The Third International Symposium on Optimization and Systems Biology (OSB'09)
Zhangjiajie, China, September 20–22, 2009

New Approaches for Computer-Assisted Skin Cancer Diagnosis

M. J. Ogorzałek[1] G. Surówka[1] L. Nowak[1]
C. Merkwirth[1]

[1]Department of Information Technologies, Faculty of Physics, Astronomy and Applied Computer Science, Jagiellonian University, ul. Reymonta 4, 30-059 Kraków, Poland

Abstract With the wide-spread availability of advanced digital cameras dermoscopy has become nowadays a standard technique to help the doctors taking diagnosis for many types of skin lesions. Further, computer-assisted techniques and image processing methods can be used for image filtering and for feature extraction and pattern recognition in the selected images. Apart from standard approaches based on geometrical features and color/pattern analysis we propose to enhance the computer-aided diagnostic tools by adding non-standard image decompositions and applying classification techniques based on statistical learning and model ensembling. Ensembles of classifiers based on the extended feature set show improved performance figures suggesting that the proposed approach could be used as powerful tool assisting medical diagnosis.

1 Introduction

As in recent years there has been a significant raise in the number of melanoma cases recorded world-wide, early diagnosis and efficient tools helping the dermetologists become a necessity. The dermoscope or epiluminescence microscope [1] facilitates the diagnosis process. Image atlas of reference pictures are widely available [3], [4]. The Second Consensus Meeting on Dermoscopy has been held in 2000 and its main conclusions were that four algorithms: pattern analysis , ABCD rule, Menzies scoring method and 7-point check list are good ways of evaluation of skin lesions using dermoscopy. All four methods share some common concepts and allow for selection of specific features possible to be done with the aid of computer.

The ABCD [14] allows for computation of so-called TDS factor (Total Dermoscopy Score). Four features of the image are taken into account:

1. **A**symmetry – The dermoscopic image is divided by two perpendicular axes positioned to produce the lowest possible asymmetry score. If the image shows asymmetric properties with respect to both axes with regard to colors and differential structures, the asymmetry score is 2. If there is asymmetry on one axis the score is 1. If asymmetry is absent the score is 0.
2. **B**order – The images of the lesions are divided into eighths and a sharp, abrupt cut-off of pigment pattern at the periphery within one eighth has a score 1. A gradual, indistinct cut-off within one eighth has a score of 0.

3. **C**olor – Six different colors: white, red, light-brown, dark-brown, blue-gray, and black, are counted in the color score. White is only counted if the area is lighter than the adjacent skin. Cancerous skin changes are usually characterized by three or more colors. About 40% of melanomas have five or six colors are present.
4. **D**ifferential structure – Stolz proposed five types of these: pigment network, structureless or homogeneous areas, streaks, dots, and globules. The more structures are present in the picture, the higher the probability of the lesion being a cancer.

Once the ABCD values are evaluated for the image under consideration one can compute the TDS factor: $TDS = A*1.3+B*0.1+C*0.5+D*0.5$
$TDS < 4.75$ gives indication of a benign lesion, $4.75 < TDS < 5.45$ is nonconclusive while $TDS > 5.45$ gives strong indication that the lesion is cancerous.

As an alternative Menzies [11] proposed a scoring method based on inspection of the lesion image. Asymmetry of pattern, more then one color and presence of one to nine positive specific features are simply adding one point to the score. The positive features are blue-and-white vail,multiple brown dots, pseudopods, radial streaming, scar-like depigmentation, peripheral black dots/globules, multiple colors (five or six), multiple blue/gray dots, broad pigment network. Most of the features proposed by Menzies are present also in the ABCD rule and serve for TDS calculation. The third method used by dermatologists is so-called seven point checklist. Here also most of the descriptive features are similar to those present in the ABCD scale.

Evaluating the approaches Argenziano [1] and Johr [10] made a comparisons of different methods finding that the seven point checklist had sensitivity in the range of 95% and specificity of 75% while for pattern analysis these number were 91% and 90% respectively going down to 85% and 66% respectively for the ABCD rule.

Collaboration of dermatologists, computer scientists and image processing specialists has led to significant automation of analysis of dermoscopic images and improvement in their classification [7], [8] [13]. Several approaches to tele-diagnostic systems can be also mentioned when images to be evaluated and classified are sent eg. via internet [9]. There are several areas in dermoscopic image analysis where specific approaches, algorithms and methods from the image processing and computational intelligence toolkit could be used [2], [5], [13]. In this paper we propose to combine several methods from the image processing and computational intelligence areas providing the medical doctors with a computer-assisted diagnosis toolkit.

