Computer-Aided Breast Cancer Diagnosis Based on the Analysis of Cytological Images of Fine Needle Biopsies

Paweł Filipczuk, Thomas Fevens, Adam Krzyżak*, *Fellow, IEEE*, and Roman Monczak

***Abstract*—The effectiveness of the treatment of breast cancer depends on its timely detection. An early step in the diagnosis is the cytological examination of breast material obtained directly from the tumor. This work reports on advances in computer-aided breast cancer diagnosis based on the analysis of cytological images of fine needle biopsies to characterize these biopsies as either benign or malignant. Instead of relying on the accurate segmentation of cell nuclei, the nuclei are estimated by circles using the circular Hough transform. The resulting circles are then filtered to keep only high-quality estimations for further analysis by a support vector machine which classifies detected circles as correct or incorrect on the basis of texture features and the percentage of nuclei pixels according to a nuclei mask obtained using Otsu's thresholding method. A set of 25 features of the nuclei is used in the classification of the biopsies by four different classifiers. The complete diagnostic procedure was tested on 737 microscopic images of fine needle biopsies obtained from patients and achieved 98.51% effectiveness. The results presented in this paper demonstrate that a computerized medical diagnosis system based on our method would be effective, providing valuable, accurate diagnostic information.**

***Index Terms*—Breast cancer, classification, computer-aided diagnosis, pattern analysis.**

I. Introduction

ACCORDING to the International Agency for Research on Cancer, breast cancer is the most common cancer among women. In 2008, there were 1 384 155 diagnosed cases of breast cancer and 458 503 deaths caused by the disease worldwide [1], [2]. There has also been an increase in the incidence of breast cancer by 3%–4% a year since the 1980s. The effectiveness of treatment largely depends on the timely detection of the disease. An important and often used diagnostic method is the so-called triple-test, which is based on three medical examinations, that is used to achieve high confidence in the diagnosis [3]. The triple-test includes self examination (palpation), mammography or ultrasonography imaging, and fine needle biopsy (FNB) [4]. FNB is an invasive examination that consists in obtaining material directly from the tumor. The collected material is then examined under a microscope to determine the prevalence of cancer cells. This approach requires extensive knowledge and experience of the cytologist or pathologist responsible for the diagnosis. Automatic morphometric diagnosis can help make the results objective and assist inexperienced specialists. It also allows for screening on a large scale where only difficult and uncertain cases would require further examination by the specialist. Along with the development of advanced vision systems and computer science, quantitative cytopathology has become a useful method for detection of diseases, infections as well as many other disorders [5]–[9].

Recently there has been an increase in interest in computer-aided cytology and digital pathology. Several researchers have researched the analysis of cytological images of breast tumors, from studying the segmentation of the images, proposing new features, or considering various classification algorithms. However, only a few of these researchers have tested effectiveness of their methodology in a comprehensive computerized cancer classification system. Jeleń *et al.* [10] presented an approach based on the level sets segmentation method. Classification effectiveness was tested on 110 (44 malignant, 66 benign) images with results reaching 82.6%. Niwas *et al.* [11] presented a method based on the analysis of nuclei texture using a wavelet transform. Classification effectiveness with the k-nearest neighbor algorithm on 45 (20 malignant, 25 benign) images reached 93.33%. Another approach was presented by Malek *et al.* [12]. They used active contours to segment nuclei and classified 200 (80 malignant, 120 benign) images using fuzzy c-means algorithm achieving 95% effectiveness. Breast cancer diagnosis was also studied by Xiong *et al.* [13]. Partial least squares regression was used to classify 699 (241 malignant, 458 benign) images yielding 96.57% effectiveness. However, the authors did not describe the segmentation method used to extract nuclei.

This paper presents recent progress in the development of a comprehensive fully automatic breast cancer diagnostic system

Manuscript received May 01, 2013; revised July 12, 2013; accepted July 18, 2013. Date of publication July 29, 2013; date of current version November 25, 2013. The work of P. Filipczuk and R. Monczak was supported by the Polish National Science Centre under Grant NN518287440. The work of T. Fevens and A. Krzyżak was supported by the Natural Sciences and Engineering Research Council of Canada under Grant RGPIN 249849-2011 and Grant RGPIN 270-2010. *Asterisk indicates corresponding author.*

P. Filipczuk is with the Institute of Control and Computation Engineering, University of Zielona Góra, 65-246 Zielona Góra, Poland (e-mail: p.filipczuk@issi.uz.zgora.pl).

T. Fevens is with the Department of Computer Science and Software Engineering, Concordia University, Montreal, QC, H4B 1R6 Canada (e-mail: fevens@cs.concordia.ca).

*A. Krzyżak is with the Department of Computer Science and Software Engineering, Concordia University, Montreal, QC, H4B 1R6 Canada (e-mail: krzyzak@cs.concordia.ca).

R. Monczak is with the Department of Pathomorphology, Regional Hospital in Zielona Góra, 65-046 Zielona Góra, Poland (e-mail: r.monczak@issi.uz.zgora.pl).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2013.2275151

based on the analysis of cytological images of FNB material. The task at hand is to classify a case (FNB of a patient) as benign or malignant. This is done by using morphometric, textural and topological features of nuclei isolated from microscopic images of the tumor. In previous work [14]–[17], we used a segmentation method based on the combination of adaptive thresholding in grayscale and pixel classification in color space using one of four clustering algorithms: k-means, fuzzy c-means, Gaussian mixture model, and competitive neural networks. This previous approach did not work satisfactorily for a new challenging database of higher-resolution images, which were captured using entirely new technology. Also, due to the low resolution of the previous images we had not considered textural features.

The new images offer radically improved level of detail in comparison to the previous database of images. However, cytological images in the new database are very challenging with regards to nuclei segmentation. In contrast to the previous images, one can clearly observe clumps of chromatin (especially in malignant cases), nucleoli as well as nuclear membranes. A nucleus is no longer represented by nearly uniform pixels but rather appears as a much more complex object with fine structures and more varied texture. Therefore, attempts to adapt the previously used segmentation algorithms based on clustering combined with adaptive thresholding or the level set method did not yield satisfactory results for nuclei segmentation. Eventually we settled on the fast and robust approach proposed in this paper to determine the locations of cell nuclei based on circle detection using the Hough transform followed by circle filtration using a support vector machine designed to retain only those circles that are most likely to represent cell nuclei. From the selected circles, we extract a set of 25 features which are then tested by four different classifiers. The entire automatic diagnostic procedure was tested on microscopic images of fine needle biopsies achieving 98.51% effectiveness. The details of the conducted experiments and a full analysis of their outcomes are presented in the paper. The results presented in this paper demonstrate that a computerized medical diagnosis system based on our method would be effective and can provide valuable, accurate diagnostic information.

The paper is divided into seven sections. Section I presents an introduction into breast cancer diagnosis and outlines previous work. Section II describes the acquisition process of the medical images used for testing. Segmentation, feature extraction, and classification are described in Sections III, IV, and V, respectively. Section VI shows the experimental results obtained by the proposed method. The paper ends with our conclusions.

