots for train and verify are almost overlapped and do not significantly differ from each other ($P=0.78$,

TABLE III. Classification results obtained by the telespectrophotometric system using ANN and MDA.

	ANN			MDA			P (ANN vs MDA)	
Diagnoses	Train	Verify	P	Train	Verify	P	Train	Verify
Sensitivity	80%	78%	0.59	80%	76%	0.28		
Specificity	72%	76%	1.68	60%	57%	0.50		
Accuracy	74%	77%	1.55	65%	61%	0.36		
Contingency table[a]							<0.01	<0.01
AUC[b]	0.852	0.847	0.780	0.810	0.764	0.01	0.04	0.01

[a]2×2 contingency table between diagnoses by ANN and MDA, McNemar chi square test.
[b]P values evaluated according to Hanley and McNeil (Refs. 28 and 29).

Table III). In contrast, the difference observed between train and verify ROC curves is significantly less satisfactory [$P = 0.01$, Table III; Fig. 3(b)] when MDA is considered, thus indicating for MDA a generalization ability poorer than the one of ANN.

To simulate an actual application, the automated diagnosis was trained at a specific learning level of sensitivity (80%), and this was accomplished by selecting, through a proper choice of the discriminant threshold in the model, a specific point in the corresponding ROC diagram of the train set (open circles, Fig. 3). Actually, the 80% level is a rather arbitrary choice. However, in spite of the absence of a reference value accepted for clinical sensitivity, this 80% figure can be roughly considered as an indication of the diagnostic capability of an experienced clinician.[33,34] With this particular selection, results between train and verify diagnostic rates from MDA were not significantly different (Table III), apparently in contrast to the previous discussion, based on AUC comparisons, about the generalization performance of MDA. However, it should be considered that AUC is more representative of the curve shape than a single threshold choice. As an example, the difference between train and verify sensitivities becomes significant ($P = 0.02$) if a different (e.g., 0.75) sensitivity threshold is selected for learning, leading to a sensitivity in the verify set equal to 0.65 [squares, Fig. 3(b)]. An analysis on model stability can be only found in the paper by Wallace *et al.*[18] The authors report a classification of new cases (28 cases, 11 melanomas), based on an instruction operated with previous data (47 cases, 15 melanomas). Classification, however, was performed by inserting new lesions one at a time, in chronological order, each time reinstructing the network. Similarly, the same procedure was applied to MDA whose discriminant function was updated at each inclusion of a new case. The overall accuracy during the stepwise process of new lesions inclusion was monitored at each step by running the LOO method, including all the lesions available at the current step. Accuracy was shown to fluctuate around a mean value without a significant trend for ANN, whereas an average accuracy decrease of -0.53% per case entered was shown for MDA, thus further supporting, in agreement with our results, a reduced ability of MDA to generalize for new cases.

Our results show that factor analysis allowed a significant reduction in the number of the original variable set. An interesting consequence is that, for instance, mean reflectance or variegation spectra, obtained from 17 different images, can be to a great extent (at least in a proportion of 90%) described by only two or at most three different principal components. It is interesting to note that, for reflectance and variegation, factor loadings related to the first two respective components, share common values approximately near or shortly after the 600 nm wavelengths (see the plots in Fig. 2), i.e., in a region just after the one characterized by the influence of the oxy- and deoxyhaemoglobin absorption peaks between 542 and 577 nm.[35] By observation of such ''intersection points'' in all plots of Fig. 2, an interpretation about the nature of the different selected factors can be attempted. For example, $MR_{1,2}$ [Fig. 2(a)] could be roughly associated with the behavior of lesion mean reflectance mainly in the wavelength interval between 420 and 625 nm (MR_1), or between 625 and 1040 nm (MR_2). In the same way, $VI_{1,2,3}$ [Fig. 2(b)] roughly describe, respectively, lesion variegation in the three spectral intervals of (420–600), 600–850, and (850–1040) nm. By this interpretation, it follows that the selected variegation component in our model (VI_2) is mainly related to a spectral range, including those visible and infrared wavelengths with a reduced blood absorption contribution ($\lambda > 600$ nm) and where melanin distribution is more likely to be better appreciated. In fact, the melanin absorption spectrum presents a slowly decreasing trend toward infrared wavelengths in contrast to the peaked shape of blood absorption.[35]

