sciendo

Int. J. Appl. Math. Comput. Sci., 2018, Vol. 28, No. 4, 759–770
DOI: 10.2478/amcs-2018-0058

amcs

THE FEATURE SELECTION PROBLEM IN COMPUTER–ASSISTED CYTOLOGY

MAREK KOWAL [a,*], MARCIN SKOBEL [a], NORBERT NOWICKI [a,b]

[a]Institute of Control and Computation Engineering
University of Zielona Góra, ul. Szafrana 2, 65-516 Zielona Góra, Poland
e-mail: M.Kowal@issi.uz.zgora.pl

[b]Department of Medical Physics
University Hospital in Zielona Góra, ul. Zyty 26, 65-046 Zielona Góra, Poland

Modern cancer diagnostics is based heavily on cytological examinations. Unfortunately, visual inspection of cytological preparations under the microscope is a tedious and time-consuming process. Moreover, intra- and inter-observer variations in cytological diagnosis are substantial. Cytological diagnostics can be facilitated and objectified by using automatic image analysis and machine learning methods. Computerized systems usually preprocess cytological images, segment and detect nuclei, extract and select features, and finally classify the sample. In spite of the fact that a lot of different computerized methods and systems have already been proposed for cytology, they are still not routinely used because there is a need for improvement in their accuracy. This contribution focuses on computerized breast cancer classification. The task at hand is to classify cellular samples coming from fine-needle biopsy as either benign or malignant. For this purpose, we compare 5 methods of nuclei segmentation and detection, 4 methods of feature selection and 4 methods of classification. Nuclei detection and segmentation methods are compared with respect to recall and the F1 score based on the Jaccard index. Feature selection and classification methods are compared with respect to classification accuracy. Nevertheless, the main contribution of our study is to determine which features of nuclei indicate reliably the type of cancer. We also check whether the quality of nuclei segmentation/detection significantly affects the accuracy of cancer classification. It is verified using the test set that the average accuracy of cancer classification is around 76%. Spearman's correlation and chi-square test allow us to determine significantly better features than the feature forward selection method.

Keywords: nuclei segmentation, feature selection, classification, breast cancer, convolutional neural network.

1. Introduction

Breast cancer diagnosis is a complex and time-consuming process. It is based on the so-called triple-test, which includes three medical examinations: palpation, mammography or ultrasonography imaging, and fine needle biopsy (FNB). In this work we focus on the processing of results of FNB examination. FNB is performed by an experienced pathologist under the control of an ultrasonograph. As a result of the biopsy we get cellular material which is then stained using basophilic hematoxylin (blue dye) and eosinophilic eosin (red dye). The cytological preparation can be scanned using a microscopic scanner and automatically analyzed with computerized systems. The latter can assist the pathologist by cell counting, determining their morphometric features, classifying tumors, or discovering new diagnostics rules invisible to the naked eye.

It is reported in medical studies that morphometric, textural and topological features of cell nuclei are important indicators of the cancer type. Automatic extraction of these features is possible if cell nuclei are precisely segmented and then detected. Thus, the nuclei segmentation/detection process is crucial for successful computer assisted cytology (CAC). Unfortunately, nuclei segmentation/detection is difficult because tissue samples are composed of complex cellular structures with pervasive occlusions. Sometimes, even for a human, it can be hard to extract nuclei from clumps. Nuclei had until recently been segmented using classical segmentation methods such as intensity thresholding, the watershed

*Corresponding author

method or active contours (Irshad *et al*., 2014; Yang *et al*., 2006; Więcławek and Piętka, 2015; Koyuncu *et al*., 2016; Piórkowski, 2016; Paramanandam *et al*., 2016; Kłeczek *et al*., 2017).

However, we can observe that convolutional neural networks (CNNs) are becoming the state-of-the-art method of nuclei segmentation (Sadanandan *et al*., 2017; Khoshdeli *et al*., 2017). The main advantage of this approach is that the CNN learns from training data a hierarchy of filters to extract invariant features to represent an image. This approach has proven to be more accurate for semantic segmentation than methods based on features engineered by hand. However, we must be aware of the fact that manual labeling images to generate training data is very labor-expensive and tedious. Moreover, the classical CNN segments a single image by separately processing every local patch generated for every pixel. Thus, such an approach is very slow and strongly limits the use of sliding-window-based architectures. Fortunately, fully convolutional neural networks (FCNNs) are free from most of these disadvantages. In particular, FCNN works even if very few training images are available (typical scenario in medical applications).

