# Exploiting Classifier Combination for Early Melanoma Diagnosis Support

E. Blanzieri, C. Eccher, S. Forti, and A. Sboner

ITC-irst
Via Sommarive, 18, 38100 – Trento, Italy
{blanzier,sboner}@itc.it

**Abstract.** Melanoma is the most dangerous skin cancer and early diagnosis is the main factor for its successful treatment. Experienced dermatologists with specific training make the diagnosis by clinical inspection and they reach 80% level of both sensitivity and specificity. In this paper, we present a multi-classifiers system for supporting the early diagnosis of melanoma. The system acquires a digital image of the skin lesion and extracts a set of geometric and colorimetric features. The diagnosis is performed on the vector of features by integrating with a voting schema the diagnostic outputs of three different classifiers: discriminant analysis, k-nearest neighbor and decision tree. The system is build and validated on a set of 152 skin images acquired via D-ELM. The results are comparable or better of the diagnostic response of a group of expert dermatologists.

## 1 Introduction

Combination and integration of different models has been a very popular area in recent years and techniques such *bagging* and *boosting* have gained a lot of attention (see for references Chan et al. 1999). In such schemata training of the same algorithm on different data generates different models. However, as noted by Merz (1999), the combination of different kinds of algorithms proved to be effective in increasing accuracy.

In real-world applications, namely when a user is involved in the development (following the definition of Saitta and Neri, 1998), accuracy is not the main issue. Comprehension and readability of the results are particularly relevant when the final user is a skilled physician. Presenting their experience in a medical application, Morik et al. (1999) emphasized the relevance of understandability and embeddedness of the learning component into the overall application system. Finally, in a medical diagnosis application, sensitivity and specificity are far more relevant of accuracy, as noted by Kukar et al. (1999) cost-sensitive algorithms should be used.

In this work we claim that combination of different algorithms can be also useful in order to solve problems that arises in a real application in a sensitive domain such as cancer diagnosis. In particular we present MEDS (Melanoma Diagnosis System) a system for early melanoma diagnosis support developed by ITC-IRST in collaboration with the Department of Dermatology of Santa Chiara Hospital, Trento.

R. López de Mántaras, E. Plaza (Eds.): ECML 2000, LNAI 1810, pp. 55-62, 2000.

The main goal of MEDS is to provide support to a physician for early diagnosis of melanoma, the most dangerous skin cancer. Although it is diagnosed in about 5% of the overall skin cancers, melanoma is responsible of the 91% of the deaths and its incidence in Europe is increasing of 3%-5% yearly. The early diagnosis of melanoma is the principal factor for the prognosis of this disease. The diagnosis is difficult and requires a well-trained physician, because the early lesion looks like a benign one. Several studies have shown that the diagnostic accuracy of a specialist is about 69% for early melanomas, and it reduces to 12% for non-specialists (Clemente et al. 1998). One of the digital techniques that had considerable success in clinical practice is *digital epi-luminescence microscopy (D-ELM*) (see for a review Zsolt, 1997). It allows the determination of several morphological and structural characteristics of skin lesions without remove them. Several automatic systems were proposed for the early diagnosis of melanoma (Shindewolf 1992, Green 1994, Ercal 1994, Binder 1994, Takiwaki 1995, Seidenari 1998) and recently D-ELM has been exploited by Bischof et al. (1999).

MEDS make the diagnosis on a D-ELM image, so it faces the problem of processing the image for feature extraction. A second problem is that collecting data on melanoma is difficult: melanoma cases are not common and characteristics of the D-ELM images depend on the type of acquisition system chosen. Therefore, the application has to been built with small data sets and unbalanced classes. A third and major problem arises when loss functions are considered. Sensitivity shows the ability of the system to recognize the malign lesion, while specificity describes how the system recognizes the benign lesion. Depending on the application (screening by a general practitioner or diagnosis support of an expert dermatologist) very different levels of sensitivity and specificity are required and the system should provide a tuning mechanism. Finally, another critical issue in order to build a usable system is gaining the trust of the user, comprehensibility of the results are one of the major issue.