2 Basic image processing

Before any other type of analysis of dermoscopic images could be carried out the acquired images have to be pre-processed. Pre-processing will include image normalization (size and color or gray shades), filtering and artefact removal.

As the next step the geometric features such as borders can be determined. The shape of the lesion can be found automatically by executing properly selected binarization of the image under consideration. Figure 1 shows typical problems associated with the choice of the threshold level. Misinterpretations are possible especially when the image exhibits multiple colorings or multiple gray shades. In such cases one can obtain very different shapes, borders and areas of the lesion depending on the selected threshold value.

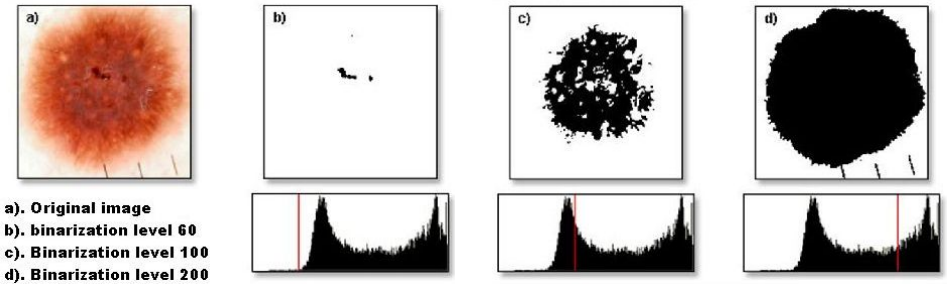
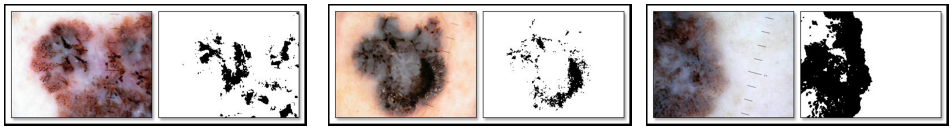


Figure 1: Varying the threshold value one can single out from an image different structural components. For the binarization threshold equal 60 one can see the black spots present in the image. Choosing the binarization level 100 the fine pattern is becoming visible while for threshold level equal 200 the whole area of the lesion and its outer shape and boundaries can be determined.

Information important for medical diagnosis is carried in the coloring of the image. Decomposition of color images into color components is one of the best known image analysis algorithms. Commonly two kinds of color decompositions are widely used namely RGB (additive) and CMY (subtractive). In the ABCD-scale existence of more then three out of six colorings gives a strong indication of cancerous lesion. Computerized methods provide a way to find the colorings which is user-independent. One can define thresholds for RGB levels to find the cancer specific colorings as specified eg. in the caption of Fig. 2. Figures 2a-c show three examples of how the RGB thresholding



(a) (b) (c)

Figure 2: Dangerous color components: (a) white areas can be found by RGB thresholding setting $R > 110$, $G > R - 26$ and $B > R - 20$; (b) black areas can be found by RGB thresholding setting $R < 90$, $G - 10 < B < G + 25B$ or $R/5 - 4 < G < R/2 + 12$, $B > R/4 - 3$; (c) grey-bluish areas can be found by RGB thresholding setting $R > 60$, $R - 46 < G < R + 15$ or $G > B - 30$, $R - 30 < B < R + 45$.

could automatically extract white, black and grey-bluish area in the image. The levels as shown in the experiments were adjusted to give best diagnostic results and are based on past experience and close joint work with clinical dermatologists.

3 Feature selection

The images were collected using a Minolta Dimage Z5 digital camera equipped with an epiluminescence lens with white halogen lighting. The settings of the camera were fixed on resolution of 2272x1704 pixels and quantization depth of 24 bits (RGB-8). Later

all the pictures were scaled down to 800x600 pixels. For differentiation of the images we used some standard features such as the geometrical properties (height, width, symmetry, border length) and standard color representations (RGB or CMYK) as presented above.

We propose to use also as features statistical properties (averages) of signal components used in the image/video decompositions: luminance, C_r and C_b components, average H, S, V, Y, I, Q, average H, S, V, Y, I, Q of background, average luminance and C_b, C_r components of background. These signal components carry important information and can be used as additional features bringing more information about the lesion images. These components up to our knowledge have so far never been used in analysis of dermoscopic images. Also various binary compositions of of colors and colors of the background will be considered.