II. Medical Images Database

The testing database contains 737 images of the cytological material obtained by FNB. The material was collected from 67 patients (cases) at the Regional Hospital in Zielona Góra, Poland. Biopsies without aspiration were performed under the control of an ultrasonograph with a 0.5-mm-diameter needle. Smears from the material were fixed in spray fixative (Cellfix by Shandon) and dyed with hematoxylin and eosin $(\mathrm{h}+\mathrm{e})$. The time between preparation of smears and their preservation in fixative never exceeded three seconds. Cytological preparations were then digitalized into virtual slides using the Olympus VS120 Virtual Microscopy System. The system consists of a 2/3 in CCD camera and $40\times$ objective giving together $0.172\mu\mathrm{m}/\mathrm{pixel}$ resolution. The average size of the slides is approximately $200\,000 \times 100\,000$ pixels. The scans were prepared using Extended Focal Imaging (EFI). EFI is performed by scanning a preparation several times with the focus plane located at different positions along the Z axis. These frames are then compiled while only retaining these regions in each frame that are in sharp focus. This allows for extended focal depths impossible to obtain using optics alone. Next, on each slide a pathologist manually selected 11 distinct areas which were converted to 8 bit/channel RGB TIFF files of size 1583×828 pixels compressed with the lossless LZW algorithm. Note that the areas were not selected for the medical information, but only in terms of a sufficient amount of cytological material. The number of areas per patient used was recommended by the pathologists at the hospital and allows for a correct diagnosis. The database contains 25 (275 images) benign and 42 (462 images) malignant cases. All cancers were histologically confirmed and all patients with benign disease were either biopsied or followed for a year. Fig. 1 summarizes the acquisition process. Fig. 2 presents sample images of the $40\times$ enlargement and digitally obtained $10\times$ enlargement.

III. Nuclei Segmentation

To determine whether the tumor is benign or malignant, cell nuclei need to be isolated from the background and from other objects on the image (e.g., red blood cells). Then from the nuclei certain features can be extracted and the malignancy determined. Because all subsequent analysis is based on the results obtained in the segmentation step, it is very important that the nuclei were properly extracted. In the literature, many different approaches have been already proposed to extract cells or nuclei from microscope images [18]–[24]. This task is usually done automatically or semi-automatically, using one of the well known methods of image segmentation [25]–[27]. However, reliable nuclei segmentation is a challenging task. FNB images are particularly difficult due to the way they are prepared. The material is extracted by a needle and smeared on a glass. This may result in partial destruction of the tissue structure, and sometimes even of nuclei. The cells are usually not uniformly distributed on the preparation. They often form 3-D structures, and they can possibly be in contact with and/or occluded by other cells. Typically, there are many areas on the slides where even a human expert is unable to distinguish individual nuclei. Therefore, attempts to generalize the segmentation approaches proposed in literature usually fail because such methods work correctly only for specific types of images, preferably with clear disjoined nuclei. Moreover, the new set of images used in this study are far more detailed than the ones used in our previous work, so they provide more information. At the same time the fact that the individual components of nuclei are visible complicates the process of segmentation. Experimentation shows that neither our previous approach based on adaptive thresholding and clustering, nor algorithms based on marker-controlled watershed or level sets work properly on the new database (see Fig. 3).

In this paper, we propose a different approach that involves the estimation of nuclei by easily defined and identifiable shapes

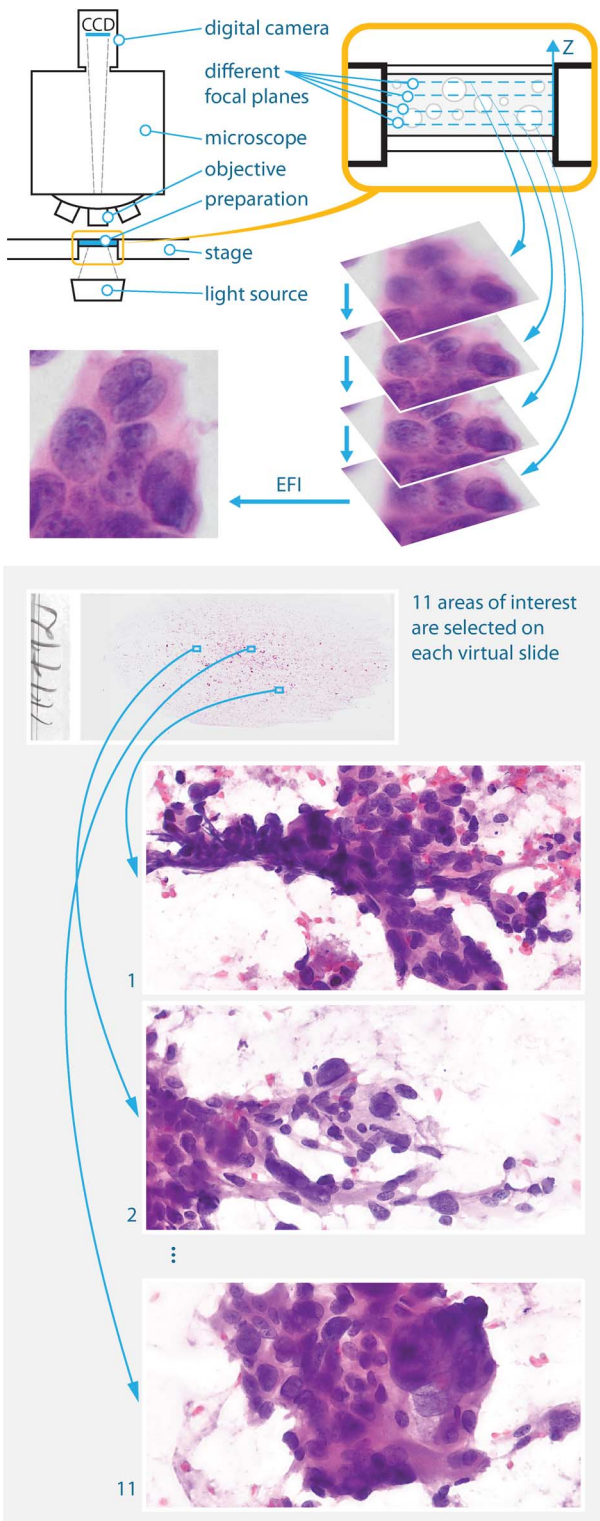

Fig. 1. Scanning process using extended focal imaging (EFI) (top, figure not to the scale) and sample virtual slide with areas of interest selection (bottom).

instead of attempting the precise determination of edges of cell nuclei which are difficult to obtain precisely and reliably. We rather focus on reliably determined nuclei and remove from consideration all dubious nuclei. This approach attempts to mimic the behavior of a pathologist who first tries to locate nuclei that

