feature selection,
classification, breast cancer

Paweł FILIPCZUK[1], Marek KOWAL[1], Andrzej MARCINIAK[1]

FEATURE SELECTION FOR BREAST CANCER MALIGNANCY CLASSIFICATION PROBLEM

The paper provides a preview of some work in progress on the computer system to support breast cancer diagnosis. Diagnosis approach is based on microscope images of the FNB (Fine Needle Biopsy) and assumes distinguishing malignant from benign cases. Studies conducted focus on two different problems, the first concern the extraction of morphometric parameters of nuclei present in cytological images and the other concentrate on breast cancer nature classification using selected features. Studies in both areas are conducted in parallel. This work is devoted to the problem of feature selection from the set of determined features in order to maximize the accuracy of classification. Morphometric features are derived directly from a digital scans of breast fine needle biopsy slides and are computed for segmented nuclei. The quality of feature space is measured with four different classification methods. In order to illustrate the effectiveness of the approach, the automatic system of malignancy classification was applied on a set of medical images with promising results.

1. INTRODUCTION

Breast cancer is the most common cancer among women. The prognosis in breast cancer is strongly dependent on the disease development before any treatment is applied so the chance of recovery is a function of time of the detection of cancer. Modern medicine does not provide one hundred percent reliable, if possible cheap and at the same time non-invasive diagnostic methods for the diagnosis of breast pathology. As a result, in practice the important function acting in breast cancer diagnosis is the so-called triple-test, which is based on the summary of results of three medical examinations with different degrees of sensitivity and it allows to achieving high confidence of diagnosis. The triple-test includes self examination (palpation), mammography or ultrasonography imaging and fine needle biopsy [15]. Fine needle biopsy is collecting nucleus material directly from tumor for microscopic verification. Next, the material (collected cells) is examined using microscope in order to confirm or exclude the presence of cancerous cells. The present approach requires a deep knowledge and experience of the cytologist responsible for diagnosis. In short, some pathologists can diagnose better than others. In order to make the decision independent of the arbitrary factor, morphometric analysis can be applied. Objective analysis of microscopic images of cells has been a goal of human pathology and cytology since the middle of the 19th century. Early work in this area consisted of simple manual measurements of cell and nuclear size. Along with the development of advanced vision systems and computer science, quantitative cytopathology has become a useful method for the detection of diseases, infections as well as many other disorders. In the literature one can find approaches to breast cancer classification [2, 4, 5, 10, 12, 13, 16, 18]. Mentioned approaches are concentrated on classifying FNA (Fine Needle Aspiration) or FNB biopsy slides as benign or malignant.

In this work, we present a method that allows distinguish malignant cells from the benign cells. The classification of the tumor is based on morphometric examination of cell nuclei. In contrast to normal and benign nuclei, which are typically uniform in appearance, cancerous nuclei are characterized by irregular morphology that is reflected in several parameters. Morphometric measurements characterizing the size, cell grouping and color changes within the nuclei have been mainly used for feature extraction. It was decided not to use shape features because previous work showed that shape factors do not have good discriminative properties [8].

The quality of feature subset is measured using the set of classifying algorithms. The measure is based on classification accuracy obtained by leave-on-out cross-validation. In this work four different

[1] University of Zielona Góra, Institute of Control and Computation Engineering.
{M.Kowal,P.Filipczuk,A.Marciniak}@issi.uz.zgora.pl.

classification methods was used to rate the feature subsets: k-nearest neighbor, naive Bayes classifier, decision trees and classifiers ensemble [1, 9]. Taking into account the fact that exhaustive search of feature space is generally impractical, sequential forward selection was applied to add best feature in each step of the search algorithm.

The paper is divided into three sections. Section 1 gives an overview of breast cancer diagnosis techniques. Section 2 describes the process of acquisition of images used to breast cancer diagnosis. Section 3 deals with feature selection problem. Section 4 shows the experimental results obtained using the proposed approach. The last part of the work includes a conclusions and bibliography.

