2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

On optimal wavelet bases for classification of skin lesion images through ensemble learning

Grzegorz Surówka and Maciej Ogorzałek

***Abstract*—In order to recognize early symptoms of melanoma, the fatal cancer of the skin, systems for computer aided melanoma diagnosis have been developed for years. In this work we analyze an ensemble-based binary classifier for discriminating melanoma from dysplastic nevus utilizing wavelet-based features of the dermatoscopic skin lesion images. The multiresolution decomposition of the dermatoscopy images is done through wavelet packets. We search for the optimal wavelet base maximizing the quality of the classifier in terms of AUC (Area Under Curve) for models optimized by some common quality measures: accuracy, precision, F1-score, FP-rate, specificity, BER and recall. Within the statistics of our experiments reverse bi-orthogonal wavelet rbio3.1 makes the best wavelet model of melanoma.**

I. Introduction

The increasing rate of cutaneous melanoma worldwide has been a big epidemiologic problem for years due to its early metastases and high mortality rate [1]. Transformations of the pigment cells in the epidermis may lead to: benign (melanocytic nevus), atypical (dysplastic nevus) and malignant stages (malignant melanoma)[2]. Medical doctors use some descriptive measures based on a visual examination to classify the stage of atypia: ABCD(E), the 7-Point Checklist or Menzies to mention the most common [3]. Those checklists have certain geometric or coloristic criteria that contribute to the total score. Observations of the moles are made with help of ELM (Epiluminescence Microscopy) i.e. dermatoscopy. This is a non-invasive technique that consists in optical enlarging and illuminating the skin by white (halogen) light. The magnified field of the lesion can be digitally photographed or displayed on a computer screen for analysis of its surface structure [4]. Some dedicated instruments allow for trans-illumination where the light is directed into the skin at an angle of 45^0 or use a set of wavelengths to penetrate deeper layers of the skin and thus attempting to reveal its 3D structure. The most common are however the cheapest, plain dermatoscopes.
Dermatoscopy images recorded and stored on a computer can be compared for how lesions develop in time, transmitted to a clinic/remote specialist for a (tele)consultation or analyzed by dermatoscopy management software [5].

Grzegorz Surówka is with Faculty of Physics, Astronomy and Applied Computer Science, Jagiellonian University, 30-151 Kraków, Poland (phone: +48126635590; fax: +48126337086; e-mail: grzegorz.surowka@uj.edu.pl).
Maciej Ogorzałek is with Faculty of Physics, Astronomy and Applied Computer Science, Jagiellonian University, 30-151 Kraków, Poland (phone: +48126635827; e-mail: maciej.ogorzalek@uj.edu.pl).
This work was supported by the Polish National Science Center under Grant **N** N518 419038.

Benign melanocytic nevi are usually well recognized in dermatoscopy images whereas discrimination between dysplastic nevus and melanoma may be very difficult even for experienced specialists. It is in force especially at the earliest stages of malignancy, when resection can be a life-saving factor [6]. Biopsy and the subsequent microscopy examination is the only fully reliable method to identify nevi and melanoma lesions. Based on histologic criteria for melanoma two main staging schemes have been proposed: Clark and Breslow. The Clark's level (I-V) differentiates the degree of tumor penetration quantitatively while the Breslow's depth is an actual micrometer measurement of the lesion depth and is grouped into four categories (<0.75mm, 0.75-1.5 mm, 1.5-4.0mm and >4.0mm) which determines the prognosis of the case. Since biopsy of all suspicious moles is not feasible (economy, surgical complications, ANS-Atypical Nevus Syndrome), early detection of malignant moles is the key of effective treatment. This, however, is with high accuracy still unsatisfactory. Quality of dermatoscopic diagnosis depends on the appearance of classic dermatoscopic features, but early stages of melanoma are mainly featureless [7]. In this case clinical descriptive measures and geometric/coloristic segmentation is not sensitive enough.
Methods for wavelet based decomposition of skin lesion images have been proposed since the 90-ies of the XX century [8]. They assume analysis of frequency and scale information found in the skin texture to be a sensitive probe of the pigmented skin atypia and the melanoma progression. Discrete wavelet transforms are closely related to the theory of digital filtering so the properties of the decomposition filters (the choice of the basis, degree of regularity, the sub-bands of interest) play an important role in the skin texture characteristics [9], [10].
Various factors play a role in the discrete wavelet analysis [11]:
i) decomposition path: recursive decomposition of the low-frequency (averaged) signal (=the pyramidal algorithm) or a selective tree-structured analysis where the consecutive decomposition is applied to the output of any channel (=wavelet packets/trees),
ii) wavelet base: this choice has a diverse impact on the texture classification,
iii) wavelet order: decomposition over an optimal finite range of resolutions,
iv) model constraints: orthogonality (the wavelet transform is energy preserving and nonredundant) versus bi-orthogonality (separate filters for decomposition and synthesis are present, wavelets are more compact and symmetric at the cost of orthogonality),

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

iv) sampling of 2D signals: the Mallat algorithm [12].