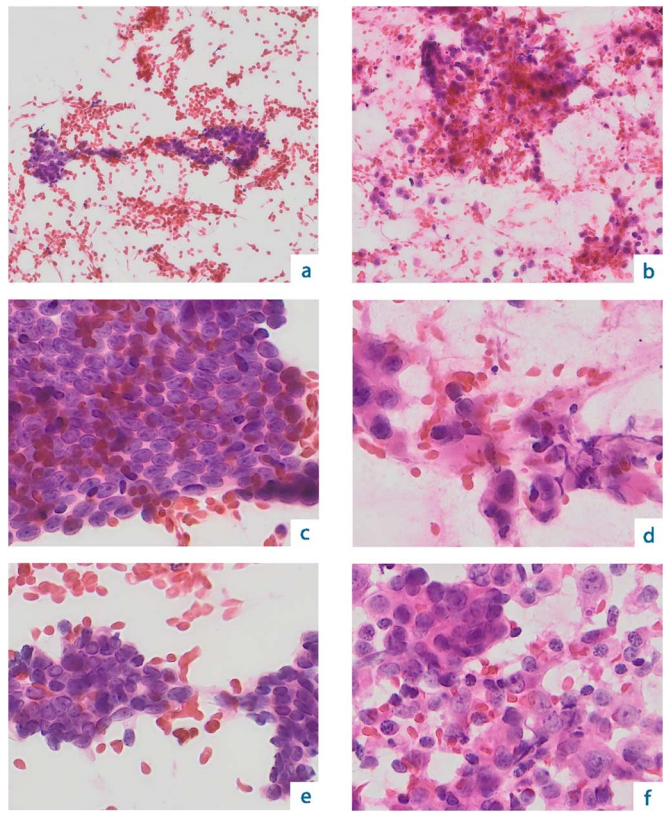

Fig. 2. Sample images: digitally obtained 10× enlargement (a), (b) and 40× enlargement (c)–(f). Images a, c and e are from a benign case, and images b, d, and f are from a malignant case.

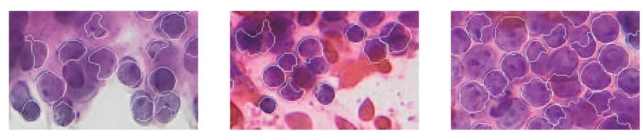

Fig. 3. Three examples of incorrectly segmented nuclei using the level set method initialized by the circular Hough transform detected circles.

are clearly visible and then uses his expert knowledge to determine whether the tumor is benign or malignant. In our approach, as a first estimation of the nuclei, we first search for as many candidates as possible with potentially good quality nuclei. This task is done using the Hough transform [28]. Considering the fact that most of the nuclei are approximately round, circles will be used in the Hough transform to detect nuclei. Ellipses might provide a better and more accurate approximation but the computational complexity of ellipse detection is significantly higher due to the fact that ellipse is described by five parameters while circle only by three. Then we remove false positives, red blood cells and touching nuclei using a quadratic discriminant trained on a set of 300 manually determined nuclei.

A. Circle Detection

Since most of the nuclei are circular or elliptical in shape, the circular Hough transform (CHT) [29], [30] is used to detect them in the images. The Hough transform was originally designed to detect lines, but later extended to the identification of arbitrary shapes [31]. The advantage of this method is its robustness to high levels of noise and irregularity of objects.

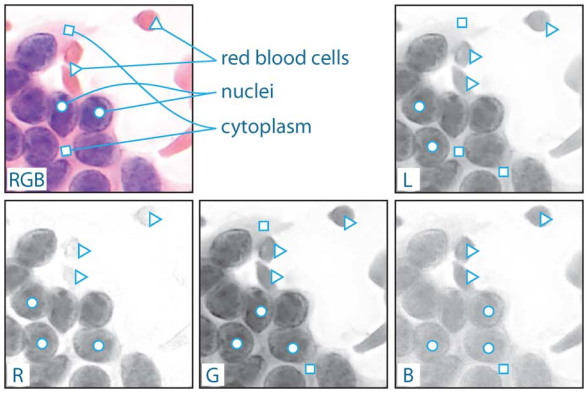

Fig. 4. Channels R, G, B, and L of sample section of a slide with marked nuclei (circles), red blood cells (triangles), and cytoplasm (squares). In the red channel, the difference between nuclei and red blood cell is the greatest and the cytoplasm is barely visible.

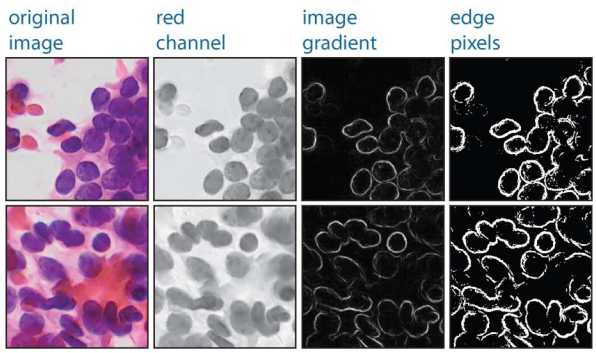

Fig. 5. Two examples of edge detection required for the circular Hough transform.

The first step of the algorithm is to detect edges in the image. Let I be a grayscale image. We define edge indicator E by

$$E = \begin{cases} 1, & \text{if } (\nabla I)^2 > t \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where ∇ is the gradient operator and t is the threshold value. The parameter t was chosen experimentally on the test set of images and set to 8. In the processing we use only the red channel where the difference in values between nuclei and red blood cells is the greatest. Also the cytoplasm, which surrounds the nuclei, is barely visible. Fig. 4 shows a sample excerpt from a slide in channels R, G, B, and the often used luminance channel L. Examples of detected edges are shown in Fig. 5.

In our application of CHT, we search for circle radii r in the range $1.72\ \mu\text{m} \leq r \leq 7.224\ \mu\text{m}$ with step size $\varepsilon = 0.172\ \mu\text{m}$. In pixels, this range corresponds to values $10 \leq r \leq 42$ and $\varepsilon = 1$. Fig. 6 presents examples of detected circles.

B. Circle Filtration

The variation of nuclear sizes is relatively high. As a result, sometimes a circle detected by CHT comprises two or more nuclei simultaneously. There are also other objects present in the images such as red blood cells. Although much brighter in the red channel than the nuclei, they are occasionally detected by the Hough transform. Another issue are false positives caused by, for example, geometric arrangements in the background being

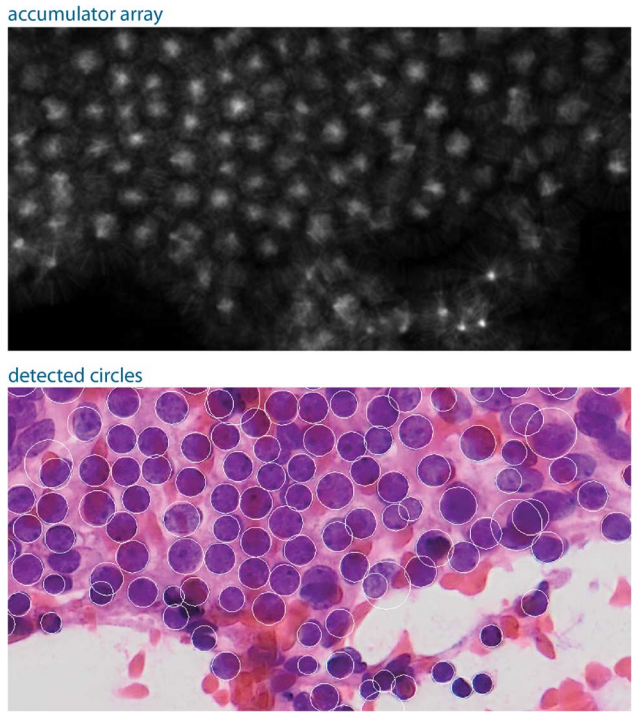

Fig. 6. Sample result of the circular Hough transform: the final accumulator array (top) and detected circles imposed on the original image (bottom).