In the HSV (Hue, Saturation and Value) decomposition the V-value is the intensity of a particular pixel, S - determines the saturation $S = \frac{Max-Min}{Max}$, and Hue can be directly related to the RGB components as: $H = (\frac{G-B}{Max-Min})/6$ if $R = Max$, $H = (2+\frac{B-R}{Max-Min})/6$ if $G = Max$ and $H = (4+\frac{R-G}{Max-Min})/6$ if $B = Max$.

The YIQ or YUV space components widely used in Video/TV eg. in the NTSC standard are related to the RGB via a matrix transformation:

$$\begin{bmatrix} Y \\ I \\ Q \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ 0.586 & -0.274 & 0.322 \\ 0.211 & -0.523 & 0.312 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{1}$$

or

$$\begin{bmatrix} Y' \\ U \\ V \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ -0.14713 & -0.28886 & 0.436 \\ 0.615 & -0.51499 & -0.10001 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{2}$$

Another useful signal decomposition which has been tested is into luminance-chrominance components YC_bC_r:

$$\begin{bmatrix} Y \\ C_b \\ C_r \end{bmatrix} = \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} + \begin{bmatrix} 0.25678824 & 0.50412941 & 0.09790588 \\ -0.1482229 & -0.29099279 & 0.43921569 \\ 0.43931569 & -0.36778831 & -0.07142737 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{3}$$

In total we tested up to 45 features. Among these are lesion geometry descriptors (9) such as: estimated size (px), symmetry (%), area of the lesion (%), area of the lesion (px), area of background (px), height (px), width (px), estimated borderline and borders. Further, color components (8) are also being used: average red, average green, average blue, white color (px), black color (px), gray-blue (px), grey-blue and sum of color components. On the feature list are also various binary compositions of colors and colors of the background (10): Binary sum of GBR, RGB, GRB, RBG, BGR and BRG composition, average red in background, average green in background, average blue in background and sum of background color components. Apart from those commonly used characterizations we use also as features statistical properties of signal components (18): average luminance, average C_r component, average comp. C_b, average H, average S, average V, average Y, average I, average Q, average Y of background, average I of background, average Q of background, average H of background, average S of background, average V of background, average luminance of background, average C_r component of background, average C_b component

of background. All the above-mentioned characterizations have been thoroughly tested on skin lesion images with the goal of finding those features which carry most information useful for diagnostic purposes. To build a useful diagnosis assisting system an efficient classification tool which will use the proposed feature set is needed.

4 How to build a good classifier?

The approach we propose for building classifiers is to use statistical learning techniques for data-based model building. All classifiers we use are model-based. To construct an extremely efficient classifier we build ensembles of well trained but diverse models. There exist a vast variety of available models described in the literature which can be grouped into some general classes

- Global Models (eg. Linear Models, Polynomial Models, Neural Networks (MLP), Support Vector Machines)
- Semi–global Models (eg. Radial Basis Functions, Multivariate Adaptive Regression Splines (MARS), Decision Trees (C4.5, CART))
- Local Models such as k–Nearest–Neighbors
- Hybrid Models such as Projection Based Radial Basis Functions Network (PRBFN)

Implementation of any of such modeling methods leads typically to solution of an optimization problem.

5 Ensemble Methods

Building an ensemble consists of averaging the outputs of several separately trained models eg. $\bar{f}(\vec{x}) = \frac{1}{K}\sum_{k=1}^{K} f_k(\vec{x})$. Krogh et al. derive the equation $E = \bar{E} - \bar{A}$ which relates the ensemble generalization error E with the average generalization error $\bar{E}$ of the individual models and the variance $\bar{A}$ of the model outputs with respect to the average output. When keeping the average generalization error $\bar{E}$ of the individual models constant, the ensemble generalization error E should decrease with increasing diversity of the models $\bar{A}$. Hence we try to increase A by using two strategies:
• Re-sampling: Each model is trained on a randomly drawn subset of 80% of all training samples. The number of models trained for one ensemble is chosen so that all samples of the training set are covered at least once by the different subsets.
• Variation of model type.
On average the ensemble of well trained but diverse models will perform better then any of the members of the ensemble. This does not mean however that in specific cases performance of a selected model will not be better.

In order to select models for the final ensemble we use a cross validation scheme for training. As the models are initialized with different parameters (number of hidden units, number of nearest neighbor, initial weights, etc.), cross validation helps us to find a proper value for these model parameters. The cross validation is done in several training rounds on different subsets of the entire training data. The trained models are compared by evaluating their prediction errors on the unseen data of the test set. The model with the smallest test error is taken out and becomes a member of the ensemble. This is repeated several times.