The compactness factor (Cmp_2) can be roughly considered representative of the lesion compactness, or border fragmentation, between 461 and 578 nm [Fig. 2(c)]. In this range, the highest contrast between a lesion and surrounding skin was observed when evaluated as the ratio between skin and lesion reflectance over the whole dataset. In order to better assess the importance of the factors included in the ANN, input variables were ranked according to the error that would be obtained if each input was eliminated, divided by the network's own error (error ratio). After this operation, performed in the train set, the variables were ordered as A_1 (error ratio $= 1.11$), MR_2 (error ratio $= 1.08$), VI_2 (error ratio $= 1.05$), MR_1 (error ratio $= 1.02$) and Cmp_2 (error ratio $= 1.01$). These results would assign a major importance to the lesion dimension and, second, to reflectance evaluated at visible (mainly beyond 625 nm) and infrared wavelengths. However, since the correlation coefficient r between A_1 and the other factors was found to be always less than 0.3, the contribution of MR_2 in the model could have been underestimated with respect to A_1 due to the consistent correlation occurring between MR_2 and VI_2 ($r = -0.63$). After removing VI_2 from the model and reinstructing the ANN, the factors order changed as follows: MR_2 (error ratio $= 1.14$), A_1 (error ratio $= 1.09$), MR_1 (error ratio $= 1.03$), and Cmp_2 (error ratio $= 1.02$).

The above considerations lead to the conclusion that lesion dimension and reflectance at wavelengths beyond (approximately) 625 nm (red–infrared part of the electromagnetic spectrum), are the most important features employed by our instrument for classification, followed by variegation between 600 and 850 nm, lesion reflectance between 420 and 625 nm and, finally, lesion compactness, or border fragmentation. The importance of lesion reflectance in the infrared has been supported by other studies on an automated diagnosis based on multispectral imaging.[21,22,24,36,37] In addition, our findings are consistent with the clinical observation that a melanoma is generally darker (i.e., with lower reflectance) than nonmelanoma lesions [Fig. 1(a); Table II].

Variegation features also support the clinical finding that melanoma is more variegated than other lesions [Fig. 1(b); Table II]. Although the mean value of VI_1 was lower for melanomas (Table II), this feature, with the available spectrophotometric data, was not found to be significant, thus suggesting that the observed increase of CM variegation is

principally due to the distribution of pigments, including melanin, other than blood, whose influence on absorption is more important in the wavelength region below 600 nm, where VI_1 is mainly related. There is also agreement between clinical observation and the other shape describing features included in the ANN model, represented in Fig. 1(c), Fig. 1(e), and listed in Table II, i.e., compactness, and area, indicating a significantly higher border fragmentation and a greater dimension for melanoma in respect with other skin pigmented lesions. Data reported in literature for multispectral imaging measurements[22,23,37] match with these results about lesion shape descriptors.

The reliability of a classification should be assessed with reference to the main clinical and histological characteristics of the evaluated lesions. In particular, a diagnosis would be of value if it is able to distinguish melanoma at an early stage. It has been recently suggested[38] that comparisons between different automated systems for melanoma detection should be based on standardized criteria applied to the enrolled cases, whose description should include, for instance, lesion dimension and melanoma thickness. Of the 132 melanomas in our series, 91 (69%) were thin lesions (tumor thickness<1 mm or level I), and 14 (11%) were of small (maximum diameter$\leqslant 6$ mm) size. Our system was able to correctly classify 64 (70%) out of the 91 thin melanomas of which only two were small lesions. No one of the 41 thick lesions were missed by the ANN diagnosis. Of the 27 misdiagnosed melanomas, 11 (41%) were small lesions. By these last observations it can be argued that the system does not behave efficiently when facing small lesions. This topic appears to be important because small melanomas seem to represent a considerable subset of all CM.[39]

Since pixel dimension of our images is about 250–300 μm, the previous failure could be, to some extent, caused by the not adequate spatial resolution, preventing a correct evaluation of small lesions' inner structures or border details.

Dermatoscopy, or epiluminescence microscopy, has been introduced as an aid to clinical diagnosis and it is widely accepted that it improves diagnostic accuracy, especially for small lesions.[40,41]

The use of an improved acquisition system, especially with better spatial resolution characteristics, such as, for example, the one described in the work of Elbaum *et al.*[24] could allow the evaluation of dermatoscopic features, and provide useful additional diagnostic informations, especially using an ANN classifier.