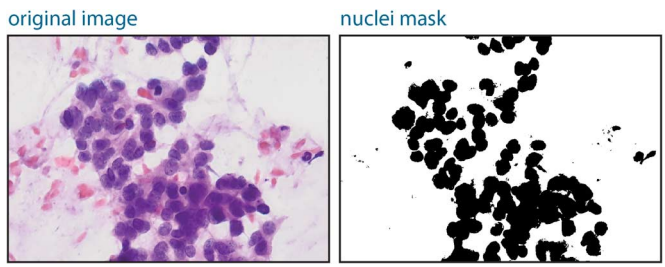

Fig. 7. Example of nuclei mask obtained by Otsu's thresholding: original image (left) and the nuclei mask (right).

incorrectly identified as the boundary of a nucleus. In order to remove all these nonnuclear objects we use a support vector machine as follows.

First, for all detected circles we calculate three features: the mean value of pixels inside the circle in the blue channel, long run high gray-level emphasis determined using gray-level run-length matrix (see Section IV for detailed description), and the percentage of nuclei pixels according to the nuclei mask. The nuclei mask is obtained by conducting Otsu's thresholding on the red channel of the image. The result is a binary mask where dark objects like nuclei are zeros and bright background pixels are ones. The final value of the feature is

$$\text{PNM} = \frac{n_{\text{mask}}}{n_{\text{all}}} \tag{2}$$

where n_{mask} is the number of pixels inside the circle, for which the mask value is 0, and n_{all} is the number of all pixels inside the circle. An example of nuclei mask is shown in Fig. 7. The circles are then classified as correct or incorrect using a support vector machine [32] with a Gaussian radial basis function kernel with scaling factor $\sigma = 0.8$. The classifier was trained on a manually

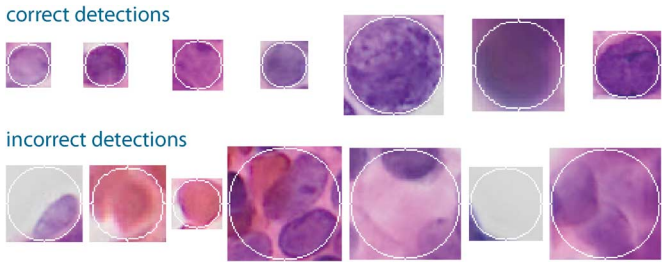

Fig. 8. Representative samples from 300 manually selected correct and incorrect circles obtained using the circular Hough transform. Incorrect cases include false positives, red blood cells, and joined or overlapped nuclei.

TABLE I
RESULTS OF CLASSIFICATION OF CIRCLES AS CORRECT OR INCORRECT USING VECTOR SUPPORT MACHINE ON A DATABASE OF 300 MANUALLY SELECTED CIRCLES. THREE FEATURES WERE USED: THE MEAN VALUE OF PIXELS ON THE BLUE CHANNEL, LONG RUN HIGH GRAY-LEVEL EMPHASIS DETERMINED USING GRAY-LEVEL RUN-LENGTH MATRIX, AND THE PERCENTAGE OF NUCLEI PIXELS ACCORDING TO THE NUCLEI MASK

Effectiveness	Conf. matrix		Sensitivity	Specificity
95.33%	139	11	0.93	0.98
	3	147		

prepared database of 300 circles. The database contained 150 properly detected nuclei and 150 incorrect detections, which included red blood cells, joined and overlapped nuclei, as well as false positives (see samples in Fig. 8).

The procedure described above was settled upon after an investigation with a series of experiments. First, the above mentioned database of 300 circles was prepared, and for all the circles 15 textural features were calculated, based on gray-level co-occurrence matrix and gray-level run-length matrix (see Section IV), the mean and variance of pixel values in each RGB channel, and the percentage of nuclei pixels according to the nuclei mask. Then five different classifiers were tested: k-nearest neighbor [33], naive Bayes classifier [34], [35], decision trees [36], support vector machines [32], and neural networks [37]. For each classifier, the sequential forward selection [38] was performed to find an optimal subset of features. The classification rate was obtained by using the leave-one-out cross-validation technique [39]. The recognition rate was the percentage of successfully recognized circles (as correct or incorrect) among all 300 circles. The best combination of a classifier and feature subset gave 95.33% effectiveness. The effectiveness is the percentage ratio of successfully classified circles, as correct or incorrect, to the total number of circles. Table I presents the detailed results.

The average number of nuclei per case (patient) after circle filtration is 490 with standard deviation 188. The worst case was represented by 146 nuclei which are still enough to make a diagnostic decision.

IV. FEATURE EXTRACTION

After the isolation of nuclei from the images, as determined by the circles classified as correct in the previous step, 50 global features are extracted and used in the classification procedure. First, for each nucleus we calculate all 25 features, described below. Note that each nucleus is represented in the features calculations by the pixels in the interior and on the boundary of the circle that determined it. For each of the 25 features, the mean and variance are determined giving a total of 50 global features. The mean and variance are computed either for a single image, or across all 11 images for a single case, depending on whether the classification is to be determined for an image or a case. The classification schemes are discussed in detail in Section V.

Rodenacker and Bengtsson [40] presented a comprehensive overview of features for cytological analysis. In our approach, the features chosen reflect the observations of cytologists. The key features associated with the diagnosis of breast cancer as used by specialists can be divided into three groups. The first group are features related to the size of the nuclei. A large variation of sizes in the image suggests malignancy of the tumor. Small uniform nuclei argue for a benign case. This group is represented by the area and perimeter of a nucleus. Another important feature is the distribution of chromatin in the nuclei of healthy cells. Frequent occurrences of distinct lumps of chromatin may indicate presence of cancer [41]. The second group of features represent this dependance with texture features based on gray-level co-occurrence matrix (GLCM) [42] and gray-level run-length matrix (GLRLM) [43], as well as the mean and variance of pixel values in each RGB channel. The last group of features is related to the distribution of nuclei in the image. Healthy tissues usually form single-layered structures, while cancerous cells tend to break up which increases the probability of encountering separated nuclei. To express this relation, we use features representing the distance to centroid of all nuclei, and the distance to c-nearest nuclei.

Below is a detailed description of all the features.

- *Area*: The actual number of pixels of the nucleus.
- *Perimeter*: The sum of distances between each adjoining pair of pixels around the border of the nucleus.
- *Distance to centroid of all nuclei*: The distance between the geometric center of the nucleus and centroid of all nuclei.
- *Distance to c-nearest nuclei* (distance to c-NN): Sum of distances between the geometric center of the nucleus and geometric centers of c-nearest nuclei; after conducting experiments with different values of c, we decided to set this parameter to 1,
- *Mean R value, mean G value, mean B value*: The mean value of pixels of the nucleus in channel R, G, and B, respectively,
- *Variance of R value, variance of G value, variance of B value*: The variance of pixel values of the nucleus in channel R, G, and B, respectively.