Pioneering contributions on wavelet based decomposition of melanoma dermatoscopy images belong to Patwardhan et al. [9], [10]. This group successfully studied binary classification models for benign nevus and melanoma by decomposing different frequency scales of the skin texture (wavelet packets). This approach corresponds to the observations that the significant sub-bands of the pigmented skin texture belong to the middle frequency range and the standard (recursive) analysis of the low-frequency sub-band only, is less optimal than the wavelet packets (also called selective wavelet trees).

II. Motivation

Melanoma binary classifiers from Patwardhan [9], [10] and following contributions [13], [14], [15], [16] were using only one wavelet base (Daubechies 3) to build classification models. In this work we study different wavelet bases and analyze how they affect the quality of the classification models. As a framework to test wavelet features we use a moderate ensemble of six different model types. We don't aim at optimization (fine-tuning) of any single model or an ensemble of models beyond the standard machine learning procedure (cross-validation) e.g. through feature selection, but use the ensembling technique as a 'blind' learning environment to find optimal wavelet bases in terms of classification quality measures: accuracy, precision, recall, specificity, false-positive rate (FP-rate), F-score and balanced error rate (BER).

In the following sections we show methodology of our machine learning experiments and present the results. Discussion on mathematical implications (wavelet properties) of the results is beyond the scope of this work.

III. Methodology

A. Image acquisition and preparation

Dermatoscopy images of 2272x1704 pixel resolution and RGB 24-bit color depth were collected from different 185 anonymous patients with Minolta Z5 digital camera with an extra dermatoscopy extension. After resection and histopathology examination all the melanoma cases (102) were coded with '1' and dysplastic nevus cases (83) as '-1'. Our experimental setup was coded in Matlab 8.1.0.604 (R2013a) [24] with help of 'Image processing Toolbox', 'Wavelet toolbox' and Entool [17]. Since the Wavelet Toolbox supports only indexed images with linear, monotonic color maps, all the JPG dermatoscopy images had to be transformed into this format. Finally our dataset consisted of one 185x2272x1704 matrix of double precision numbers.

Since each iteration of the wavelet decomposition downscales the input image by a factor of 2 in the rows or columns, after three iterations the width and length of the resulting filtered images were still integer numbers 284x213 so no initial padding was required.

In this study no preprocessing tasks to the images were done e.g. removal of artefacts like tiny hairs, remaining droplets of immersion fluid, etc. This was to eliminate any bias on the final wavelet base selection.

B. Wavelet packets

Wavelet analysis of signals is well established in theory after studies of Gabor, Morlet, Daubechies, Mallat and the others [11], [12]. It is also widely used especially for discrete signals in the form of DDWT-Discrete Dyadic Wavelet Transform to analyze the signal structure, signal de-noising and compression capabilities. Images are two-dimensional signals so one iteration of the Mallat filtering algorithm produces 4 sub-images which can be considered as LL, LH, HL and HH filters (L-low-pass, H-high-pass filter) after one-dimensional wavelet transform on the rows and then on columns. Since we used the wavelet packets, the further iterations were done on each of the four parent filters. Altogether in three iterations 1+4+16=21 different transformation branches were produced. In one branch the following 12 features were calculated: (e_i, i=1,2,3,4) - energies of the sub-images (energy is defined as a sum of absolute values of the pixels), (e_i/e_{max}, i=1,2,3,4) - maximum energy ratios, ($e_i/\Sigma e_k$, $k\neq i$, i=1,2,3,4) - fractional energy ratios [9], [10], [13]. This procedure was repeated for each wavelet base, producing different sets of 21x12=252 attributes.