2. ORIGIN AND ACQUISITION OF THE IMAGES

It is necessary to have appropriate amount of real case data to test new developed as well as existing image analysis algorithms. Probably, the most popular database of FNB images and nuclei features is Wisconsin Database of Breast Cancer (WDBC). However, the quality of images delivered in the set is unsatisfactory for image analysis methods described in the paper. Because of that we decided to use our own data set.

The database contains 500 images of the cytological material obtained by FNB. The material was collected from 50 patients of outpatient clinic ONKOMED in Zielona Góra. It gives 10 images per case which was recommended amount by specialists from the Regional Hospital in Zielona Góra [8]. This number of images per single case allows correct diagnosis by a pathologist. The set contains 25 benign and 25 malignant lesions cases. Biopsy without aspiration was performed under the control of ultrasonograph with a 0.5 mm diameter needle. Smears from the material were fixed in spray fixative (Cellfix of Shandon company) and dyed with hematoxylin and eosin (h+e). The time between preparation of smears and their preserving in fixative never exceeded three seconds. The images were recorded by SONY CDD IRIS color video camera mounted atop an AXIOPHOT microscope. The slides were projected into the camera with 10 and 160× objective and a 2,5× ocular. One image was generated for enlargement 100× and nine for enlargement 400×. Images are BMP files, 704×578 pixels, 8 bit/channel RGB (Fig. 1). All cancers were histologically confirmed and all patients with benign disease were either biopsied or followed for a year.

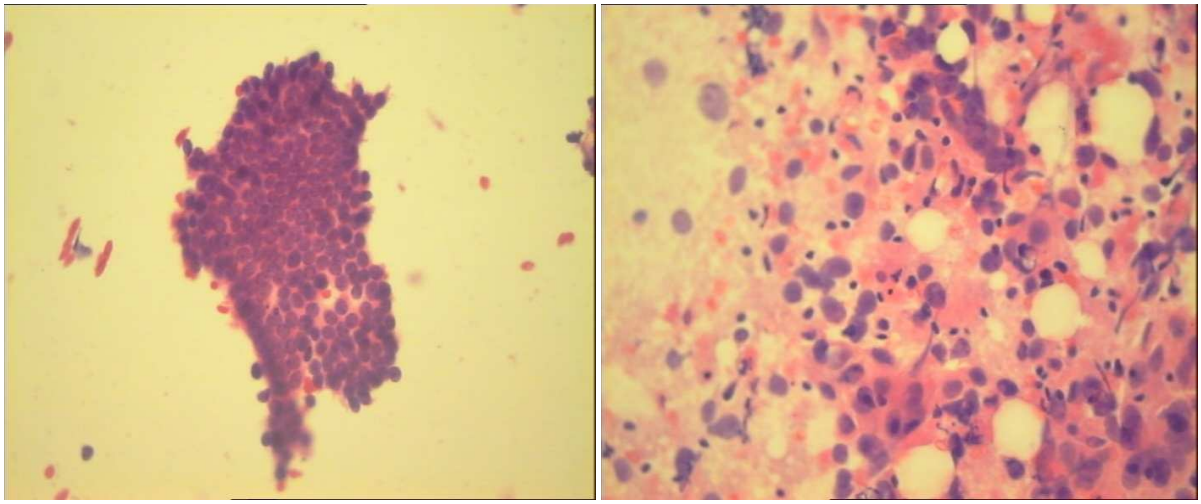

Fig. 1. FNB microscope images - benign case (left), malignant case(right).

3. FEATURE SELECTION ALGORITHM

3.1. FEATURES

Analyzing FNB images one might see that benign cells seem to be uniform in appearance. In the other hand, malignant cells are distinguished by much bigger diversity in shape and texture. In our research we have tried to find features best describing the differences between two sorts of cells.

Table 1. Features extracted from images.

Feature	Description
area	the actual number of pixels of the nucleus
perimeter	the distance around the boundary of the nucleus
eccentricity	the ratio of the distance between the foci of the ellipse and its major axis length
major axis length	the length of the major axis of the ellipse that has the same normalized second central moments as the region
minor axis length	the length of the minor axis of the ellipse that has the same normalized second central moments as the region
luminance gradient sum	the sum of luminance gradients in the image of the nucleus
luminance mean	the mean of luminance in the image of the nucleus
luminance variance	the variance of luminance gradients in the image of the nucleus
distance from the centroid	the Euclidean distance between the geometric center of the nucleus and mean of geometric centers of all the nuclei in the image

For each cell following features have been extracted: area, perimeter, eccentricity, major axis length, minor axis length, luminance gradient sum, luminance mean, luminance variance and distance from the centroid of all nuclei on the image. Detailed description of each used feature is delivered in table 1.