After successful nuclei segmentation/detection we are able to extract their morphometric, textural and topological features. We can generate an enormous number of different features to describe nuclei as well as their statistics to describe the cytological sample. Nonetheless, most of them will not contain useful information to classify the cancer type. Therefore, we need to select the ones that are most informative for our task. Unnecessary features have to be rejected because they can interfere with the classification process.

The topic of feature selection is a popular and well-recognized problem in scientific literature (Kowal and Filipczuk, 2014; Roffo, 2016; Szemenyei and Vajda, 2017). However, the problem of feature selection for cytological images is rather rarely dealt with (Jeleń *et al*., 2008; Filipczuk *et al*., 2013; Araújo *et al*., 2017). Jeleń *et al*. (2008) applied a manual feature selection procedure and then they were discriminating among breast cancer cases as medium-malignant and very malicious. They achieved classification errors from 5.76% to 24.71%. The experiment was carried on using only 110 images. Filipczuk *et al*. (2013) present an automatic feature selection procedure based on a feature forward selection scheme. The authors were able to reach the classification accuracy equal to 98.51% (determined for patients, not for single images) using an image database consisting of 737 cytological images. Unfortunately, it was not reported if the experiment was verified using a test set. Araújo *et al*. (2017) used a CNN to classify the malignancy of breast cancer. In this case features were learned from data. The training data set consisted of 249 images and the test data set included 20 images. For 4 levels of breast cancer malignancy, the authors report that the classification accuracy was 77.8% and for 2 levels of malignancy the accuracy was 83.3%.

In this work we are addressing the problem of breast cancer classification based on cytological images. However, we are not trying to outperform the existing methods, but we would like to establish a reliable reference baseline of accuracy for the breast cancer classification problem (a two-class problem: benign or malignant). We tested 5 segmentation methods, 4 feature selection methods and 4 classification techniques in all possible configurations. To make the results as reliable as possible, every experiment was repeated 4 times. For this purpose, we generated randomly 4 folds with training images and test images. Test images were never used to select features or train classifiers. Experiments carried out showed that using state-of-the art segmentation, feature selection and classification methods, we can expect approximately 76% classification accuracy. We also checked how the number of features used in the classification affects its accuracy. It was observed that for more than 3 features there was no significant improvement in the accuracy of classification.

The remainder of this paper is organized as follows. The material used in the experiments are described in Section 2. The details of the method are presented in Section 3. Section 4 describes the experiment and results. Discussion and concluding remarks are given in Section 5.

2. Medical image database

Breast cancer fine needle biopsy samples were obtained from 50 patients of the University Hospital in Zielona Góra, Poland. The set contains 25 benign and 25 malignant cases. All cancers were histologically confirmed, and all patients with a benign disease were either biopsied or followed for a year. Smears from the cellular material were fixed in a spray fixative and dyed with hematoxylin and eosin. Cytological preparations were then digitized into virtual slides using a virtual microscopy system.

The system consists of a 2/3 in CCD camera and a $40\times$ objective. The average size of the slides is approximately $200\,000 \times 100\,000$ pixels. The scans were prepared using the extended focal imaging technique. Next, on each slide a pathologist manually selected 11 distinct regions of interests, which were converted to 8 bit/channel RGB TIFF files of the size of 512×512 pixels compressed with the lossless LZW algorithm. All images were labeled as benign or malignant cases. Moreover, 100 images (50 malignant and 50 benign) were extracted from this collection and manually segmented. These images were used to train and validate the convolutional neural network (50 images) and to compare the accuracy of methods used for semantic segmentation (50 images).

Feature selection procedures and the classification process were conducted using 500 images.

3. Methods

3.1. Preprocessing. Cell nuclei are crucial diagnostic objects in cytology. Therefore, it seems desirable to preprocess the image to filter out cytoplasm and red blood cells and leave nuclei. The cellular material is dyed with hematoxylin and eosin. The former is mainly absorbed by nuclei and the latter by cytoplasm. As a result, nuclei are blue color and cytoplasm is red. However, nuclei structures also deposit eosin to some extent. Absorption spectra of hematoxylin and eosin overlap in RGB space, but color deconvolution allows us to evaluate to some extent the contribution of hematoxylin and eosin at each pixel (Ruifrok and Johnston, 2001; Nurzynska, 2018). Three separate intensity images are created as a result of deconvolution; the first represents the hematoxylin density, the second the eosin density, and the third the residuals. For further processing, we are using images of the hematoxylin density. They emphasize nuclei and suppress cytoplasm as well as red blood cells which absorbs mainly eosin.