Within the debate about the role that automated melanoma diagnosis devices should have from a practical clinical standpoint (e.g., for screening, to aid experts or to instruct nontrained clinicians),[38] the hope is that such enhanced multispectral systems, in the future, prove themselves to be of some help in the fight against melanoma.

V. CONCLUSIONS

This study was performed to retrospectively assess the potentials of a telespectrophotometric system in discriminating cutaneous melanoma from other skin lesions. Data reelaboration of spectroscopic images and lesion classification by a nonlinear classifier based on neural networks showed interesting diagnostic scores, providing with 80% sensitivity and 72% specificity in the train set; 78% sensitivity and 76% specificity in the validation set. AUC was found to be 0.852 and 0.847 for the train and verify set, respectively. A comparison was also done with a different classification method, based on standard linear MDA. This classification procedure was found to be significantly worse than the ANN approach, both in the diagnostic performance and in the model generalization capability.

The substantial lack of a common or standardized basis about the enrolled lesions strongly limits the comparison of our classification results with those from the literature. Although encouraging, our results should be confirmed by a larger study. In addition, the reported results are based on a nonoptimized device and evidence a lesser ability of the system in detecting small melanomas. This could partially be ascribed to the reduced spatial resolution of our instrument.

Within the debate about the role that automated melanoma diagnosis devices should have in clinical practice, the hope is that enhanced multispectral systems provided with nonlinear classifiers, such as those based on neural networks, prove themselves to be of substantial help in the fight against melanoma.

ACKNOWLEDGMENTS

This research was partially supported by the Associazione Italiana per la Ricerca sul Cancro (AIRC), and Lega Italiana per la Lotta Contro i Tumori.

[a]Author to whom correspondence should be addressed. Telephone: +39-2-23902192; fax: +39-2-23902124; electronic mail: stefano.tomatis@istitutotumori.mi.it

[1] "NIH Consensus Conference: Diagnosis and treatment of early melanoma," J. Am. Med. Assoc. **268**, 1314–1319 (1992).
[2] C. M. Balch, S. J. Soong, J. E. Gershenwald, J. F. Thompson, D. S. Reintgen, N. Cascinelli, M. Urist, K. M. McMasters, M. I. Ross, J. M. Kirkwood, M. B. Atkins, J. A. Thompson, D. G. Coit, D. Byrd, R. Desmond, Y. Zhang, P. Y. Liu, G. H. Lyman, and A. Morabito, "Prognostic factors analysis of 17,600 melanoma patients: Validation of the American Joint Committee on Cancer melanoma staging system," J. Clin. Oncol. **19**, 3622–3634 (2001).
[3] R. J. Friedman, D. S. Rigel, and A. W. Kopf, "Early detection of malignant melanoma: the role of the physician examination and self examination of the skin," Ca-Cancer J. Clin. **35**, 130–151 (1985).
[4] H. Pehamberger, A. Steiner, and K. Wolff, "*In vivo* epiluminescence microscopy of pigmented skin lesions: I. Pattern analysis of pigmented skin lesions," J. Am. Acad. Dermatol. **17**, 571–583 (1987).
[5] H. Pehamberger, M. Binder, A. Steiner, and K. Wolff, "*In vivo* epiluminescence microscopy: improvement of early diagnosis of melanoma," J. Invest. Dermatol. **100**, 356s–362s (1993).
[6] A. Green, N. Martin, J. Pfitzner, M. O'Rourke, and N. Knight, "Computer image analysis in the diagnosis of melanoma," J. Am. Acad. Dermatol. **31**, 958–964 (1994).
[7] N. Cascinelli, M. Ferrario, R. Bufalino, S. Zurrida, V. Galimberti, L. Mascheroni, C. Bartoli, and C. Clemente, "Results obtained by using a computerized image analysis system designed as an aid to diagnosis of cutaneous melanoma," Melanoma Res. **2**, 163–170 (1992).
[8] A. J. Sober and J. M. Burstein, "Computerized digital image analysis: an aid for melanoma diagnosis: Preliminary investigations and brief review," J. Dermatol. **21**, 885–890 (1994).
[9] L. Andreassi, R. Perotti, P. Rubegni, M. Burroni, G. Cevenini, M. Biagioli, P. Taddeucci, G. Dell'Eva, and P. Barbini, "Digital dermoscopy