Features of the nuclei can be extracted from the image after the nuclei are correctly segmented. In parallel to presented studies, research is being carried out to develop the nuclei segmentation system using fuzzy clustering with shape constraints, active contours and region growing methods [2, 4, 6, 7, 11, 14, 17, 19]. The accuracy of the segmentation process obtained in developed approaches is promising, however current methods are not able handle properly overlapped nuclei [2, 7, 8, 11]. In order to eliminate such segmentation inaccuracy, it was decided to use during feature selection procedure reference images that was manually segmented (Fig. 2).

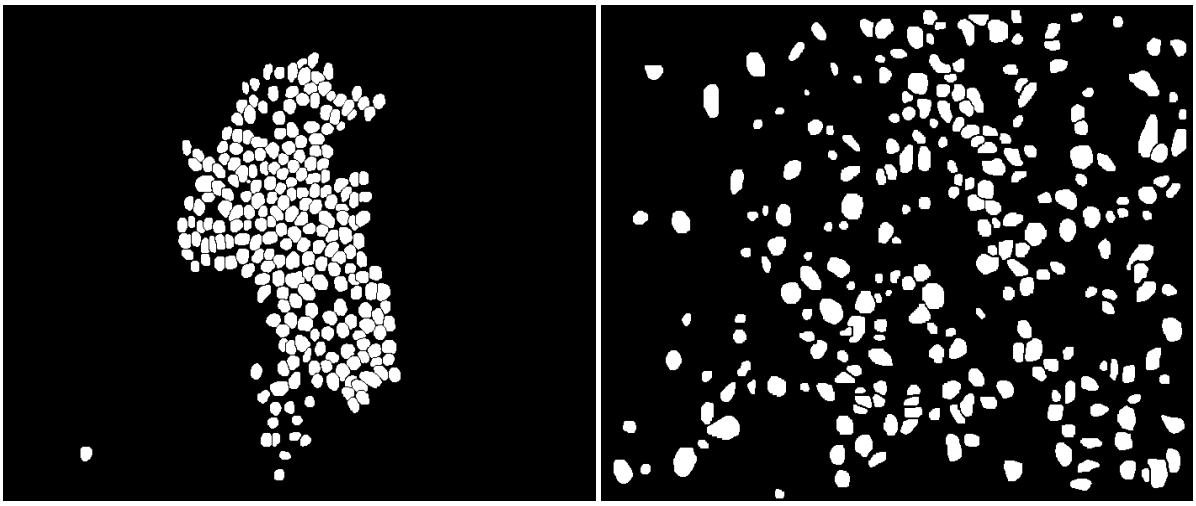

Fig. 2. Manually segmented FNB images - benign case (left), malignant case (right).

Such an approach allows us to carry out feature selection experiments independent of inaccuracies in the segmentation process. Of course, works on segmentation overlapped cell nuclei are well advanced and soon results obtained during automatic segmentation procedure will be used to select discriminant features.

3.2. CLASSIFICATION METHODS

A set of 4 different classifiers was used to test the effectiveness of the features in diagnosing new samples. It was decided to use well known classification algorithms such as k-nearest neighbor (with k = 9), naive Bayes classifier (with normal kernel distribution), decision trees (with GINI criterion) and classifiers ensemble [1, 9]. The idea behind using such number of classification techniques was to check how the method can influence the classification accuracy. However, it must be mentioned that ensemble of classifiers is not a separate classification technique and its classification procedure is based on the results of others classifiers used in the experiments. Simply, the answer of classifiers ensemble is determined by voting procedure and class that gathers majority vote wins and represents the answer of the classifiers ensemble.