3.2. Nuclei segmentation. We have chosen for our study 5 well-known methods of nuclei segmentation: Image's nuclei segmentation algorithm (IJ), CellProfiler pipeline for nuclei segmentation based on an RGB input image (CPRGB), CellProfiler pipeline for nuclei segmentation based on an input image preprocessed by deconvolution (CPD), marker controlled watershed (MCW) with conditional erosion, and 2 U-Net convolutional networks with a marker controlled watershed (2UNETW). Usually, the accuracy of nuclei segmentation is measured using the distances between detected nuclei and reference nuclei (manually marked). We are able to evaluate segmentation in term of distances between nuclei, but also in term of classification accuracy.

3.2.1. Marker-controlled watershed. The watershed is the state-of-the-art method of image segmentation. It is widely used for cell nuclei segmentation (Yang *et al.*, 2006; Cheng and Rajapakse, 2009; Jung and Kim, 2010; Irshad *et al.*, 2014; Koyuncu *et al.*, 2016). The classical watershed algorithm treats the image as a topographic surface I_{TM}. It segments the image by flooding basins from seeds until basins attributed to different seeds meet on watershed lines. The input of the algorithm is usually a distance transform of a binary mask of the image. The binary mask is obtained by intensity thresholding, and local maxima from the distance transform are used as seeds for flooding topographic surface I_{TS}. Unfortunately, the algorithm in classical form tends to create an excessive number of micro-segments (Yang *et al.*, 2006). To deal with this problem, we used the MCW version of the watershed, which uses nuclei seeds generated by conditional erosion to refine topographic map I_{TS}.

In our approach, we are transforming the input image by color deconvolution. Next, the preprocessed image is binarized using the Otsu thresholding because the nuclei are dark and the other objects are quite bright. It can be observed that at this stage of segmentation some nuclei are properly segmented, but there is also a lot of clustered nuclei which create large clumps in the binarized image. They are stuck together and must be further processed to be separated. To tackle this problem, we process the binary map of nuclei using conditional erosion. The method is used to separate clustered nuclei and to find their centers. It has two steps. Both perform repetitive operations of morphological erosion. During the first step, repetitive erosion is conducted using a coarse structuring element. The coarse erosion tends to keep the actual shape of nuclei but reduces their size quickly. To prevent objects from disappearing, all those whose area was reduced below a predefined threshold T_c are further eroded using fine structuring element. Fine erosion is less likely to make objects disappear, but it leads to a loss of their original shape. Fine erosion is conducted for an object as long as its area is below the predefined minimum threshold T_f. The whole process is ended when all objects have areas below this threshold.

As a result we get a binary image of nuclei seeds I_S. It is combined with the original topographic map I_{TM} to improve the segmentation accuracy of the watershed method. Topographic surface I_{TS} is modified according to found seeds I_S using morphological reconstruction $\rho_{I_{TS}}(I_S)$ (Vincent, 1993). The algorithm is based on repeated dilations of a seed mask I_S until the contour of this mask fits under topographic map I_{TS}

$$I'_{TM} = \rho_{I_{TS}}(I_S) = \bigcup_{n \geq 1} \delta^{(n)}_{I_{TS}}(I_S). \tag{1}$$

The geodesic dilation of the n-th level is given by

$$\delta^{(n)}_{I_{TS}}(I_S) = \underbrace{\delta_{I_{TS}}(\ldots \delta_{I_{TS}}(\delta_{I_{TS}}(I_S)))}_{n \text{ times}}, \tag{2}$$

and the elementary geodesic dilation is described by the following relationship:

$$\delta_{I_{TS}}(I_S) = (I_S \oplus B) \cap I_{TS}, \tag{3}$$

where $(I_S \oplus B)$ is a one-step standard dilation followed by an intersection (pointwise minimum $\cap$) and B is the 4-connected neighborhood structural element with pairs of horizontal and vertical connected pixels. The modified topographic surface I'_{TS} preserves regional minima at the locations specified by the seeds I_S but suppresses others. Such a version of MCW allows splitting the clustered nuclei, avoiding over-segmentation.