MEDS elaborate D-ELM images extracting features that could be meaningful for the expert dermatologist, following the so-called ABCD Rule (Nachbar 1994) in order to improve comprehensibility. The features are the input of three different classifiers, namely Discriminant Analysis, Decision Tree and k-Nearest Neighbor, which MEDS integrates by means of voting schemata. The classifiers permit different explanations of the results. The combination improves the performance in terms of sensitivity and specificity given the small number of data. Finally, the simple voting mechanism is clear and well understood by the user and it is possible to use the voting schema as a tuning parameter comprehensible by the expert.

The paper is organized as follows: section 2 describes the system, technical and clinical validations are presented in section 3 and 4 respectively, section 5 is devoted to related works and finally, conclusions are drawn in section 6.

## 2 System Description

MEDS architecture has three main components: the D-ELM Image Acquisition component whose goal is to acquire the image of the pigmented skin lesion, the Image Processing component that elaborates the digital image producing the vector of

features, the Multi-classifier that applies and combines Discriminant Analysis, Decision Tree and k-Nearest Neighbor. The system functional architecture is shown in Fig.1. Physically, the three main components run on different machines and we transfer information by means of files. We are working to integrate the components in a client/server application. The image processing and the multi-classifier components will reside on a centralized server, while the D-ELM image acquisition component will be deployed to several clients.

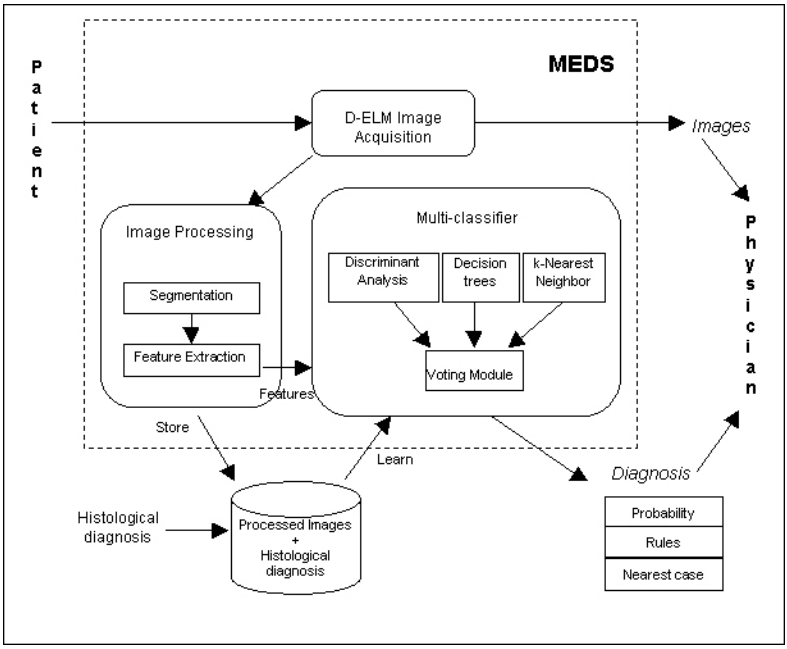


**Fig. 1.** Overview of the overall system

For the D-ELM Image Acquisition, we used a stereomicroscope Leica WILD M-650, with a color camera SONY 3CCD DXC-930P. The camera is linked with an acquisition board AT-Vista Videographics, which allows digitizing the analog image of the microscope. The software for the image acquisition is DBDERMO MIPS (Dell'Eva/Burroni Studio, Florence/Siena, Italy). For image processing we used the morphometer Leica Q570. The colored image is usually divided in three gray images, which represent the red, the green and the blue component respectively (RGB color space). During the image processing, the image is converted from the usual RGB colors space to the hue, saturation and value (HSV colors space). The HSV colors space is particularly useful because it reflects the human perception of colors.