Classifiers inputs are formulated as different statistics calculated for features presented in table 1. Each single case is described by a collection of structures that store features computed for all nuclei extracted from the image. Statistics such as mean, median and variance were computed for each single case. All input variables were normalized (scaled) to the range 0 to 1 in order to eliminate the effect of different variable ranges in the Euclidean space. Classifiers outputs were declared by fixed labels that describe the malignant or benign case. In figure 1 sample feature space with 3 input variables is presented.

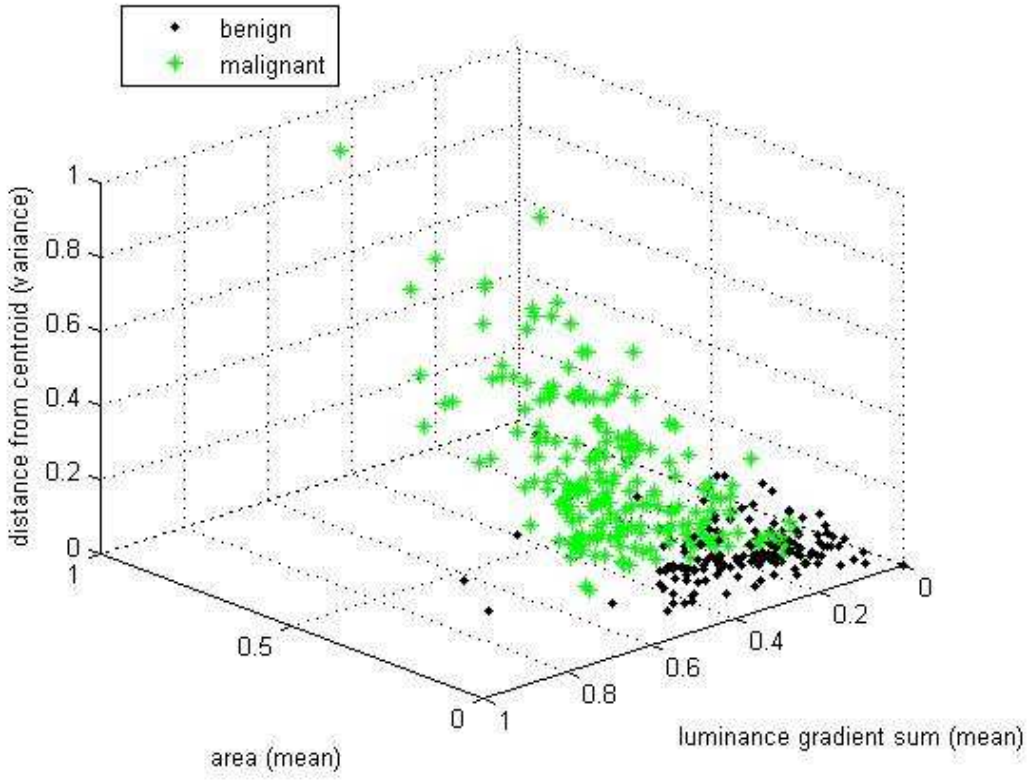

Fig. 3. Sample feature space.

The prospective accuracy of the resulting classifiers was tested using the leave-one-out validation technique. In this approach, if N samples are available, N partitions are formed by leaving one single pattern for testing, and using the remaining $N - 1$ to build the classifier. The N performance results obtained this way is then averaged and gives an accurate and unbiased estimate of the method's prospective accuracy. It is a measure of generalization ability of classifier (generalization to unseen

samples). Since the number of samples is relatively small, using chosen classification algorithms with leave-one-out is computationally tractable and allows for accurate estimation of the error.

4. EXPERIMENTAL RESULTS

In order to illustrate the effectiveness of proposed feature selection procedure an experimental results was collected and presented in the tables below. The discriminative power of individual features was estimated with previously indicated classifiers and results in form of recognition rates are presented in Table 2. Recognition rate is defined as percentage of successfully recognized cases to the total number of all cases. Images described in section 2 were used during the experiments but it must be pointed out that each image was treated as a separate case in recognition rate computation procedure.

Table 2. Recognition rates for single input variables.