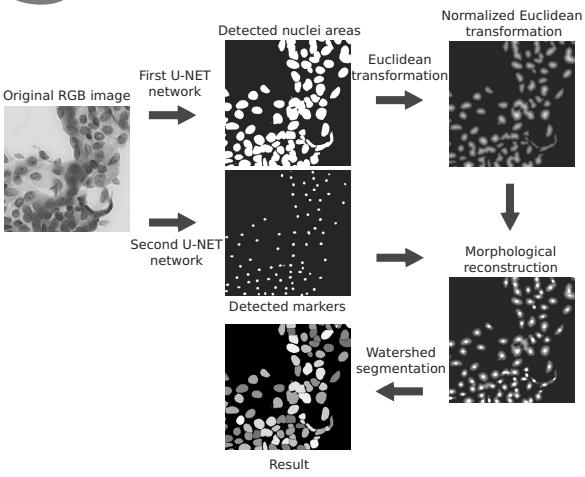

Fig. 1. Watershed segmentation using the U-NET network.

3.2.2. Hybrid approach based on U-Net neural networks and a marker-controlled watershed. FCNNs are currently by far the most popular approach for semantic segmentation. For this reason, we decided to verify their segmentation performance using our cytological images. For this purpose, we developed a hybrid system for nuclei detection based on two U-Net networks and MCW. We assumed that both types of networks will have the same layers and topology as presented in the original paper by Ronneberger *et al*. (2015). The training images were manually annotated using four labels: nuclei interiors, nuclei edges, nuclei centers, and background. Both networks were trained using 50 input images and their corresponding segmentation maps of the size of 512×512 [px]. To process arbitrarily sized images, the overlap-tile strategy for seamless segmentation was used as described by Ronneberger *et al*. (2015). The setup of the training process was also taken from that.

The first network was trained to predict if pixels belong to nuclei interiors (class 1) or nuclei edges and background (class 2). Based on the semantic segmentation provided by this network, we extracted nuclei interiors in the form of binary images. These were used to determine topographic surface I_{TS} for MCW using a distance transform. The second U-Net network was trained to detect the nuclei centers. The network was predicting if pixels belong to the nuclei center (class 1) or to nuclei interior, edge and background (class 2). The centers (nuclei seeds) were extracted from semantic segmentation results in the form of binary images I_S.

In the final step, topographic surface I_{TS} was combined with centers I_S (nuclei seeds) using morphological reconstruction. Modified topographic surface I'_{TS}, was processed by the classical watershed transform in order to detect nuclei.

3.2.3. ImageJ. ImageJ is already classical software for processing medical images. It includes a package for segmenting nuclei based on the watershed method (ImageJ, 2015). This algorithm has a few steps. At the beginning, the input image is processed using the color deconvolution procedure to extract nuclei (nuclei mostly absorb blue dye, hematoxylin). Next, the image is slightly blurred using the Gaussian filter and then thresholded using the Otsu method (Otsu, 1979). Finally, the watershed method is applied to detect instances of nuclei.

3.2.4. CellProfiler. CellProfiler is a well-known software for the processing and analysis of medical images. It offers a lot of predefined pipelines which implement well-known image processing procedures. The nuclei segmentation procedure implemented in CellProfiler is based on the watershed method. We used CellProfiler to process color RGB images (CPRGB) and images after color deconvolution (CPD), hence they are treated as two separate segmentation methods.

3.3. Feature extraction. Based on the results of nuclei detection, we can determine various features describing a particular cell nucleus. In this work, we are proposing the set of 42 features. Unfortunately, we are not able to classify separate nuclei based on these features because we do not have training data labeled on this level of detail. Pathologists would have to put too much effort into labeling every nucleus in 500 training images. Thus, at our disposal are only labels given for the whole images. For this reason, we need to aggregate nuclei features for every image. These aggregates can be used to classify images. Aggregations are defined as the mean, median, variance, standard deviation, kurtosis, interquartile range, and skewness. This gives a total number of 294 features for each image. They are computed for every image based on the features of nuclei coming from this image.

The set on nuclei features can be divided into three groups. The first one is related to the size and shape of the nuclei. This is represented by area (A), perimeter (P), shape factor (SF), convex deficiency (CD), eccentricity (E), major axis length (MjAL), minor axis length (MnAL), bending energy (BE).

The second group of features is related to the distribution of nuclei in the image. Benign cells usually form single-layered structures, while malignant cells tend to break up which increases the probability of encountering separated nuclei. To express this relation, we use features representing the spatial distribution of nuclei: the distance to the centroid of all nuclei (D2C), and the distance to k-nearest nuclei (D2KNN).