Different wavelet bases decompose the skin texture with variable classification accuracy. Also numeric properties and the management of computer resources are important factors when choosing certain wavelet families. For discrete signals, orthogonal or bi-orthogonal and compactly supported wavelet functions are usually taken into account when analyzing patterns. It has to do with accuracy of the signal reconstruction, monotonic behavior affecting the convergence and number of vanishing moments (representation density) [11]. We tested orthogonal wavelets:

i) Daubechies db1-db10 (wavelet number=2-11),
ii) symlets: sym2-sym8 (wavelet number=12-18),
iii) coiflets: coifN (wavelet number=19-23),

and bi-orthogonal/reverse bi-orthogonal wavelets:

iv) biorNr.Nd (wavelet number=24-38),
v) rbioNr.Nd (wavelet number=38-53).

(Reverse) Bi-orthogonal wavelets (wavelet pairs) have the property of perfect reconstruction i.e. X=A+D, where: X-image, A-reconstructed image of approximation and D-reconstructed image of details. This property is possible due to two separate filter sets, one for decomposition and another one for image reconstruction. Those wavelets are not orthogonal. Orthogonal wavelets, on the other hand, fulfil the formula $X^2=A^2+D^2$. Symlets, coiflets and (reverse) bi-orthogonal wavelets are symmetric functions, whereas Daubechies - asymmetric [11].

C. Ensembling

An ensemble is a set of single machine learning models f_k whose predictions are combined by voting or weighted averaging [18]:

$$\overline{f}(\vec{x}) = \sum_{k=1}^{K} w_k f_k(\vec{x}) \qquad \text{Eq.1}$$

(x: data matrix=(cases)*(attributes), y: output class, $\Sigma_k w_k=1$).

It is known that the generalization error of the ensemble:

$$e(\vec{x}) = \left(y(\vec{x}) - \overline{f}(\vec{x})\right)^2 = \overline{\varepsilon}(\vec{x}) - \overline{\delta}(\vec{x}) \qquad \text{Eq.2}$$

can be decomposed into an average error of the individual models:

$$\overline{\varepsilon}(\vec{x}) = \sum_{k=1}^{K} w_k \left(y(\vec{x}) - f_k(\vec{x})\right)^2 \qquad \text{Eq.3}$$

and average ambiguity of the ensemble:

$$\overline{\delta}(\vec{x}) = \sum_{k=1}^{K} w_k \left(f_k(\vec{x}) - \overline{f}(\vec{x})\right)^2 \qquad \text{Eq.4}$$

The ensemble generalization error (Eq.2) is always smaller than the mean of the generalization error of the single ensemble members (Eq.3), which makes this technique a good tool to maximize the classification performance. In order to increase the ensemble ambiguity (Eq.4) it should consist of well trained but diverse models of any type (no assumptions are made about the constituent models).

To build an ensemble of models starting from an empty ensemble we were selecting step-by-step the best models by a cross-validation scheme for model training (the so called OOT-Out-of-Train procedure (after Breiman's Out-Of-Bag technique) [19]. The cross-validation was done in several training rounds on different training sets, because this increases the ambiguity of the ensemble and leads to better generalization. This is one way of introducing diversity of models because training on slightly different data sets leads to different models. Another advantage of this method is that one gets an unbiased estimator of the ensemble generalization error. The whole procedure consisted of the following steps:

0) data are divided into training/testing set (80%) and validation set (20%), cross-validation partitions: 5, here final quality (AUC) of the trained ensemble was calculated as the mean of the five trained samples,
1) training/testing data are divided into training set (80%) and testing set (20%), cross-validation partitions: 10
2) several models are trained on the training set,
3) these models are compared by evaluating the prediction errors on the testing set,
4) the best models are taken out and become ensemble members,
5) data are divided again in a way that the new testing set has minimal overlap with the former ones,
6) the procedure stops if the ensemble has the desired size.

Training in step 2) was performed with the following six model families:

-Penalized Fisher's Linear Discriminant Analysis: classical LDA classifier with spatial constraints on many highly correlated predictors (a model for pixels in an image) [20],
-Kernel Ridge Regression: a model with the Tikhonov-Phillips regularization capable of controlling bias-variance trade-off, with a polynomial kernel k(x,x') = (a+x.x')^b and coefficient a and b [21],
-Multi Layer Perceptron: trained with the first order weight update mechanisms (RPROP descent), with the changeable number of nodes [22],
-Perceptron: trained with a second order gradient decent [22],
-Decision Trees: based on C4.5 algorithm, with pruning procedures based on cross-validation scheme [23],
-Matlab data trees (dtree) [24].