The next four textural features are co-occurrence features proposed by Haralick [42] calculated using GLCMs for 0°, 45°, 90°, and 135°, and eight gray-levels.

- *Contrast*: The intensity contrast between a pixel and its neighbor over the whole image.
- *Correlation*: The correlation of a pixel to its neighbor over the whole image.
- *Energy*: In literature also known as uniformity—the sum of squared elements in the GLCM.
- *Homogeneity*: The closeness of the distribution of elements in the GLCM to the GLCM diagonal.

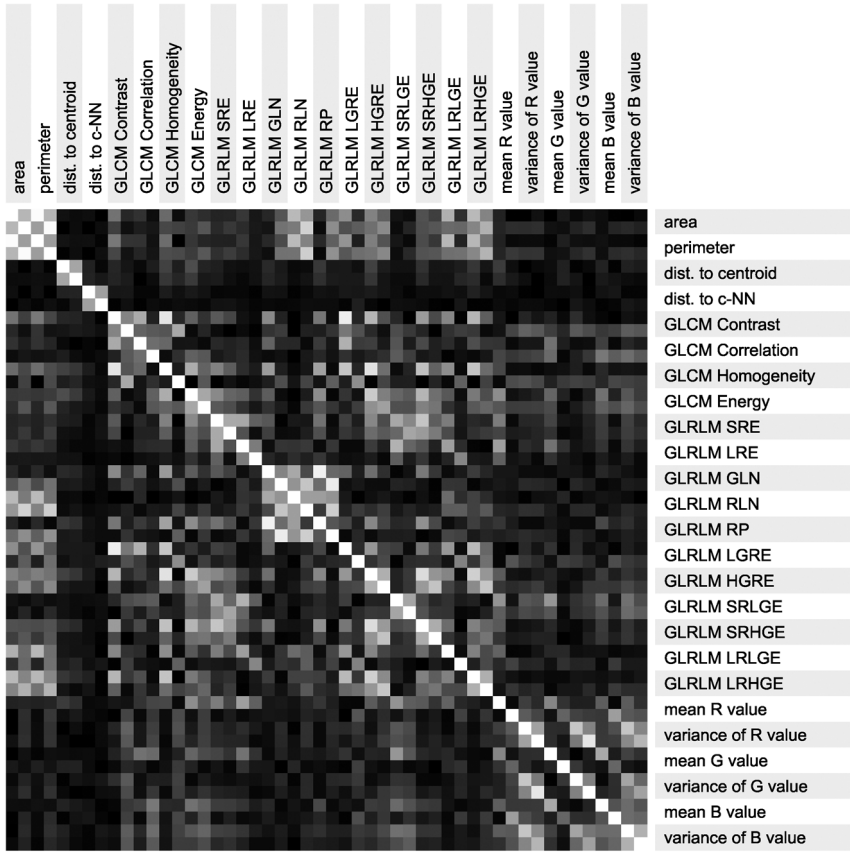

Fig. 9. Correlation matrix for all 50 global features determined for image classification scheme. $|Corr(X,Y)| = 1$ (white) means the highest correlation, 0 (black) no correlation. For readability both statistics, mean and variance, are described only by the name of the feature for which they were calculated (e.g., area is actually mean area and variance of area, perimeter is mean perimeter, and variance of perimeter, and so on).

The last eleven textural features are run-length features [43]–[45] calculated using GLRLMs for 0°, 45°, 90°, and 135°, and eight gray-levels:

SRE	Short run emphasis.
LRE	Long run emphasis.
GLN	Gray-level nonuniformity.
RLN	Run length nonuniformity.
RP	Run percentage.
LGRE	Low gray-level run emphasis.
HGRE	High gray-level run emphasis.
SRLGE	Short run low gray-level emphasis.
SRHGE	Short run high gray-level emphasis.
LRLGE	Long run low gray-level emphasis.
LRHGE	Long run high gray-level emphasis.

V. Classification

In our study, we considered the use of two different classification schemes.

1) *Image classification scheme*: In this approach, we calculate the mean and variance of the features for each image separately. The images are also classified individually. The final classification is obtained by majority vote of the results of the classification of the images belonging to a given patient.
2) *Patient classification scheme*: In this scheme, we calculate mean and variance of the features for patients, which means that we collect information from all 11 images representing a given case.

After determining the statistics, all input variables are standardized as follows:

$$z_i = \frac{x_i - \mu_i}{\sigma_i} \tag{3}$$

where μ_i is the mean value and σ_i is standard deviation of the feature i. Fig. 9 shows the absolute correlation matrix for all 50 global features determined for image classification scheme (for each of 25 features we calculated two statistics yielding 50 features). For classification we used four classification algorithms [35], [46]: *k-nearest neighbor* (k-NN) using $k = 5$, *naive Bayes classifier* (NB) using kernel density estimate, *decision tree* (DT), and *support vector machine* (SVM) using Gaussian radial basis function kernel with scaling factor $\sigma = 0.9$. Each of these classifiers represents a different approach to classification. The k-NN method associates objects to classes based on closest training examples in the feature space [33]. The Naive Bayes classifier is a probabilistic algorithm that applies Bayes' theorem [34], [35]. A decision tree is trained by splitting the

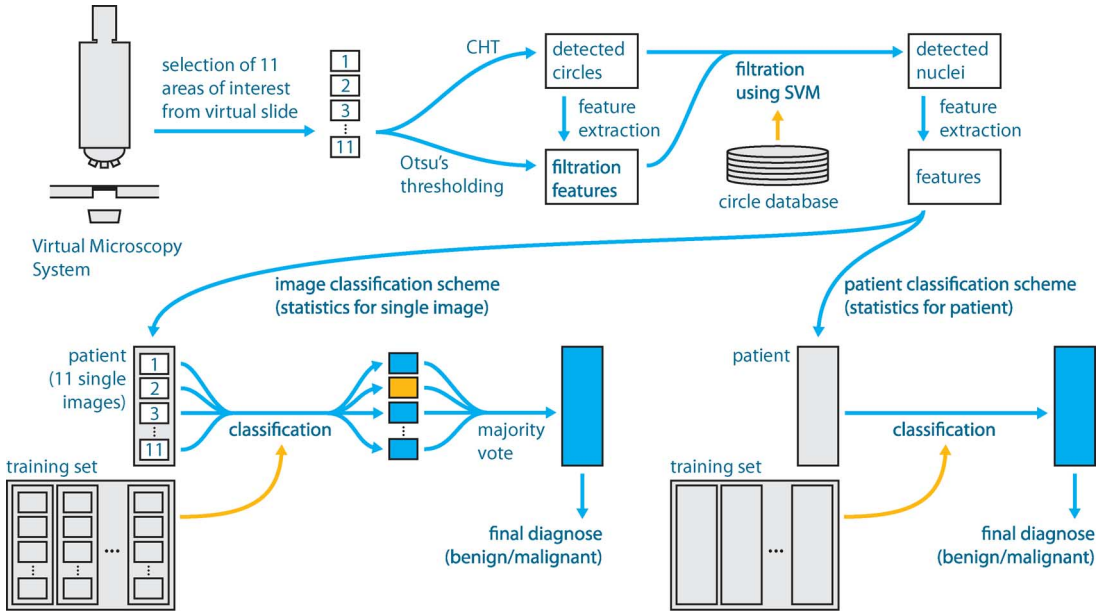

Fig. 10. Overview of the system.

training set into subsets based on an attribute value test [36]. Finally, the SVM constructs a hyperplane in a high-dimensional space that has the largest distance to the nearest training data point of any class maximum margin [32]. The parameters were chosen experimentally based on results of a full test procedure described in Section VI for various parameters of the classifiers (e.g., for k-NN we performed tests for $k = 1, 2, \ldots, 24, 25$, and chose one that gave the best result for given classification scheme).