The third group of features is related to the distribution of chromatin in the nuclei. This is

represented with textural features based on the gray-level co-occurrence matrix (GLCM) (Haralick *et al*., 1973) and the gray-level run-length matrix (GLRLM) (Tang, 1998).

The first 13 textural features are based on the GLCM. The $N \times N$ matrix P, where N is the number of gray levels, is defined over an image to be the distribution of co-occurring values of pixels at a given offset. In other words, each element of P specifies the number of times a pixel with gray-level value i occurs shifted by a given distance to a pixel with the value j. We calculate the mean of GLCM features determined for offsets corresponding to 0°, 45°, 90° and 135° using 8 gray-levels: GLCM energy (GLCM_01), GLCM contrast (GLCM_02), GLCM correlation (GLCM_03), GLCM sum of squares (GLCM_04), GLCM homogeneity (GLCM_05), GLCM sum average (GLCM_06), GLCM sum variance (GLCM_07), GLCM sum entropy (GLCM_08), GLCM entropy (GLCM_09), GLCM difference variance (GLCM_10), GLCM difference entropy (GLCM_11), GLCM information measures of correlation type 1 (GLCM_12), GLCM information measures of correlation type 2 (GLCM_13).

The next 11 textural features are based on the GLRLM. The $N \times M$ matrix p, where an N is the number of gray levels and M is the maximum run length, is defined for a given image as the number of runs with pixels of gray level i and run length j. As in GLCM, we compute run length matrices for 0°, 45°, 90° and 135° using 8 gray-levels: GLRLM short run emphasis (GLRLM_01), GLRLM long run emphasis (GLRLM_02), GLRLM gray-level nonuniformity (GLRLM_03), GLRLM run-length nonuniformity (GLRLM_04), GLRLM run percentage (GLRLM_05), GLRLM low gray-level run emphasis (GLRLM_06), GLRLM high gray-level run emphasis (GLRLM_07), GLRLM short run low gray-level emphasis (GLRLM_08), GLRLM short run high gray-level emphasis (GLRLM_09), GLRLM long run low gray-level emphasis (GLRLM_10), GLRLM long run high gray-level emphasis (GLRLM_11).

The last group of features is related to the colorimetric features of nuclei: mean red channel value (MR), mean green channel value (MG), mean blue channel value (MB), mean lightness value (ML), variance of the red channel value (VR), variance of the green channel value (VG), variance of the blue channel value (VB), variance of the lightness value (VL).

3.4. Feature selection. The number of features defined for the image is very large. Some of them are either redundant or irrelevant, and thus can be removed without much loss of information. We are using 4 methods of feature subset selection: forward selection, 2 variants of correlation feature selection (CFS) and feature selection based on the test of variable independence.

3.4.1. Forward selection method. Forward selection (FS) is a wrapper-based algorithm, which searches through the space of possible features and evaluates each subset by running a classification model on the subset. The procedure starts from the classification model, which does not have any input variables. Then, the set of input variables is recursively expanded. The variable that increases the accuracy of the model most is added to the resulting subset. Sometimes several variables increase the accuracy of the model by the same amount. They are stored in the queue. The first variable from the queue is recursively expanded until no improvement can be detected. Then, the algorithm goes backward to the last stored queue and expands the subsequent variable from the queue. The search procedure is repeated until no improvement can be detected and there are no queues storing variables for expansion.

The image set used to select features contains 500 images coming from 50 patients. Each case is described by 10 images. The images were divided into two subsets: the training set (300 images, 30 cases) and the test set (200 images, 20 cases). Cases were distributed randomly among subsets. Every feature selection experiment was repeated four times to increase the statistical significance of the results. For every experiment, new training and test sets were generated randomly.

In every step of the searching algorithm, a given classification model must be trained and validated using 300 images coming from the training set. The performance of the model was validated with the n-fold cross-validation procedure. There were 30 folds (the number of patients), and each consisted of 10 images belonging to a single patient. The classification model was trained and validated 30 times for every variable subset. The final evaluation of the variable subset was computed using a mean for 30 validation results.