The Image Processing component performs two functions: segmentation and feature extraction. The purpose of segmentation is defining the border of the lesion, separating it from the rest of the skin. We exploit the HSV images, because, in this case, the normal skin and the lesion present marked differences, especially for the hue and the saturation images. The feature extraction module produces numerical features: geometric and colorimetric ones. The geometric parameters measure the dimension of the lesion (area and perimeter) and its symmetric characteristics (roundness, aspect

ratio and full ratio). The colorimetric features quantitatively reflect concepts as the presence and symmetric or asymmetric distribution of the colors, the granularity of the colors, the irregularity of the pigmentation on the border of the lesion, etc. The extracted features reflect the ABCD rule used by the dermatologist to diagnose a skin lesion (Nachbar 1994). In fact, with this rule, the physician evaluates the Area of the lesion, the irregularity of the Border, the presence and the distribution of the Color and the presence of Differential structures.

The Multi-classifier module produces a diagnosis based on the features extracted from the D-ELM images. It used three kinds of classifiers: discriminant analysis, decision tree and k-nearest neighbor, and combine them by means of different voting schemata. The classifiers were chosen in order to permits different explanation of the results (probability, rules and a nearest similar cases respectively).

In this application, the relevant gain functions for the classifiers are sensitivity, defined as TP / (TP + FN), and specificity, defined as TN / (TN +FP). TP, TN, FP, FN are the number of melanomas correctly classified (True Positives), the number of nevi correctly classified (True Negatives), the number of nevi classified as melanomas (False Positives) and the number of melanomas classified as nevi (False Negatives), respectively. Sensitivity depends on FN and it is the most critical parameter.

In order to improve sensitivity we altered the prior probabilities of the classes as described in Kukar et al. (1999). We adopted different strategies for the three classifiers. The prior probabilities for the linear discriminant analysis were considered equal for each class. Discriminant analysis was performed via a multivariate analysis on features selected by means of a univariate analysis. For the decision tree, we adopted C4.5 and we performed a pre-processing on the data increasing the weight of malignant melanomas as described in Breiman (1984) and reported by Kukar et al. (1999). Finally, we adopted the Euclidean metric for the k-nearest neighbor (*k-NN*). To improve sensitivity, we also used a particular form of the nearest neighbor algorithm (*k-NN-Uni*), whose output is "melanoma" if at least one of the k-nearest-neighbors is a melanoma.

Comprehensibility is preserved by combining discriminant analysis, decision tree and k-nearest-neighbor with simple rules. If the combination involved k-NN-Uni, we adopted the majority rule (schema "2/3"). Otherwise, we required a total agreement on benign lesions for classify the new case as a mole (schema "1/3").

## 3 System Validation

We analyzed 152 skin images acquired by D-ELM at the Department of Dermatology of Santa Chiara Hospital, Trento. The images where classified histologically by the pathologist as 42 melanomas and 110 nevi. Breslow thickness was evaluated for 42 malign lesions. This parameter is linked to the prognosis of the disease and it can be determined only by the histological analysis. The average Breslow thickness for our lesions is $1.0 \pm 0.7$ mm, and the 90% of them are thinner than 1.70 mm. This fact confirms the earliness of the involved melanomas. We evaluate sensitivity and specificity using *10-fold cross-validation* and performed experiments with the single classifiers and their combinations.

Table 1 reports the results for the single classifiers, both sensitivity and specificity with their standard deviation are shown. Discriminant analysis, decision tree and k-NN have poor sensitivity values, while specificity range from 0.83 to 0.97. This fact is due to the major number of benign lesions relative to the melanomas. Instead, for the k-NN-Uni sensitivity is very high (from 0.69 to 1.00), while specificity significantly decreases as k become greater (from 0.75 to 0.20). This fact agrees with the adopted modified decision rule for the k-NN-Uni, which strongly promotes sensitivity.