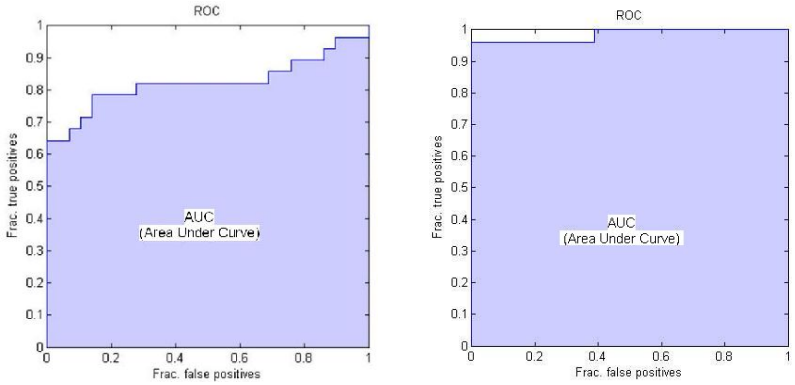
6 The ENTOOL Toolbox for Statistical Learning

The ENTOOL toolbox for statistical learning contains machine learning algorithms available under a common interface. It allows construction of single models or automatic generation of ensembles of (heterogenous) models. ENTOOL is Matlab-based with parts written in C++. Each model type is implemented as separate class in our simulator, all model classes share common interface.

Each selected model goes through three phases: Construction, Training and Evaluation. In the construction phase topology of the model is specified. The model can't be used yet – it has now to be trained on some training data set $(\vec{x}_i, y_i)$. After training, the model can be evaluated on new/unseen inputs $(\vec{x}_n)$. All the described methods are available from our web-site [6]. New features and algorithms are being added continuously to our toolkit. The toolkit has been tested on a variety of problems from ECG modeling, CNN training, financial time series, El Niño real data and many others [12].

7 Results

To compare performance of classifiers employing various features we use so-called ROC curves ie. plots of the true-positive fraction versus the false-positive fraction. A single threshold value chosen for the classifier will produce a single point on the ROC curve.



(a) (b)

Figure 3: ROC curve showing the quality of the classification when (a) using all described features AUC=0.82 (a). and (b) when using 15 best features as suggested by the sensitivity analysis AUC=0.91 (b).

In typical cases (on average for the whole data set) the built ensemble classifier outperforms all available single-model-based methods including SVM. As it is typical in the applications of ensembling methods for specific single cases it has been found that SVM gave better results. Also varying the number of features taken for classification we found that the results could be significantly affected. The sensitivity analysis performed enabled us to make a list of features following their importance for classification. It is clear that some of the feature do not contribute or bring only very minor contributions

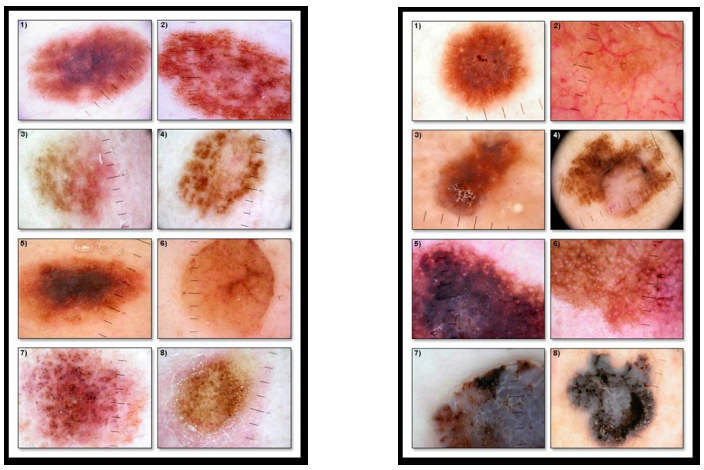


Figure 4: Using an ensemble of classifiers and 15 best selected features it has been possible to differentiate some of the difficult cases. Images in the left column are the displastic ones while those on the right are melanoma cases confirmed by histopatology studies.

or even degrade the performance index. Those features have been eliminated from the final set. It was impossible however to find the multi-parameter relations ie. how the performance changes when different sets of features were selected. Fig.3 shows two results of the classifier performance when all the features were used for classification and when the number of features is 15 (most significant ones). On this list of 15 most significant features determined using sensitivity analysis we find standard features used by the dermatologists such as symmetry, grey-bluish, black and white components, coloring - average red, green and blue, but also sum of color components of the background, average S and Y of background, average green of background, average V and Q of background. It is interesting to notice that some of the video signal components are very important, namely the average Y, Q and V and average luminance. Most striking conclusion is that the color and signal components of the background play very significant role. These features bring major contributions for the definition of borderline and also to some of the differential structures. Fig. 4 gives an example of images which have been differentiated using the proposed approach and which were wrongly classified using standard medical approaches. More thorough study of the dependence of performance of the method on the selection of feature set will be reported elsewhere due to lack of space.