D. Quality measures for supervised learning

In binary classification a confusion matrix shows the test outcome versus the true condition which means it presents instances of predicted and actual classes [25]. This can visualize performance of the model on validation data. The four statistical entities: tp=true positive, tn=true negative (they are the desired results) and fp=false positive (type I error) and fn=false negative (type II error) form a set of values out of which numerous quality measures are derived. The choice for a measure and its application in the classification scheme depends on the research purpose. In our experiments we used (one by one) seven different quality measures to control how the ensembles of primary models are constructed. Those measures were optimization factors when accumulating best constituent models. Below we list them with brief annotations [25]:

accuracy: an overall measure of all desired outcomes in the test (tp+tn)/(tp+tn+fp+fn). This is a common quality measure when no particular requirements are imposed.

precision: aka PPA (positive predictive value) is a fraction of retrieved instances that are relevant, tp/(tp+fp). This is a quality measure of exactness/quality.

recall: aka sensitivity is a fraction of relevant instances that are retrieved, tp/(tp+fn). This is a quality measure of completness/quantity.

Absence of type I and type II errors corresponds respectively to maximum precision (no false positive) and maximum recall (no false negatives).

F score: F(1) score is a harmonic mean of precision and recall 2*(precision)*(reall)/(precision+recall) =2*tp/(2*tp+fn+fp).

fp rate: false positive rate fp/(tp+fp).

specificity: is a fraction of true negative, tn/(tn+fp). A high specificity has a low type I error rate.

ber: balanced error rate is an average of the errors on each class 0.5*(fn/(tp+fn) + fp/(fp+tn)).

The ensembles trained according to the above mentioned quality measures were finally (step 0) tested on validation data. For the quality measure at this step we chose AUC - the area under the ROC curve (Receiver Operating Characteristic) [26] obtained by plotting sensitivity=tp/(tp+fn) against (1-specificity)=tn/(tn+fp), for each confidence value. The ROC curve is better in presenting the quality of the classification system than any single quality measure since one can obtain sensitivity and specificity as a function of a confidence level (thresholds between single values of calculations from the model). The values of AUC presented in all the figures were calculated from the ROC

curve using the trapezoid method.

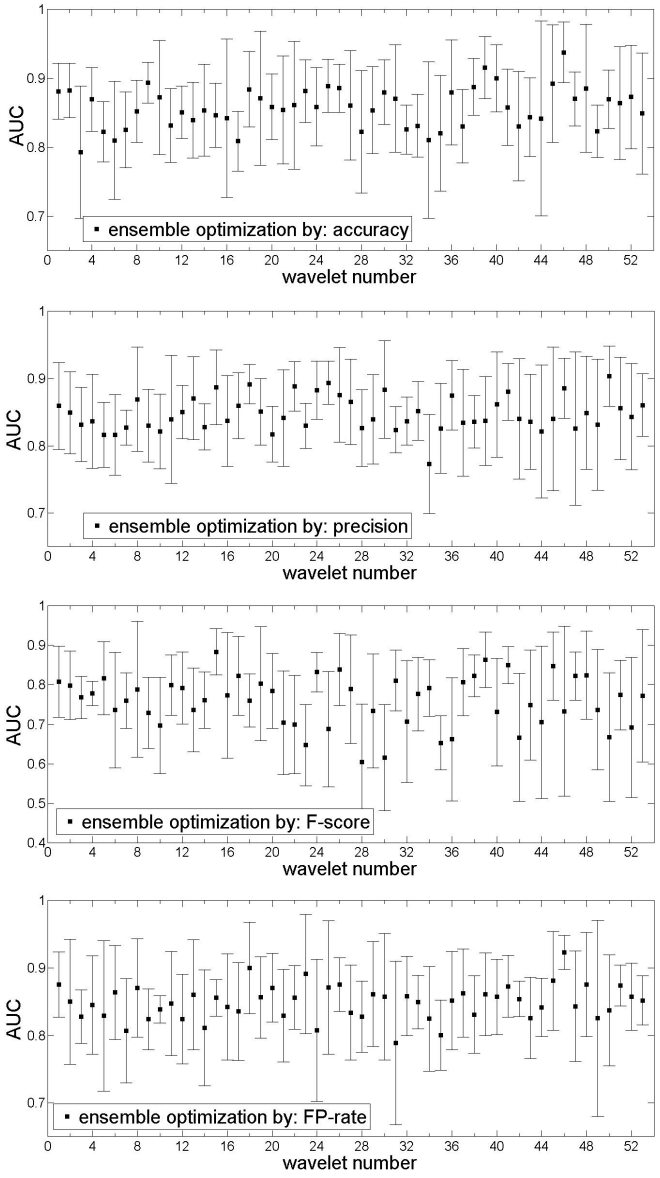

Fig. 1. AUC (Area under Curve) for the ensemble classifiers optimized for different quality measures (accuracy, precision, F-score, FP-rate) as a function of wavelet bases.