3.4.2. Correlation feature selection. CFS evaluates features on the basis of their correlation with the target variable. We used Spearman's rank correlation to evaluate the correlation between every input and output variable (Spearman, 1904). Spearman's coefficients were computed using 300 images coming from the training set, the remaining 200 images were used for testing. Every experiment was repeated 4 times using different training and test sets (generated randomly). Variables were sorted with respect to the absolute value of Spearman's coefficients. Based on this ranking, we can select a subset of variables which are strongly correlated with the target variable. The finally chosen subset is tested using a given classification model. The model is trained using 300 images from training set and then tested using 200 images from the test set. The final evaluation of the variable subset is described by the classification accuracy of the model.

3.5. Correlation feature selection using non-redundant features. The method is very similar to the CFS strategy. The only novelty is the pre-processing step used to find and remove redundant input variables. The redundancy of variables is evaluated using Spearman's rank correlation matrix. Every variable pair with the absolute value of Spearman's correlation higher than 0.9 is considered to be strongly correlated. Having two input variables that are strongly correlated, we delete the one that is less correlated with the target variable. Such a pre-processing procedure removed 149 redundant variables and the CFS procedure was applied for other 145 variables.

3.6. Chi-square test of independence. The chi-square test of independence can be used to check whether there is a significant relationship between two variables. We have to formulate a null hypothesis that two variables are independent and an alternative hypothesis that variables are not independent. The test statistic, is a chi-square random variable χ^2. Based on the value of the χ^2 statistic it is possible to compute the V-Cramer coefficient, which measures the association between two categorical variables:

$$VC = \sqrt{\frac{\chi^2}{n \min(\sqrt{(r-1)(f-1)})}}, \qquad (4)$$

where r is the number of levels of the first feature, f is the number of levels of the second feature, n is the number of samples, χ^2 is the value of the chi-square statistic. The VC coefficient takes values in the range from 0 to 1. Higher values of the VC mean stronger dependence between variables.

Unfortunately, the VC coefficient can be used only for categorical variables and in our case we deal with continuous variables. We can convert our variables using a binning algorithm. However, classical binning methods did not work well in our case. That is why we proposed a heuristic method that divides the range of the variable into seven bins. Examination of histograms showed that we should bin the variables more densely in the vicinity of their expected values. To do this, the experimental distribution is approximated by the normal distribution (Fig. 2). Next, the procedure is dividing the sigma area using 5 equally long intervals and form 2 single intervals for values below and above the sigma area, respectively (cf. Fig. 3). Finally, the method generates 7 intervals, including 5 uniform intervals around the expected value and 2 broader intervals for extreme values.

3.7. Classification models. All presented feature selection methods are evaluated and compared with respect to image classification accuracy. It is given as the ratio of the number of successfully classified images to the total number of images. Classification models were trained using 300 images and then tested using 200 images. Every such experiment was repeated 4 times. Experiments were conducted using different training images, and test images because both image sets were generated randomly for each experiment. For classification, we used four classification algorithms: naive Bayes classifier (NB), k-nearest neighbor (kNN) (Cover and Hart, 1967) using $k = 9$, decision tree (DT) (Breiman *et al*., 1984), and support vector machine (SVM) (Cortes and Vapnik, 1995) using a third-order polynomial kernel and scale factor $\sigma = 0.9$.

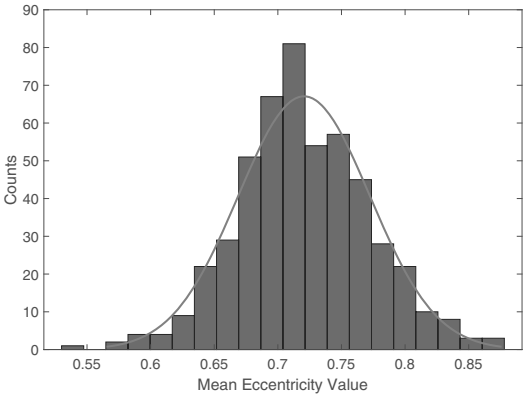

Fig. 2. Example of the binning problem.

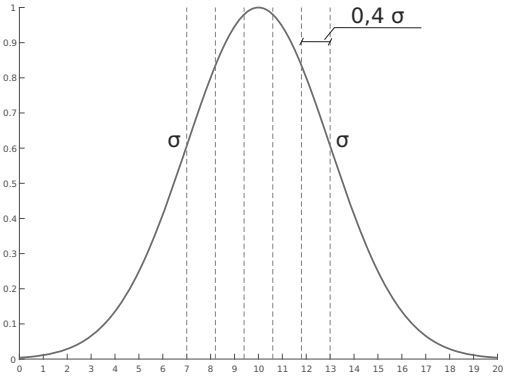

Fig. 3. Heuristic approach to the binning problem.