analysis for the differentiation of atypical nevi and early melanoma,'' Arch. Dermatol. **135**, 1459–1465 (1999).
[10] T. Schindewolf, W. Stolz, R. Albert, W. Abmayr, and H. Harms, ''Comparison of classification rates for conventional and dermatoscopic images of malignant and benign melanocytic lesions using computerised colour image analysis,'' Eur. J. Dermatol. **3**, 299–303 (1993).
[11] M. Binder, H. Kittler, S. Dreiseitl, H. Ganster, K. Wolff, and H. Pehamberger, ''Computer-aided epiluminescence microscopy of pigmented skin lesions: the value of clinical data for the classification process,'' Melanoma Res. **10**, 556–561 (2000).
[12] M. Binder, H. Kittler, A. Seeber, A. Steiner, H. Pehamberger, and K. Wolff, ''Epiluminescence microscopy-based classification of pigmented skin lesions using computerized image analysis and an artificial neural network,'' Melanoma Res. **8**, 261–266 (1998).
[13] O. Debeir, C. Decaestecker, J. L. Pasteels, I. Salmon, R. Kiss, and P. Van Ham, ''Computed assisted analysis of epiluminescence microscopy images of pigmented skin lesions,'' Cytometry **37**, 255–266 (1999).
[14] P. Rubegni, A. Ferrari, G. Cevenini, D. Piccolo, M. Burroni, R. Perotti, K. Peris, P. Taddeucci, M. Biagioli, G. Dell'Eva, S. Chimenti, and L. Andreassi, ''Differentiation between pigmented Spitz naevus and melanoma by digital dermoscopy and stepwise logistic discriminant analysis,'' Melanoma Res. **11**, 37–44 (2001).
[15] G. R. Day and R. H. Barbour, ''Automated skin lesion screening—a new approach,'' Melanoma Res. **11**, 31–35 (2001).
[16] F. Ercal, A. Chawla, W. V. Stoecker, H. C. Lee, and R. H. Moss, ''Neural network diagnosis of malignant melanoma from color images,'' IEEE Trans. Biomed. Eng. **41**, 837–845 (1994).
[17] V. P. Wallace, D. C. Crawford, P. S. Mortimer, R. J. Ott, and J. C. Bamber, ''Spectrophotometric assessment of pigmented skin lesions: methods and feature selection for evaluation of diagnostic performance,'' Phys. Med. Biol. **45**, 735–751 (2000).
[18] V. P. Wallace, J. C. Bamber, D. C. Crawford, R. J. Ott, and P. S. Mortimer, ''Classification of reflectance spectra from pigmented skin lesions, a comparison of multivariate discriminant analysis and artificial neural networks,'' Phys. Med. Biol. **45**, 2859–2871 (2000).
[19] R. Marchesini, M. Brambilla, C. Clemente, M. Maniezzo, A. E. Sichirollo, A. Testori, D. R. Venturoli, and N. Cascinelli, ''*In vivo* spectrophotometric evaluation of neoplastic and nonneoplastic skin pigmented lesions—I. Reflectance measurements,'' Photochem. Photobiol. **53**, 77–84 (1991).
[20] R. Marchesini, N. Cascinelli, M. Brambilla, C. Clemente, L. Mascheroni, E. Pignoli, A. Testori, and D. R. Venturoli, ''*In vivo* spectrophotometric evaluation of neoplastic and nonneoplastic skin pigmented lesions—II. Discriminant analysis between nevus and melanoma,'' Photochem. Photobiol. **55**, 515–522 (1992).
[21] R. Marchesini, S. Tomatis, C. Bartoli, A. Bono, C. Clemente, C. Cupeta, I. Del Prato, E. Pignoli, A. E. Sichirollo, and N. Cascinelli, ''*In vivo* spectrophotometric evaluation of neoplastic and nonneoplastic skin pigmented lesions—III. CCD camera-based reflectance imaging,'' Photochem. Photobiol. **62**, 151–154 (1995).
[22] S. Tomatis, C. Bartoli, A. Bono, N. Cascinelli, C. Clemente, and R. Marchesini, ''Spectrophotometric imaging of cutaneous pigmented lesions: Discriminant analysis, optical properties and histological characteristics,'' J. Photochem. Photobiol., B **42**, 32–39 (1998).