IV. Results and Discussion

Fig. 1 and Fig. 2 present seven different AUC values as a function of the wavelet base number. The wavelet numbers are mapped to particular wavelet names in Section III.B.

Each wavelet was used to decompose a set of dermatoscopic images and to calculate a corresponding feature set. The wavelet features were learnt by an ensemble of models in this way, that the ensemble optimized (one by one) seven different quality measures: accuracy, precision, F1-score, fp rate, specificity, ber and recall. The final model was validated on a separate unseen set of data (pulled out at Step 0).

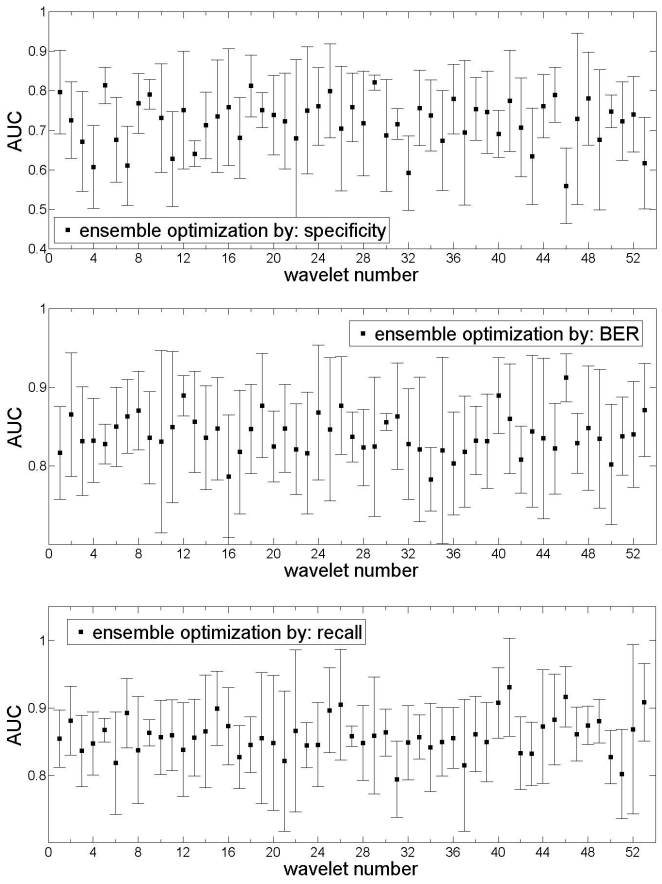

Fig. 2. AUC (Area under Curve) for the ensemble classifiers optimized for different quality measures (specificity, BER, recall) as a function of wavelet bases.

The AUC values in Fig.1, 2 have error bars that reflect standard deviation of the AUC value over different validating rounds. Our first observation refers to the magnitude of the error bars. For most of the quality measures they are bigger than the fluctuations of AUC over different wavelet bases. This confirms that the learning environment plays an important role in the stability of the models and this role may outperform or at least screen the influence of a wavelet base.

Each quality measure seems to 'prefer' its own mean level of magnitude plus fluctuates with the wavelet number. Where performance of a given measure has a local hill, the error bar tends to be smaller (solution in a local extremum 'gets stuck').

AUC differs in the absolute level between different quality measures and also fluctuates among the wavelet bases. For the gathered statistics, the run of 'accuracy' seems to yield the maximum AUC values overall. 'Recall' (i.e. sensitivity) has also a high level of performance (the latter two have hills in different wavelet numbers) but its error bars are slightly bigger. Quality measures between AUC=0.8 and 0.9 have apparent zones of correlation (where they follow each other in their monotonic runs) and decorrelation. Correlation of different learning schemes proves stability of the learning environment. For a better comparison between the runs Fig. 3 presents experimental points connected with a line. Fig. 3

helps us follow the absolute level and variability of AUC for different ensemble development scenarios.