VI. Experimental Results

The system was tested for the effectiveness which we specifically define as the percentage ratio of true positive and true negative cases to the total number of cases. There were 67 real medical cases (patients): 42 malignant and 25 benign. Each case was represented by 11 areas selected from its virtual slide (see Section II). From the areas, which were RGB images, nuclei were segmented using the method described in Section III. Then 25 features (see Section IV) were extracted from the nuclei. The classification was performed using four classification algorithms and two different schemes as described in Section V. Depending on the scheme, feature statistics were determined either for entire case or for single areas. In the image scheme, case classification was obtained by a majority voting of the classification of individual images belonging to the case (e.g., if six images were classified as benign and five as malignant, then the final result for the case would be benign). Complete system outline is shown in Fig. 10.

The initial set of candidate features presented in Section IV is relatively large. Feature selection can improve the classification accuracy and the reliability of the estimate of effectiveness [38]. In order to find the optimal subset of features for each classifier we applied the sequential forward selection algorithm [38]. In this method, the effectiveness is calculated for each single feature. The feature for which the result was the best is added to the optimal subset. Then, the rate is calculated for each remaining feature combined with the current optimal subset and the best one is included to the subset. The procedure is repeated until the addition of any feature does not improve the effectiveness. This approach does not guarantee finding the global optimum, but allows for finding a suboptimal set of features that gives satisfactory recognition results with a relatively small computational effort (as compared to an exhaustive search of all possible subsets of 50 global features).

First, we compared both classification schemes. The effectiveness was tested using the leave-one-out cross-validation technique [47], [48], which is a special case of the n-fold (when n equals the number of cases). It must be mentioned that for the image classification scheme a full set of 11 images representing one case was removed from the training set. This means the images belonging to the same case were never present at the same time in the training and testing set.

The following features have been selected by the sequential forward selection [the letter in brackets next to feature name indicates mean (m) or variance (v)]:

- *k-Nearest Neighbor*:
  - — image scheme: area (v), dist. to centroid (m), dist. to c-NN (m), variance of R value (m), GLRLM SRE (v), GLRLM LRLGE (v);
  - — patient scheme: area (v), dist. to c-NN (m), variance of R value (m), GLRLM SRLGE (m).
- *Naive Bayes*:
  - — image scheme: area (v), dist. to centroid (m), dist. to c-NN (m), GLCM contrast (v), GLCM correlation (v), GLCM homogeneity (v);
  - — patient scheme: dist. to centroid (m), dist. to c-NN (m), variance of R value (m), GLRLM LRLGE (v), GLRLM LRHGE (m).
- *Decision Tree*:
  - — image scheme: dist. to c-NN (m), mean of G value (v), variance of G value (m), GLCM contrast (v), GLRLM LGRE (v), GLRLM LRLGE (m), GLRLM LRHGE (v);

TABLE II
DETAILED RESULTS FOR BOTH CLASSIFICATION SCHEMES. MCC IS THE MATTHEWS CORRELATION COEFFICIENT

Patient Scheme

Classifier	Effectiveness	Sensitivity	Specificity	MCC
k-NN	**94.03%**	0.84	1.00	0.88
NB	**94.03%**	0.88	0.98	0.87
DT	91.04%	0.88	0.93	0.81
SVM	92.54%	0.88	0.95	0.84

Image Scheme

Classifier	Effectiveness	Sensitivity	Specificity	MCC
k-NN	**98.51%**	1.00	0.98	0.97
NB	**98.51%**	1.00	0.98	0.97
DT	97.01%	1.00	0.95	0.94
SVM	**98.51%**	1.00	0.98	0.97

— patient scheme: dist. to centroid (m), variance of R value (m), GLCM correlation (v), GLRLM HGRE (v).

- *Support Vector Machine*:
  — image scheme: dist. to centroid (m), dist. to c-NN (m), GLRLM LGRE (v), GLRLM LRHGE (v);
  — patient scheme: area (v), variance of R value (m), GLCM contrast (m), GLCM correlation (v).

Analyzing the subsets of the features it can be seen that in most cases they include a representative from each group of features described in Section IV. There is always the distance to c-nearest nuclei describing the distribution of nuclei in the image. The choice of the texture features does not seem so obvious but clearly the texture features provide important information and each from the determined subsets includes them.

Table II shows the classification results for both classification schemes. The patient scheme gave results that were around 5%–6% less effective. This may be due to the fact that the areas taken from the virtual slides were selected regardless of the diagnostic information they contain. Single images may not provide sufficient information and cause noise in the patient classification scheme. On the other hand, the image scheme better reflects the behavior of a medical expert.

We compared the method presented in this paper with our previous approaches: the combination of adaptive thresholding and clustering using the k-means (KM) [14] and Gaussian mixture model (GM) [16]. Classification was performed for image classification scheme using leave-one-out cross-validation. Table III presents the results, which can be compared with the results obtained using our new approach presented in Table II (image scheme). We also tested the methods using n-fold cross-validation for $n = 2$, 5, 10, and 33. Since in n-fold the samples are divided into n training and testing sample subsets randomly, for each classifier and for each n we performed 150 runs. The mean of the effectiveness and standard deviation are presented in Table IV.

The results show that the proposed method gives significantly better results. In the conducted experiments, all tested classification algorithms gave comparable satisfactory effectiveness for the method. However, the decision tree was found to be very sensitive to the size of the training and testing sets. For low n, in n-fold cross validation, when the training set is relatively small, the decision tree delivered a surprisingly low effectiveness rate.