Table 2 reports the results for the 3-Classifiers systems. The combination of three classifiers has a general improvement of sensitivity without a strong decrease of specificity. Comparison of the 3-classifiers with the single systems (in particular for the voting schema "1/3") shows a statistical improvement of sensitivity, while specificity shows no statistical differences.

**Table 1.** Single classifier systems – For each classifier, sensitivity and specificity, with the standard deviation, are shown

| *CLASSIFIER* | *Sens.* | *SD* | *Spec.* | *SD* | *CLASSIFIER* | *Sens.* | *SD* | *Spec.* | *SD* |
|---|---|---|---|---|---|---|---|---|---|
| *Discriminant analysis* | | | | | *k-NN-Uni* | | | | |
| DiscrAn | 0.65 | 0.30 | 0.83 | 0.11 | 2-Uni | 0.69 | 0.30 | 0.75 | 0.10 |
| *Decision trees* | | | | | 3-Uni | 0.81 | 0.19 | 0.61 | 0.08 |
| C4.5 | 0.64 | 0.28 | 0.84 | 0.05 | 4-Uni | 0.86 | 0.19 | 0.53 | 0.16 |
| *k-Nearest-Neighbor* | | | | | 5-Uni | 0.88 | 0.16 | 0.45 | 0.17 |
| 1-NN | 0.68 | 0.30 | 0.90 | 0.10 | 6-Uni | 0.99 | 0.05 | 0.35 | 0.16 |
| 3-NN | 0.49 | 0.36 | 0.91 | 0.07 | 7-Uni | 1.00 | 0.00 | 0.29 | 0.18 |
| 5-NN | 0.46 | 0.34 | 0.97 | 0.06 | 8-Uni | 1.00 | 0.00 | 0.28 | 0.18 |
| 7-NN | 0.35 | 0.23 | 0.97 | 0.04 | 9-Uni | 1.00 | 0.00 | 0.20 | 0.18 |
| 9-NN | 0.41 | 0.25 | 0.96 | 0.04 | | | | | |

**Table 2.** 3-Classifiers systems – The columns "Comparison [vs. single] {+, =, -}" represent the comparison of the combined classifier with each single component: combined classifier better than the single (+), combined worst than single (-) and no significant differences (=). The statistical analysis is based on the paired Wilcoxon test with a p-value of 0.05

| | | *Sensitivity* | | | | | *Specificity* | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | *Comparison* | | | | | *Comparison* | | |
| *Voting* | *Combined* | | | *[vs. single]* | | | | | *[vs. single]* | | |
| *Schema* | *Classifiers* | *Value* | *SD* | + | = | - | *Value* | *SD* | + | = | - |
| | DiscrAn C4.5 1-NN | 0.86 | 0.32 | ααα | | | 0.64 | 0.11 | | ααα | |
| | DiscrAn C4.5 3-NN | 0.84 | 0.32 | ααα | | | 0.65 | 0.11 | | ααα | |
| *1/3* | DiscrAn C4.5 5-NN | 0.84 | 0.32 | ααα | | | 0.68 | 0.12 | | ααα | |
| | DiscrAn C4.5 7-NN | 0.84 | 0.32 | ααα | | | 0.68 | 0.10 | | ααα | |
| | DiscrAn C4.5 9-NN | 0.85 | 0.32 | ααα | | | 0.68 | 0.11 | | ααα | |
| | DiscrAn C4.5 2-Uni | 0.75 | 0.31 | | ααα | | 0.89 | 0.11 | | ααα | |
| | DiscrAn C4.5 3-Uni | 0.75 | 0.31 | | ααα | | 0.84 | 0.09 | | ααα | |
| | DiscrAn C4.5 4-Uni | 0.77 | 0.31 | | ααα | | 0.81 | 0.11 | | ααα | |
| *2/3* | DiscrAn C4.5 5-Uni | 0.77 | 0.31 | | ααα | | 0.81 | 0.11 | α | αα | |
| | DiscrAn C4.5 6-Uni | 0.82 | 0.31 | αα | | α | 0.78 | 0.10 | α | αα | |
| | DiscrAn C4.5 7-Uni | 0.84 | 0.32 | αα | α | | 0.75 | 0.09 | α | αα | |
| | DiscrAn C4.5 8-Uni | 0.84 | 0.32 | αα | α | | 0.75 | 0.09 | α | αα | |
| | DiscrAn C4.5 9-Uni | 0.84 | 0.32 | αα | α | | 0.71 | 0.12 | α | αα | |