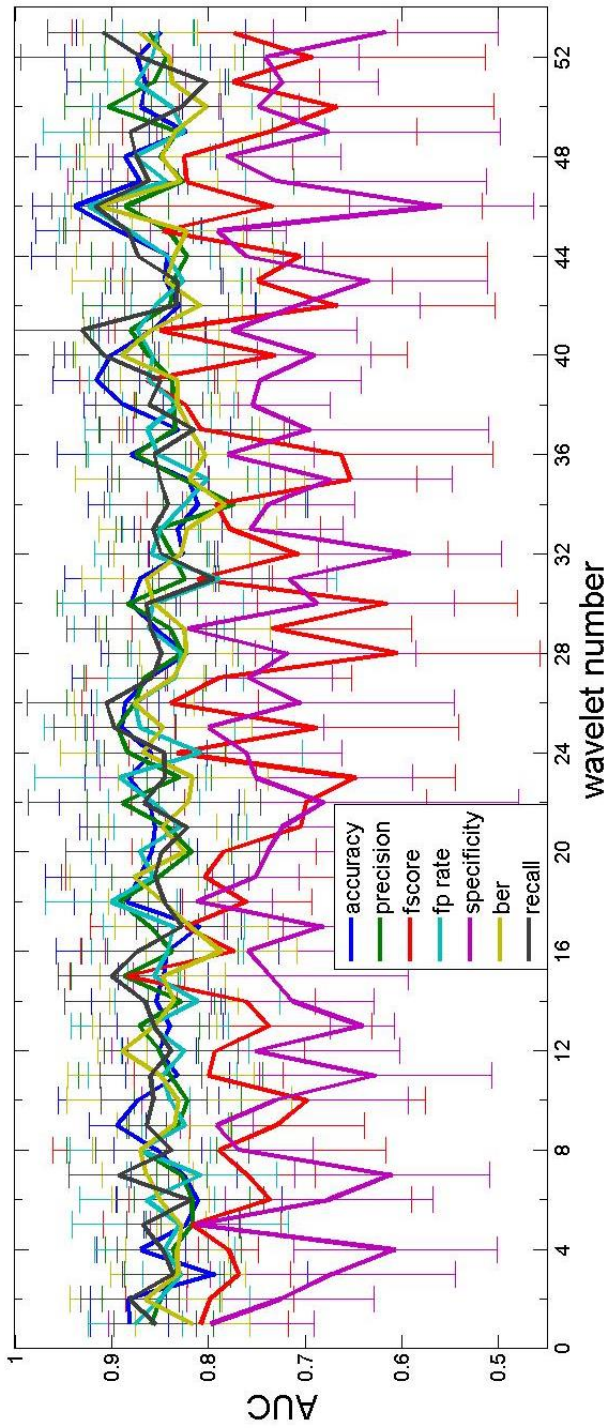

Fig. 3. AUC (Area under Curve) for the ensemble classifiers optimized for different quality measures as a function of wavelet bases - a comparison.

The most optimal wavelet base in our experiments is wavelet number 46 i.e. reverse bi-orthogonal wavelet rbio3.1. At wavelet number 46 all the examined measures achieve (global) extrema (except for 'recall' which has its global extremum at wavelet number=41 i.e. for rbio1.5).

Feature set based on rbio3.1 introduces much stability to the solutions for all of the optimization measures. At this point also error bars of AUCs are diminished simultaneously for all the ensemble learning scenarios.

V. Conclusion and Outlook

We performed some machine learning experiments to search for optimal wavelet bases for decomposition of dermatoscopic images of melanoma (102 cases) and dysplastic nevus (83 cases). This is motivated by the medical problem of pattern recognition of early stages of the cutaneous melanoma. In literature a lot of attempts have been done to find an optimal representation of melanoma in order to maximize its classification performance. Wavelet bases seem to outperform other melanoma representations (geometrical and coloristic) so experiments on how different wavelet bases affect quality of binary classification of dermatoscopic images of skin lesions are very important. We collected results of seven ensemble learning cases optimized for different quality measures. Different wavelet bases, we used, affect the training process of the ensembles of models in a different way according to different supervision of the quality measure on the learning environment.

Since in this work we focused on a 'blind' selection of the most optimal wavelet base in terms of its classification performance, future experiments may i) imply certain mathematical properties of the wavelet base to explain its performance and ii) extend model pool and/or reduce dispersion of the final models to draw more detailed/exact conclusions about the optimal wavelet bases for the wavelet model of melanoma.