4. Experimental results

4.1. Segmentation results. In this study, we used 4 methods to segment cell nuclei. To compare their accuracy, they were employed to detect nuclei in 50 test images. The accuracy of automatic detection is measured with the help of reference manual segmentation. For each test image, we are given a list of manually labeled nuclei in the form of binary masks. At our disposal are also

binary masks of objects generated by MCW, 2UNETW, CPD, CPRGB and IJ. Thus, it is possible to measure distances between ground-truth nuclei and automatically generated objects. For this purpose the Jaccard index is very often employed,

$$JI = \frac{A \cap B}{A \cup B}. \tag{5}$$

For each reference nucleus, the method tries to find the closest object detected automatically. However, the reference nucleus and the closest object can be matched if their Jaccard index is above the predefined threshold $T = 0.5$. Otherwise, they cannot be paired and the reference nucleus stays without the accompanying object. To conclude, 3 scenarios are possible for each reference nucleus: it can be matched with the nearest object detected, and such a case is classified as true positive (TP); no object can be found to match the reference nuclei, and such a case is classified as false negative (FN), detected object can stay without the corresponding reference an nucleus and this is classified as false positive (FP).

The accuracy of segmentation is measured using the true positive rate (TPR) and F1-score. The former is defined as the ratio of the number of correctly detected nuclei (TP) and the number of all reference nuclei:

$$TPR = recall = \frac{TP}{TP + FN}. \tag{6}$$

The latter is given by the following formulas:

$$F1\ score = \frac{2 \times precision \times recall}{precision + recall}, \tag{7}$$

where

$$precision = \frac{TP}{TP + FP}.$$

Sample results of nuclei segmentation for all tested methods are presented in Fig. 4

TPR values and F1-score values aggregated for 50 images are given in Table 1. These accuracy measures are presented for all methods with respect to the cancer type. The results indicate that 2UNETW outperforms all other methods for both benign and malignant cases.

Table 1. Segmentation accuracy with regard to the cancer type.

	2UNETW	CPD	CPRGB	IJ	MCW
TPR:					
Benign	**0.82**	0.69	0.61	0.70	0.74
Malignant	**0.93**	0.66	0.63	0.65	0.75
F1 score:					
Benign	**0.88**	0.70	0.58	0.66	0.71
Malignant	**0.86**	0.75	0.50	0.73	0.70

4.2. Feature selection and classification results. Four methods of feature selection were tested. The accuracy of feature selection is measured in terms of classification accuracy. The latter was measured using four classifiers. Every classification experiment was repeated 4 times with the different test set. Moreover, we used 5 segmentation algorithms to extract nuclei features. In total, each feature selection procedure was run 80 times for all combinations of segmentation methods, classification methods, and test sets. The number of features was rigidly set to 5. In Table 2, we can see the summarized results of these runs. The measure named 'Best result' indicates how many times a given method of feature selection gave the best score. Sometimes more than one method was reaching the same best score.

We can observe that all FCS based methods have similar accuracy and it is clear that the FS method is significantly worse than FCS based methods.

Table 3 presents the results for the same experiments as the previous one, but this time the results are also broken down by the method of classification. The best mean result was obtained by the SVM, but we can observe that the type of classification method has a small impact on classification accuracy. The results presented in Table 4 are broken down by feature selection, the type of classifier and the type of segmentation. Every cell in this table represents mean accuracy of classification calculated on the basis of results from 4 experiments conducted for

Table 2. Aggregated results of feature selection for different combinations of segmentation, classification and the test set.

	FS	CFS	CFSNR	CHI2
Mean	72.32	75.75	75.66	**75.92**
Maximum	88.50	91.50	91.50	**93.00**
Minimum	51.50	56.50	57.00	**58.50**
Best result	14	26	23	**33**

Table 3. Feature selection results with regards to the feature selection method and the classification method.