## 4 Clinical Validation

We involved in this study a group of eight dermatologists in order to compare the performances of the system with those of clinicians. Part of the group of the dermatologist was experienced in digital epiluminescence. The average sensitivity and specificity of the dermatologist were 0.83 and 0.66 respectively and the diagnosis were performed only on a video device, reproducing a teledermatology setting.

Table 3 and Table 4 show the results obtained comparing the dermatologists to the classifiers: for single classifiers and 3-Classifiers systems respectively. Each table shows the number of physicians (among the eight dermatologists involved in this study) that performed better, equal or worse than the classifier. We used the paired Wilcoxon test to measure statistical significance. The p-value considered for discriminate the results was 0.05. Table 3 shows that the single classifier systems do not reach useful performances for the early diagnosis of melanoma. When they perform better for one parameter, for example sensitivity, the physicians perform better for the other. Table 4 shows that the 3-Classifiers systems perform as well as the eight dermatologists for what concern sensitivity, while they perform better for what concern specificity for, at least, one physician. Moreover, these systems never have poor performances compared with each dermatologist for both sensitivity and specificity, and the mis-classified melanomas are different from those of the physicians. This fact confirms the possibility to use these systems as a diagnosis support, also for the well-trained dermatologist.

**Table 3.** Dermatologists vs. single classifier systems. (+ means that the classifier is better than the physician, = means that there is non statistical difference between the classifiers and the physician and – means that the physician is better than the classifier.)

| *Dermatologists* | *Sens.* | | | *Spec.* | | | *Dermatologists* | *Sens.* | | | *Spec.* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *vs.* | + | = | - | + | = | - | *vs.* | + | = | - | + | = | - |
| DiscrAn | 0 | 4 | 4 | 5 | 3 | 0 | 2-Uni | 0 | 8 | 0 | 3 | 5 | 0 |
| C4.5 | 0 | 7 | 1 | 6 | 2 | 0 | 3-Uni | 0 | 8 | 0 | 2 | 4 | 2 |
| 1-NN | 0 | 8 | 0 | 7 | 1 | 0 | 4-Uni | 0 | 8 | 0 | 0 | 4 | 4 |
| 3-NN | 0 | 3 | 5 | 7 | 1 | 0 | 5-Uni | 0 | 8 | 0 | 0 | 3 | 5 |
| 5-NN | 0 | 2 | 6 | 8 | 0 | 0 | 6-Uni | 3 | 5 | 0 | 0 | 1 | 7 |
| 7-NN | 0 | 1 | 7 | 8 | 0 | 0 | 7-Uni | 3 | 5 | 0 | 0 | 1 | 7 |
| 9-NN | 0 | 1 | 7 | 8 | 0 | 0 | 8-Uni | 3 | 5 | 0 | 0 | 1 | 7 |
| | | | | | | | 9-Uni | 3 | 5 | 0 | 0 | 0 | 8 |