Feature	Statistics	kNN	Naive Bayes	Decision trees	Ensemble classifiers
area	mean	81.9	84.7	82.5	83.9
	variance	86.4	84.4	82.8	86.4
	median	84.2	83.3	78.9	84.2
perimeter	mean	84.4	84.4	81.4	83.6
	variance	83.3	83.3	80.0	83.3
	median	81.6	79.9	78.2	81.0
eccentricity	mean	52.5	56.1	50.6	50.0
	variance	52.8	57.2	54.4	54.7
	median	50.8	55.9	52.7	53.1
major axis length	mean	81.4	83.6	80.3	81.9
	variance	79.7	80.6	77.2	78.9
	median	78.2	81.3	75.5	79.2
minor axis length	mean	83.9	83.6	79.4	83.6
	variance	83.9	83.9	83.6	84.2
	median	81.6	82.0	81.5	82.0
luminance gradient sum	mean	67.8	67.5	66.1	69.2
	variance	54.2	56.4	54.2	56.9
	median	54.1	55.2	54.0	55.1
luminance mean	mean	67.8	67.5	66.1	69.2
	variance	64.1	67.5	61.4	65.8
	median	64.1	67.5	61.4	65.8
luminance variance	mean	54.2	56.4	54.2	56.9
	variance	55.0	54.7	56.4	56.4
	median	55.5	54.1	55.9	55.5
distance from centroid	mean	78.3	79.4	76.7	76.7
	variance	75.3	76.7	70.8	75.0
	median	75.0	76..4	69.8	75.4

Table 3. Classification results after feature selection (bold indicates the best result achieved for the classifier).

Input variables feature (statistic)	kNN	Naive Bayes	Decision trees	Ensemble classifiers
area (mean), area (variance), perimeter (mean), luminance gradient sum (mean), luminance mean (mean), luminance variance (mean), major axis length (variance), minor axis length (variance), distance from centroid (mean), distance from centroid (variance).	**93.1**	91.4	87. 8	91. 7

area (mean), luminance gradient sum (mean), major axis length (variance), distance from centroid (variance), distance from centroid (mean).	92.2	**91.7**	90.6	92. 8
area (mean), luminance gradient sum (mean), major axis length (variance), distance from centroid (variance).	90.0	91.1	**91.2**	92.8

In order to reduce the dimensionality of feature space, sequential forward selection was applied. Ignoring redundant and irrelevant features leads to great improvement in recognition rates. Taking into account the fact that different subsets can be optimal for different classifiers, two approaches was applied to forward selection. First consider the same subset of features for each classifier and classifiers ensemble was used to assess the final quality of subset, and second approach assumes that each classifier has specific optimal subset of features. The latter approach allows for a slight improvement in the classification results. Comparison of best subsets of features for each specific classification method is presented in Table 3.

The best classification rate (93.33%) was obtained for ensemble classifiers using best specific subset of features for each classifier. The recognition rate about 93% seems to be very promising taking into account the preliminary nature of conducted investigations.

5. CONCLUSIONS

The main objective of the described work was to develop an automatic feature selection system for breast cancer malignancy classification problems. The results achieved in the experiments seem to be very promising. So far inspections of the segmented nuclei showed big differences in size and color between benign and malignant cases. Hence, there are three challenges for the near future. First, the recognition rate should be improved by adding more sophisticated features not tested during current investigations. As a second challenge, the proposed approach must be applied for automatically segmented images. So, previously developed segmentation algorithms must be extended to deal properly with overlapped cells. Finally, the whole segmentation and classification system will be applied for virtual slides generated by virtual scopes which are able to produce images with the resolution of 50000x50000 or even higher [3]. Such huge slides require a long analysis, respectively so it will be very helpful if automatic system can recognize suspected fragments of the slide and automatically present them in the first place.