8 Concluding remarks

Proposed methodology for computer-aided classification of skin lesions for diagnostic support merges medical experience with several cutting-edge technologies: image processing, pattern classification, statistical learning, ensembling techniques of model-based classifiers. The proposed approach tested on the database of cases from Jagiellonian University Dermatology Clinic (containing entries with full medical history and histopathology support) proved to give excellent results of up to 98% correct classifications. As suggested in [15] other methods such as wavelet based approaches could be combined

with learning algorithms. It should be also noticed that the recognition of specific textures has not been taken into account. We claim that the signal decomposition approach adopted (such as Y, Q, V) and use of averaged quantities contains this type of information.

References

[1] Argenziano, G., Fabbrocini, G., Carli, P., DeGiorgi, V., Sammarco, PE., Delfino, M.: Epiluminescence Microscopy for the Diagnosis of Doubtful Melanocytic Skin Lesions, Arch. Dermatl., vol.134, pp.1563-1570, (1998)

[2] Burroni, M., Corona, R., Dell'Eva, G., Sera, F.,Bono, R., Puddu, P., Perotti, R., Nobile, F., Andreassi, L., Rubegni, P.: Melanoma Computer-Aided Diagnosis: Reliability and Feasibility Study, Clin. Cancer Res., vol.10, pp.1881-1886 (2004)

[3] `http://www.dermoscopy.org/atlas/base.htm`

[4] `http://www.dermis.net/dermisroot/en/home/index.htm`

[5] Dreiseitl, S., Ohno-Machado, L., Kittler, H., Vinterbo, S., Billhards, H., Binder, M.A.: Comparison of Machine Learning Methods for the Diagnosis of Pigmented Skin Lesions, J. Biomed. Inform., vol. 34, pp.28-36, (2001)

[6] `http://zti.if.uj.edu.pl/merkwirth/entool.htm`

[7] Grzymala-Busse, P.; Grzymala-Busse, J.W.; Hippe, Z.S.: Melanoma prediction using data mining system LERS. COMPSAC'2001. pp.615 - 620 (2001)

[8] Hall, P.N., Claridge, E., Smith, J.D.: Computer Screening for Early Detection of Melanoma: Is there a Future?, British J. Dermatol., vol.132, pp.325-328, (1995)

[9] Iyatomi, H., Oka, H., Hasimoto, M., Tanaka, M., Ogawa, K.: An Internet-based Melanoma Diagnostic System - Toward the Practical Application.

[10] Johr, R.H.: Dermoscopy: Alternative Melanocytic Algorithms - The ABCD Rule of Dermatoscopy, Menzies Scoring Method, and 7-Point Checklist, Clinics in Dermatology (Elsevier), vol.20, pp.240-247, (2002)

[11] Menzies, S.W.: Automated Epiluminescence Microscopy: Human vs Machine in the Diagnosis of Melanoma, Arch. Dermatol., vol.135,pp.1538-1540, (1999)

[12] Merkwirth C., Wichard J., Ogorzałek M.J. Ensemble Modeling for Bio-medical Applications, chapter in W. Mitkowski and J. Kacprzyk (Eds.): Model. Dyn. in Processes & Sys., SCI 180, pp. 119-135, Springer-Verlag (2009)

[13] Schmid-Saugeon, P., Guillod, J., Thiran, J.-P.: Towards a Computer-aided diagnosis System for Pigmented Skin Lesions, Comp. Med. Imag. Graphics, pp.65-78, (2003)

[14] Stolz, W., Riemann, A., Cognetta, A. B., Pillet, L., Abmayr, W., Hölzel, D., Bilek, P., Nachbar, F., Landthaler, M. Braun-Falco, O.: ABCD rule of dermatoscopy: a new practical method for early recognition of malignant melanoma. Eur. J. Dermatol. 7, pp.521 - 528, (1994).

[15] Surówka G., Merkwirth C., Żabińska-Płazak, Graca A., Wavelet based classification of skin lesion images, Bio-Algorithms and Med Syst., vol.2 no.4, pp. 43-50 (2006).