**Table 4.** Dermatologists vs. 3-classifiers systems

| *Dermatologists vs.* | | | *Sens.* | | | *Spec.* | | | *Dermatologists vs.* | | | *Sens.* | | | *Spec.* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Cl. 1* | *Cl. 2* | *Cl. 3* | + | = | - | + | = | - | *Cl. 1* | *Cl. 2* | *Cl. 3* | + | = | - | + | = | - |
| DiscrAn | C4.5 | 1-NN | 0 | 8 | 0 | 2 | 6 | 0 | DiscrAn | C4.5 | 2-Uni | 0 | 8 | 0 | 1 | 7 | 0 |
| DiscrAn | C4.5 | 3-NN | 0 | 8 | 0 | 2 | 6 | 0 | DiscrAn | C4.5 | 3-Uni | 0 | 8 | 0 | 1 | 7 | 0 |
| DiscrAn | C4.5 | 5-NN | 0 | 8 | 0 | 2 | 6 | 0 | DiscrAn | C4.5 | 4-Uni | 0 | 8 | 0 | 2 | 6 | 0 |
| DiscrAn | C4.5 | 7-NN | 0 | 8 | 0 | 2 | 6 | 0 | DiscrAn | C4.5 | 5-Uni | 0 | 8 | 0 | 2 | 6 | 0 |
| DiscrAn | C4.5 | 9-NN | 0 | 8 | 0 | 2 | 6 | 0 | DiscrAn | C4.5 | 6-Uni | 0 | 8 | 0 | 2 | 6 | 0 |
| | | | | | | | | | DiscrAn | C4.5 | 7-Uni | 0 | 8 | 0 | 2 | 6 | 0 |
| | | | | | | | | | DiscrAn | C4.5 | 8-Uni | 0 | 8 | 0 | 2 | 6 | 0 |
| | | | | | | | | | DiscrAn | C4.5 | 9-Uni | 0 | 8 | 0 | 2 | 6 | 0 |

## 5 Related Works

In the recent years, several systems for the diagnosis of melanoma were proposed. Shindewolf et al. (1992) used a decision tree to classify digital images of skin lesions. The images were acquired by a photo-camera, and then scanned by a color TV camera and digitized. They showed results based on a resubstitution evaluation technique. Green et al. (1994) used a discriminant analysis as classification system. In this case, the obtained results seem to refer to all the cases, without any evaluation of the prediction performance. Ercal et al. (1994) applied an artificial neural network on feature extracted by photographic images with different films. As the color is the most significant parameter for the diagnosis of early melanoma it is difficult to compare images from different films. They obtained a sensitivity of 0.86 and a specificity of 0.85 as best result. Binder et al. (1994) applied an artificial neural network to classify dermatological images. They used as inputs of the neural network the ABCDE parameters that were predefined by a physician. This is a semi-automated method, which strongly relies upon the dermatologist. Takiwaki et al. (1995) used a decision tree to discriminate among the lesions. The reported results show only the generated tree, describing the most significant features.

Some recent works are more related to MEDS. Seidenari et al. (1998) applied a discriminant analysis describing the most significant features but it is not clear the evaluation procedure they adopted. Finally, Bischof et al. (1999) used a decision tree to classify images from a D-ELM system. Their results show a cross-validated sensitivity of 0.89 and a specificity of 0.80. Using their methodology, namely training only a decision tree, we did not succeed in reaching such good results (see in Table 1), probably for the different characteristics of the data.

## 6 Conclusions

In this paper, we have presented MEDS, a system for early diagnosis support of melanoma. MEDS uses a combination of classifiers for solving some of the typical problems that are present in melanoma diagnosis applications. By combination of standard learning algorithms it is possible to improve sensitivity and specificity for reaching the performance of skilled dermatologists solving the problems related to small data sets and unbalanced classes. Different combinations can be useful in order to select the level of sensitivity or specificity required by applications like screening or support of expert dermatologists. The algorithms selected are able to suggest an explanation for the diagnosis: a probability in case of discriminant analysis, a rule in case of decision tree and a similar case using the k-nearest neighbor. Comprehensibility is improved by extracting features related to the clinical practice, and in particular to the digital epiluminescence methodology.