BIBLIOGRAPHY

[1] BREIMAN L., FRIEDMAN J., STONE C.J., OLSHEN R.A., Classification and Regression Trees, Chapman & Hall, Boca Raton, 1993.
[2] HREBIEŃ M., STEĆ P., OBUCHOWICZ A., NIECZKOWSKI T., Segmentation of breast cancer fine needle biopsy cytological images. Int. J. Appl. Math and Comp. Sci. Vol. 18, No. 2, 2008, pp. 159–170.
[3] HUANG H.K., PACS and Imaging Informatics: Basic Principles and Applications, John Wiley & Son, New Jersey, 2010.
[4] JELEŃ Ł., FEVENS T., KRZYŻAK A., Classification of breast cancer malignancy using cytological images of fine needle aspiration biopsies, Int. J. Appl. Math and Comp. Sci. Vol. 18, No. 1, 2008, pp. 75–83.
[5] JELEŃ Ł., FEVENS T., KRZYŻAK A., JELEŃ M., Discriminatory Power of Cells Grouping Features for Breast Cancer Malignancy Classification, Proc. 4th Int. Conf. on Biomedical Engineering, Kuala Lumpur, 2008, pp. 559–562.
[6] KAWA J., PIĘTKA E., Image Clustering with Median and Myriad Spatial Constraint Enhanced FCM, Proc. 4th Int. Conf. on Computer Recognition Systems CORES' 05, Springer, 2005, pp. 211–218.

[7] KOWAL M., KORBICZ J., Segmentation of breast cancer fine needle biopsy cytological images using fuzzy clustering. In Kornacki J., Raś Z, Wierzchoń S.T., Kacprzyk J. (Eds.) Advances in Machine Learning I, Springer–Verlag, Berlin – Heidelberg, 2010, pp. 405–417.
[8] MARCINIAK A., OBUCHOWICZ A., MONCZAK R., KOŁODZIŃSKI M., Cytomorphometry of Fine Needle Biopsy Material from the Breast Cancer, Proc. 4th Int. Conf. on Computer Recognition Systems CORES' 05, Springer, 2005, pp. 603–609.
[9] MITCHELL T.M., Machine Learning. McGraw–Hill, 1997.
[10] NEZAFAT R., TABESH A., AKHAVAN S., LUCAS C., ZIA M., Feature selection and classification for diagnosing breast cancer, Proc. Int. Assoc. of Science and Technology for Development International Conference, Cancun, Mexico, 1998, pp. 310–313.
[11] OBUCHOWICZ A., HREBIEŃ M., NIECZKOWSKI T., MARCINIAK A., Computational intelligence techniques in image segmentation for cytopathology. In Smoliński T.G., Milanova M.G., Hassanien A.–G. (Eds.) Computational intelligence in biomedicine and bioinformatics : current trends and applications, Springer–Verlag, Berlin, 2008, pp. 169–199.
[12] SCHNORRENBERG F., PATTICHIS C., KYRIYRIACOU K., SCHIZAS C., Detection of cell nuclei in breast cancer biopsies using receptive fields, IEEE Proc. Engineering in Medicine and Biology Society, 1994, pp. 649–650.
[13] STREET N., Xcyt: A system for remote cytological diagnosis and prognosis of breast cancer, In Jain L. (Ed.), Soft Computing Techniques in Breast Cancer Prognosis and Diagnosis, World Scientific Publishing, Singapore, 2000, pp. 297–322.
[14] SURI J. S., SETAREHDAN K., SINGH S., Advanced Algorithmic Approaches to Medical Image Segmentation, Springer–Verlag, London, 2002.
[15] UNDERWOOD J.C.E., Introduction to biopsy interpretation and surgical pathology, Springer–Verlag, London, 1987.
[16] WALKER H. J., ALBERTELLI L., Breast cancer screening using evolved neural networks, Proc. IEEE Int. Conf. on Systems, Man and Cybernetics, San Diego, USA, 1998, pp. 1619–1624.
[17] WANG S–L., LAU W–H., LIEW A.W.C., LEUNG S–H., Robust Lip Region Segmentation for Lip Images with Complex Background. Pattern Recognition, Vol. 40, No. 12, 2007, pp. 3481–3491.
[18] WOLBERG W.H., STREET W.N., MANGASARIAN O.L., Breast cytology diagnosis via digital image analysis, Analytical and Quantitative Cytology and Histology, Vol. 15, 1993, pp. 396–404.
[19] ZOLLER T., VARIGOU, T.A., Robust image segmentation using resampling and shape constraints., IEEE Trans. on Pattern Analysis and Machine Intelligence. Vol. 29, No. 7, 2007, pp. 1147–1164.