References

[1] Marks R., Epidemiology of melanoma, Clin. Exp. Dermatol. 25, 2000.

[2] R.B. Odom, W.H. James, T.G. Berger, Melanocytic nevi and neoplasms, in: Andrews' Diseases of the Skin, 9th ed., Philadelphia, 2000.

[3] R.H. Johr, Dermatoscopy: alternative melanocytic algorithms - the ABCD rule of dermatoscopy, Menzies scoring method, and 7-point checklist, Clinics in Dermatology, 20, 2002.

[4] E. Żabińska-Płazak, A. Wojas-Pelc, G. Dyduch: Videodermatoscopy in the diagnosis of melanocytic skin lesions, Bio-Algorithms and Med-Systems, 1, 2005.

[5] J.K. Robinson, B.J. Nickoloff, Digital epiluminescence microscopy monitoring of high-risk patients, Arch. Dermatol. 140, 2004.

[6] P. Carli, V. De Giorgi, B. Gianotti et al., Dermatoscopy and early diagnosis of melanoma, Arch. Dermatol. 137, 2001.

[7] A. Roesch et al., Dermatoscopy of "dysplastic nevi": A beacon in the diagnostic darkness, Eur. J. Dermatol 16(5), 2006.

[8] T. Chang, C.C.J. Kuo, Texture Analysis and Classification with Tree-Structured Wavelet Transform, IEEE Transactions on Image Processing 2, 1993.

[9] S.V. Patwardhan, A.P. Dhawan, P.A. Relue, Classification of melanoma using tree structured wavelet transforms, Computer Methods and Programs in Biomedicine 72, 2003.

[10] S.V. Patwardhan, S. Dai, A.P. Dhawan, Multi-spectral image analysis and classification of melanoma using fuzzy membership based partitions, Computerized Medical Imaging and Graphics 29, 2005.

[11] I. Daubechies, Ten lectures on wavelets, CBMS, SIAM, 61, 1994.

[12] S.G. Mallat, A Theory for Multiresolution Signal Decomposition: The Wavelet Representation, IEEE Transactions on pattern analysis and machine intelligence, 11, 1989.

[13] G. Surówka, Ch. Merkwirth, E. Żabińska-Płazak, A. Graca, Wavelet based pattern recognition analysis of skin lesion images, Bio Alg. Med Syst. 2(4), 2006.
[14] G. Surówka, K. Grzesiak-Kopeć, Different Learning Paradigms for the Classification of Melanoid Skin Lesions Using Wavelets, Proc. of. EMBC07, Lyon.
[15] G. Surówka, Supervised learning of melanocytic skin lesion images, Proc. HSI 2008, Kraków.
[16] L. Nowak, A. Alekseenko, M. Ogorzałek and G. Surówka, Modern techniques for computer-aided melanoma diagnosis, in „*Melanoma in the clinic*" ed. Mandi Murph, ISBN 978-953-307-571-6, InTech, Croatia, 2011.
[17] C. Merkwirth, J.D. Wichard and M. Ogorzalek, A software toolbox for constructing ensembles of heterogenous linear and nonlinear models, Proceedings of the 2005 European Conference on Circuit Theory and Design 3, Ireland, 2005.
[18] L.I. Kuncheva, Combining Pattern Classifiers, John Wiley & Sons, Inc., 2004.
[19] L. Breiman, Bagging Predictors, Machine Learning 24, 1996.
[20] T. Hastie, A. Buja, R. Tibshirani, Penalized Discriminant Analysis, Ann. Stat. 23, 1995.
[21] S. An, W. Liu, S. Venkatesh, Fast cross-validation algorithms for least square support vector machine and kernel ridge regression, Patt. Recog. 40, 2007.
[22] S. Haykin, Neural Networks: A Comprehensive Foundation (2ed.), Prentice Hall., ISBN 0-13-273350-1, 1998.
[23] J.R. Quinlan, Induction of Decision Trees, Machine Learning 1, Kluwer Academic Publishers, 1986.
[24] MATLAB, The MathWorks Inc., Natick, MA, 1994-2013.
[25] http://en.wikipedia.org/wiki/Binary_classification, visited: Jan 18, 2014.
[26] Receiver operating characteristic (ROC) analysis: Basic principles and applications in radiology, Eur. J. Radiology, 27, 1998.