MEDS is integrated with D-ELM that represents the state of the art of clinical practice in pigmented skin lesion diagnosis, and we plan to test the system in a clinical setting.

**Acknowledgments**

We thank M. Cristofolini and P. Bauer of the Department of Dermatology of the S.Chiara Hospital – Trento, for their significant collaboration in this study, and F. Ricci for participating to the initial phase of the project. We thank the reviewers for insightful comments, which helped to improve the paper. This work was partially supported by the Lega Italiana per la lotta contro i tumori – Sezione Trentina

## References

1. Binder M., Steiner A., Scharz M., Knollmayer S., Wolff K., Pehamberger H., Application of an artificial neural network in epiluminescence microscopy pattern analysis of pigmented skin lesions: a pilot study, British Journal of Dermatology (1994) 130, 460-465.
2. Bischof L., Talbot H., Breen E., Lovell D., Chan D., Stone G., Menzies S.W., Gutenev A., Caffin R., An Automated Melanoma Diagnosis System, in Pham B., Braun M., Maeder A. J. and Eckert M.P., editors, New Approaches in Medical Image Analysis, volume 3474, pp. 130-41, SPIE (1999).
3. Breiman L., Friedman J.H., Ohlsen R.A., Stone C.J., Classification and Regression Trees, Monterey, CA, Wadsworth (1984).
4. Chan P.K, Stolfo S.J., Wolpert D., (Eds) Machine Learning, Special Issue on Integrating Multiple Models for Improving and Scaling Machine Learning Algorithms. 36 (1999).
5. Clemente C.G., Mihm M.C., Cainelli T., Cristofolini M., Cascinelli N., Melanoma e Nevi, Effetti, Milano (1998).
6. Ercal F., Chawla A., Stoecker W.V., Lee H., Moss R.H., Neural network diagnosis of malignant melanoma from color images, IEEE Transaction on Biomedical Engineering, vol. 4, n. 9, Sept. 1994.
7. Green A., Martin N., Pfitzner J., O'Rourke M., Knight N., Computer image analysis in the diagnosis of melanoma, J Am Acad Dermatol (1994), 31:958-64.
8. Kukar M., Kononenko I., Groselj C., Krali K., Fettic J., Analysing and improving the diagnosis of ischaemic heart desease with machine learning, Artif Intell Med, vol. 16 (1999) 25-50.
9. Merz C., Using Correspondence Analysis to Combine Classifiers, Machine Learning 36, 33-58, (1999).
10. Morik K., Brockhausen P., Joachims T., Combining statistical learning with a knowledge-based approach – A case study in intensive case monitoring. ICML99, 268-277, (1999).
11. Nachbar F., Stolz. W., Merkle T., Cognetta A.B., Vogt T., Landthaler M., Bilek P., Braun-Falco O., Plewig G., *The ABCD rule* of Dermatoscopy: high prospective value in the diagnosis of doubtful melanocytic skin lesions, J Am Acad Dermatol (1994) 30:551-9.
12. Saitta L., Neri F., Learning in the "Real-World", Machine Learning 30,133-136 (1998).
13. Seidenari S., Pellacani G., Pepe P., Digital videomicroscopy improves diagnostic accuracy for melanoma, J Am Acad Dermatol (1998) 39:175-81.
14. Shindenwolf T., Stolz W., Albert R., Abmayr W., Harms H., Classification of Melanocytic Lesions with Colour and Texture Analyses Using Digital Image Processing, The international Academy of Cytology, (1992).
15. Takiwaki H., Shirai S., Watanabe Y., Nakagawa K., Arase S., A rudimentary system for automatic discrimination among basic skin lesions on the basis of color analysis of video images, J Am Acad Dermatol (1995) 32:600-3.
16. Zsolt B.A. Dermoscopy (Epiluminescence Microscopy) of pigmented skin lesions, Dermatologic clinics, Vol. 15, n.1 Jan 1997.