TABLE III
DETAILED RESULTS FOR OUR PREVIOUS APPROACHES; THE COMBINATION OF ADAPTIVE THRESHOLDING AND CLUSTERING USING THE K-MEANS (KM) [14] AND GAUSSIAN MIXTURE MODEL (GM) [16]. CLASSIFICATION WAS PERFORMED FOR IMAGE CLASSIFICATION SCHEME. MCC IS THE MATTHEWS CORRELATION COEFFICIENT

k-Nearest Neighbor

Method	Effectiveness	Sensitivity	Specificity	MCC
KM	**89.55%**	0.72	1.00	0.79
GM	82.09%	0.56	0.98	0.62

Naive Bayes

KM	**92.54%**	0.84	0.98	0.84
GM	88.06%	0.72	0.98	0.75

Decision Tree

KM	**95.52%**	0.96	0.95	0.91
GM	92.54%	0.80	1.00	0.85

Support Vector Machine

KM	**94.03%**	0.88	0.98	0.87
GM	85.07%	0.76	0.90	0.68

TABLE IV
MEAN AND STANDARD DEVIATION (STD) OF THE CLASSIFICATION RESULTS OBTAINED IN 150 RUNS OF N-FOLD CROSS-VALIDATION FOR $n = 2$, 5, 10, AND 33 USING THE PROPOSED METHOD (CHT) AND OUR TWO PREVIOUS APPROACHES (KM, GM)

k-Nearest Neighbor

Method	Statistic	$n = 2$	$n = 5$	$n = 10$	$n = 33$
CHT	mean	**91.5522**	**96.0697**	**97.5323**	**98.4378**
	std	3.5134	1.9054	1.2823	0.3601
KM	mean	80.0896	84.3582	87.0746	89.4925
	std	3.7957	3.0036	1.9779	0.8797
GM	mean	73.0746	77.1542	80.2488	82.3980
	std	3.1644	2.4206	1.8252	0.6305

Naive Bayes

CHT	mean	**90.0597**	**93.9204**	**96.0398**	**97.8308**
	std	5.1669	2.4704	1.8462	0.9406
KM	mean	85.0647	88.8458	90.7363	92.2189
	std	4.3247	2.6264	1.7404	0.7252
GM	mean	83.1642	84.4677	85.6617	87.4925
	std	2.5887	1.9107	1.5839	0.8413

Decision Tree

CHT	mean	77.0846	81.0348	**85.0846**	**89.5423**
	std	5.0121	4.1664	3.6045	3.1245
KM	mean	77.0050	79.8507	81.4030	88.4975
	std	5.4493	3.9536	3.8806	2.6313
GM	mean	**80.0000**	**82.2189**	82.8060	85.4826
	std	4.1824	4.1929	3.6995	3.3122

Support Vector Machine

CHT	mean	**93.2438**	**95.5224**	**97.1841**	**98.4279**
	std	3.3203	1.6943	1.3999	0.3365
KM	mean	87.0050	89.9403	91.3831	93.3632
	std	3.8610	2.4420	1.7983	1.0873
GM	mean	80.2786	81.9403	83.5124	84.7363
	std	3.6296	2.4255	1.8081	0.7762

The obtained effectiveness reaching 98.51% is an indication of the effectiveness of the presented approach as a diagnostic tool.

VII. Conclusion

This paper describes a new approach for computer-aided breast cancer diagnosis. The task is to automatically distinguish benign and malignant cases. The approach is based on the analysis of cytological images of FNB. Due to the fact that most of the current segmentation methods do not work properly on the new very high resolution images used in this study, we decided to dispense with accurate segmentation in favor of estimating cell nuclei by circles. For this purpose, the circular Hough transform was used together with subsequent removal of incorrect or less reliable detections using a SVM-based procedure. The presented solution allows for removal of unreadable areas from consideration and allows for the determination of features based on only certain high-quality isolated nuclei. This, together with the proposed features and classifiers gave very good results. The best obtained effectiveness reached 98.51% indicating that the presented method is effective and capable of providing valuable diagnostic information.

Future research will focus on the automatic selection of regions of interest in full virtual slides and whole slide analysis. Another problem which warrants attention is usage of ellipses that seem to provide more accurate model for nuclei, particularly for malignant cells. However, ellipse detection is much more computationally demanding than circles and will require finding a quicker and more robust algorithm than used in this paper. Finally, we would like to improve classification process by applying the adaptive splitting and selection algorithm presented in [49].

Acknowledgment

We would like to thank the reviewers for valuable comments and suggestions. P. Filipczuk is a scholar within Sub-measure 8.2.2: Regional Innovation Strategies, Measure 8.2: Transfer of knowledge, Priority VIII: within the Regional human resources for the economy Human Capital Operational Programme co-financed by the European Social Fund and the state budget.

References

[1] J. Ferlay, H. Shin, F. Bray, D. Forman, C. Mathers, and D. Parkin, Globocan 2008 v2.0, Cancer Incidence and Mortality Worldwide: Iarc Cancerbase Int. Agency Res. Cancer, Lyon, France, Aug. 30, 2012 [Online]. Available: http://globocan.iarc.fr

[2] F. Bray, J. Ren, E. Masuyer, and J. Ferlay, "Estimates of global cancer prevalence for 27 sites in the adult population in 2008," *Int. J. Cancer*, Jul. 2012.

[3] P. Britton, S. Duffy, R. Sinnatamby, M. Wallis, S. Barter, M. Gaskarth, A. O'Neill, C. Caldas, J. Brenton, P. Forouhi, and G. Wishart, "One-stop diagnostic breast clinics: How often are breast cancers missed?," *Br. J. Cancer*, pp. 1873–1878, Jun. 2009.

[4] J. C. E. Underwood, *Introduction to Biopsy Interpretation and Surgical Pathology*. London, U.K.: Springer-Verlag, 1987.

[5] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, "Histopathological image analysis: A review," *IEEE Rev. Biomed. Eng.*, vol. 2, pp. 147–171, 2009.

[6] J. Śmietański, R. Tadeusiewicz, and E. Łuczyńska, "Texture analysis in perfusion images of prostate cancer-a case study," *Int. J.Appl. Math. Comput. Sci.*, vol. 20, no. 1, pp. 149–156, 2010.

[7] M. R. Hassan, M. M. Hossain, R. K. Begg, K. Ramamohanarao, and Y. Morsi, "Breast-cancer identification using HMM-fuzzy approach," *Comput. Biol. Med.*, vol. 40, pp. 240–251, 2010.

[8] O. Lezoray, A. Elmoataz, and H. Cardot, "A color object recognition scheme: Application to cellular sorting," *Mach. Vis. Appl.*, vol. 14, no. 3, pp. 166–171, 2003.

[9] M. Plissiti, C. Nikou, and A. Charchanti, "Automated detection of cell nuclei in Pap smear images using morphological reconstruction and clustering," *IEEE Trans. Inf. Technol. Biomed.*, vol. 15, no. 2, pp. 233–241, Feb. 2011.

[10] L. Jeleń, T. Fevens, and A. Krzyżak, "Classification of breast cancer malignancy using cytological images of fine needle aspiration biopsies," *Int. J. Appl. Math. Comput. Sci.*, vol. 18, no. 1, pp. 75–83, 2010.

[11] I. S. Niwas, P. Palanisamy, and K. Sujathan, "Wavelet based feature extraction method for breast cancer cytology images," in *Proc. 2010 IEEE Symp. Indust. Electron. Appl.*, 2010, pp. 686–690.

[12] J. Malek, A. Sebri, S. Mabrouk, K. Torki, and R. Tourki, "Automated breast cancer diagnosis based on GVF-Snake segmentation, wavelet features extraction and fuzzy classification," *J. Signal Process. Syst.*, vol. 55, pp. 49–66, 2009.