[23] B. Farina, C. Bartoli, A. Bono, A. Colombo, M. Lualdi, G. Tragni, and R. Marchesini, ''Multispectral imaging approach in the diagnosis of cutaneous melanoma: Potentiality and limits,'' Phys. Med. Biol. **45**, 1243–1254 (2000).
[24] M. Elbaum, A. W. Kopf, H. S. Rabinovitz, R. G. Langley, H. Kamino, M. C. Mihm, Jr., A. J. Sober, G. L. Peck, A. Bogdan, D. Gutkowicz-Krusin, M. Greenebaum, S. Keem, M. Oliviero, and S. Wang, ''Automatic differentiation of melanoma from melanocytic nevi with multispectral digital dermoscopy: A feasibility study,'' J. Am. Acad. Dermatol. **44**, 208–218 (2001).
[25] S. Seidenari, G. Pellacani, and A. Giannetti, ''Digital videomicroscopy and image analysis with automatic classification for detection of thin melanoms,'' Melanoma Res. **9**, 163–171 (1999).
[26] D. E. Elder and G. F. Murphy, ''Melanocytic tumurs of the skin,'' Melanoma Res. in: *Atlas of Tumor Pathology* Armed Forces Institute of Pathology (Washington, DC, 1991), p. 110-9.
[27] M. L. Thompson and W. Zucchini, ''On the statistical analysis of ROC curves,'' Stat. Med. **8**, 1277–1290 (1989).
[28] J. A. Hanley and B. J. McNeil, ''The meaning and use of the area under a receiver operating Characteristic (ROC) curve,'' Radiology **143**, 29–36 (1982).
[29] J. A. Hanley and B. J. McNeil, ''A method of comparing the areas under Receiver Operating Characteristic curves derived from the same cases,'' Radiology **184**, 839–843 (1983).
[30] D. G. Kleinbaum, L. L. Kupper, and K. E. Muller, *Applied Regression Analysis and Other Multivariable Methods*, 2nd ed. (PSW-Kent, Boston, 1998).
[31] C. M. Bishop, *Neural Networks for Pattern Recognition* (Oxford University Press, Oxford, 1995).
[32] R. C. Gonzalez and R. E. Woods, *Digital Image Processing* (Addison Wesley, Reading, 1992).
[33] C. M. Grin *et al.*, ''Accuracy in the clinical diagnosis of malignant melanoma,'' Arch. Dermatol. **126**, 763–766 (1990).
[34] I. H. Wolf, J. Smolle, H. P. Soyer, and H. Kerl, ''Sensitivity in the clinical diagnosis of malignant melanoma,'' Melanoma Res. **8**, 425–429 (1998).
[35] J. W. Ashley and M. J. C. Van Gemert, *Optical Thermal Response of Laser Irradiated Tissue* (Plenum, New York, 1995).
[36] A. Bono, S. Tomatis, C. Bartoli, N. Cascinelli, C. Clemente, C. Cupeta, and R. Marchesini, ''The invisible colours of melanoma. A telespectrophotometric diagnostic approach on pigmented skin lesions,'' Eur. J. Cancer **32A**, 727–729 (1996).
[37] A. Bono, S. Tomatis, C. Bartoli, G. Tragni, G. Radaelli, A. Maurichi, and R. Marchesini, ''The ABCD system of melanoma detection. A spectrophotometric analysis of the asymmetry, border, color, and dimension,'' Cancer (N.Y.) **85**, 72–77 (1999).
[38] R. Marchesini, A. Bono, C. Bartoli, M. Lualdi, S. Tomatis, and N. Cascinelli, ''Diagnosis of melanoma based on optical imaging and computer analysis: Questions and answers,'' Melanoma Res. **12**, 279–286 (2002).
[39] A. Bono, C. Bartoli, D. Moglia, A. Maurichi, T. Camerini, G. Grassi, G. Tragni, and N. Cascinelli, ''Small melanomas: A clinical study on 270 consecutive cases of cutaneous melanoma,'' Melanoma Res. **9**, 583–586 (1999).
[40] A. Steiner, H. Pehamberger, and K. Wolff, ''*In vivo* epiluminescence microscopy of pigmented skin lesions: I. Diagnosis of small pigmented skin lesions and early detection of malignant melanoma,'' J. Am. Acad. Dermatol. **17**, 584–591 (1987).
[41] M. Binder, M. Schwarz, A. Winkler, A. Steiner, A. Kaider, K. Wolff, and H. Pehamberger, ''Epiluminescence microscopy. A useful tool for the diagnosis of pigmented lesions for formally trained dermatologists,'' Arch. Dermatol. **131**, 286–291 (1995).