[13] X. Xiong, Y. Kim, Y. Baek, D. W. Rhee, and S.-H. Kim, "Analysis of breast cancer using data mining & statistical techniques," in *Proc. 6th Int. Conf. Software Eng., Artif. Intell., Netw. Parallel/Distribut. Comput. 1st ACIS Int. Worksh. Self-Assemb. Wireless Netw.*, 2005, pp. 82–87.

[14] P. Filipczuk, M. Kowal, and A. Obuchowicz, "Automatic breast cancer diagnostis based on k-means clustering and adaptive thresholding hybrid segmentation," in *Image Processing and Communications Challenges 3*, ser. Advances in Intelligent and Soft Computing: 102, R. S. Choraś, Ed. Berlin, Germany: Springer-Verlag, 2011, pp. 295–303.

[15] P. Filipczuk, M. Kowal, and A. Obuchowicz, "Fuzzy clustering and adaptive thresholding based segmentation method for breast cancer diagnosis," in *Computer Recognition Systems 4*, ser. Advances in Intelligent and Soft Computing: 95, M. W. Red, R. Burduk, M. Kurzyński, and A. Żołnierek, Eds. Berlin, Germany: Springer-Verlag, 2011, pp. 613–622.

[16] M. Kowal, P. Filipczuk, A. Obuchowicz, and J. Korbicz, "Computer-aided diagnosis of breast cancer using Gaussian mixture cytological image segmentation," *J. Med. Inf. Technol.*, vol. 17, pp. 257–262, 2011.

[17] M. Kowal, P. Filipczuk, and J. Korbicz, "Hybrid cytological image segmentation method based on competitive neural network and adaptive thresholding," *Pomiary, Automatyka, Kontrola*, vol. 57, no. 11, pp. 1448–1451, 2011.

[18] E. Meijering, "Cell segmentation: 50 years down the road," *IEEE Signal Process. Mag.*, vol. 29, no. 5, pp. 140–145, May 2012.

[19] Y. Al-Kofahi, W. Lassoued, W. Lee, and B. Roysam, "Improved automatic detection and segmentation of cell nuclei in histopathology images," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 4, pp. 841–852, Apr. 2010.

[20] W. F. Clocksin, "Automatic segmentation of overlapping nuclei with high background variation using robust estimation and flexible contour models," in *Proc. 12th Int. Conf. Image Anal. Process.*, 2003, pp. 682–687.

[21] F. Cloppet and A. Boucher, "Segmentation of overlapping/aggregating nuclei cells in biological images," in *Proc. IEEE Int. Conf. Pattern Recognit.*, Dec. 2008, pp. 1–4.

[22] M. Hrebień, P. Steć, A. Obuchowicz, and T. Nieczkowski, "Segmentation of breast cancer fine needle biopsy cytological images," *Int. J. Appl. Math. Comput. Sci.*, vol. 18, no. 2, pp. 159–170, 2010.

[23] Y. Peng, M. Park, M. Xu, S. Luo, J. S. Jin, Y. Cui, F. W. S. Wong, and L. D. Santos, "Clustering nuclei using machine learning techniques," in *Proc. Int. IEEE/ICME Conf. Complex Med. Eng.*, 2010, pp. 52–57.

[24] M. Plissiti and C. Nikou, "Overlapping cell nuclei segmentation using a spatially adaptive active physical model," *IEEE Trans. Image Process.*, vol. 21, no. 11, pp. 4568–4580, 2012.

[25] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 3rd ed. Upper Saddle River, NJ: Prentice Hall, 2007.

[26] S. Naz, H. Majeed, and H. Irshad, "Image segmentation using fuzzy clustering: A survey," in *Proc 6th Int. Conf. Emerg. Technol.*, 2010, pp. 181–186.

[27] J. S. Suri, K. Setarhdan, and S. Singh, *Advanced Algorithmic Approaches to Medical Image Segmentation*. London, U.K.: Springer-Verlag, 2002.

[28] R. Duda and P. Hart, "Use of the Hough transformation to detect lines and curves in pictures," *Commun. ACM*, vol. 15, no. 1, pp. 11–15, Jan. 1972.

[29] D. Kerbyson and T. Atherton, "Circle detection using Hough transform filters," in *Proc. 5th Int. Conf. Image Process. Appl.*, U.K., 1995, pp. 370–374.
[30] H. Kälviäinen, P. Hirvonen, L. Xu, and E. Oja, "Probabilistic and non-probabilistic Hough transforms: Overview and comparisons," *Image Vis. Comput.*, vol. 13, no. 4, pp. 239–252, 1995.
[31] D. Ballard, "Generalizing the Hough transform to detect arbitrary shapes," *Pattern Recognit.*, vol. 13, no. 2, pp. 111–122, 1981.
[32] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
[33] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967.
[34] T. Mitchell, *Machine Learning*. New York: McGraw-Hill, 1997.
[35] C. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.
[36] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Monterey, CA: Wadsworth Brooks/Cole Advanced Books Software, 1984.
[37] G. Zhang, "Neural networks for classification: A survey," *IEEE Trans. Syst., Man, Cybern.*, vol. 30, no. 4, pp. 451–462, Nov. 2000.
[38] M. Kudo and J. Sklansky, "Comparison of algorithms that select features for pattern classifiers," *Pattern Recognit.*, vol. 33, no. 1, pp. 25–41, 2000.
[39] P. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*. London, U.K.: Prentice-Hall, 1982.
[40] K. Rodenacker and E. Bengtsson, "A feature set for cytometry on digitized microscopic images," *Anal. Cell. Pathol.*, vol. 25, no. 1, pp. 1–36, 2003.
[41] P. Filipczuk, T. Fevens, A. Krzyżak, and A. Obuchowicz, "GLCM and GLRLM based texture features for computer-aided breast cancer diagnosis," *J. Med. Informat. Technol.*, vol. 19, pp. 109–115, 2012.
[42] R. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. 3, no. 6, pp. 610–621, Nov. 1973.
[43] M. Galloway, "Texture analysis using grey level run lengths," *Comput. Graph. Image Process.*, vol. 4, pp. 172–179, 1975.
[44] A. Chu, C. Sehgal, and J. Greenleaf, "Use of gray value distribution of run lengths for texture analysis," *Pattern Recognit. Lett.*, vol. 11, no. 6, pp. 415–419, 1990.
[45] B. Dasarathy and E. Holder, "Image characterizations based on joint gray level-run length distributions," *Pattern Recognit. Lett.*, vol. 12, no. 8, pp. 497–502, 1991.
[46] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed. New York: Wiley-Interscience, 2001.
[47] A. Elisseeff and M. Pontil, "Leave-one-out error and stability of learning algorithms with applications," *NATO Sci. Ser., III: Comput. Syst. Sci.*, vol. 190, pp. 111–130, 2003.
[48] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York: Springer, 2009.
[49] K. Jackowski, B. Krawczyk, and M. Woźniak, "Cost-sensitive splitting and selection method for medical decision support system," in *Intelligent Data Engineering and Automated Learning—IDEAL 2012*, ser. Lecture Notes in Computer Science, H. Yin, J. A. Costa, and G. Barreto, Eds. Berlin: Springer, 2012, vol. 7435, pp. 850–857.