		FS	CFS	CFSNR	CHI2
NB	Mean	72.95	**77.03**	76.20	76.55
	Maximum	88.00	91.00	90.50	**93.00**
	Minimum	60.50	62.00	57.00	**59.00**
DT	Mean	68.25	71.30	72.08	**72.33**
	Maximum	83.50	86.00	**91.50**	84.50
	Minimum	51.50	57.50	**62.50**	58.50
KNN	Mean	74.33	76.43	75.75	**76.48**
	Maximum	88.50	**89.50**	89.00	89.00
	Minimum	62.50	62.00	62.50	**64.00**
SVM	Mean	73.75	78.25	**78.60**	78.33
	Maximum	88.00	**91.50**	90.50	90.50
	Minimum	60.50	56.50	60.00	**64.50**

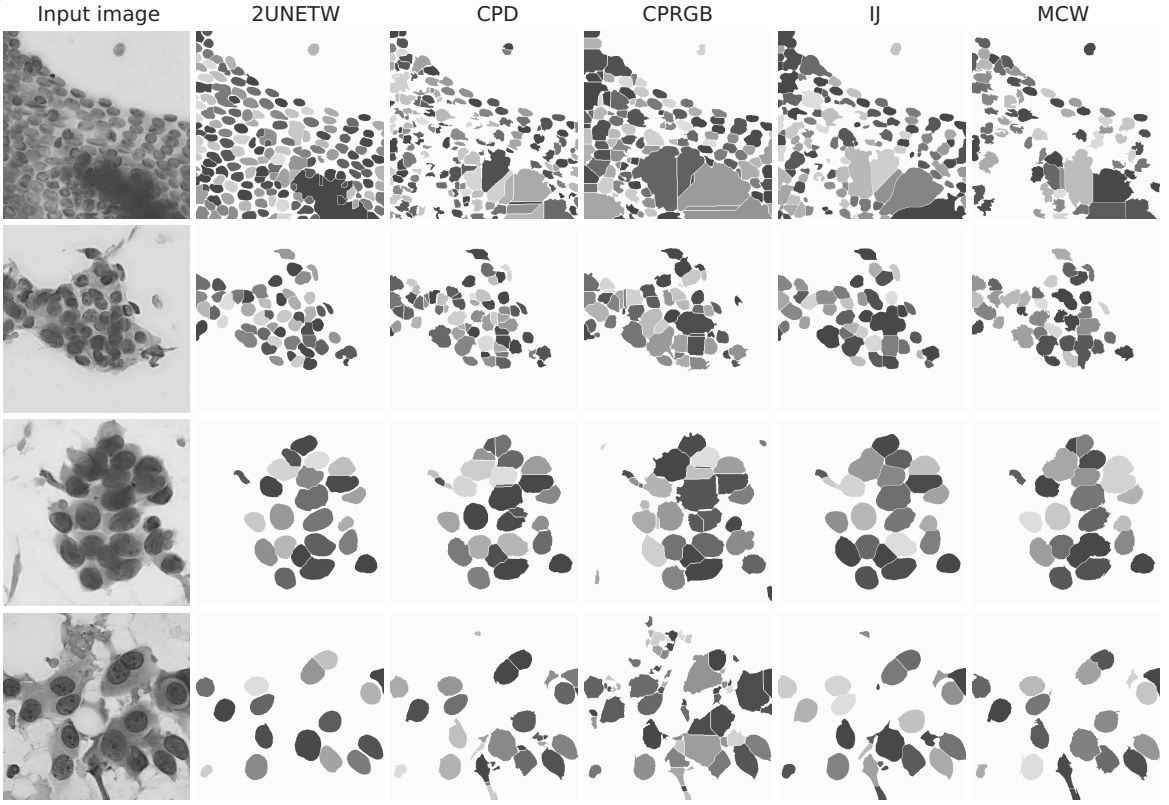

Fig. 4. Examples of segmentation results.

Table 4. Detailed results for feature selection experiments.

		FS	SC	SCIC	CHI2
NB	2UNETW	79.13	81.38	81.75	**82.75**
	CPD	75.38	**80.38**	79.75	78.88
	CPRGB	64.50	**69.50**	67.25	68.25
	IJ	70.50	**74.75**	73.75	73.50
	MCW	75.25	79.13	78.50	**79.38**
DT	2UNETW	76.00	74.88	75.88	**76.75**
	CPD	69.88	73.50	**74.63**	73.50
	CPRGB	66.25	65.25	67.50	**69.38**
	IJ	61.38	**71.25**	69.38	70.38
	MCW	67.75	71.63	**73.00**	71.63
KNN	2UNETW	**80.75**	79.88	78.63	79.38
	CPD	77.00	79.00	78.25	**79.88**
	CPRGB	**71.75**	70.63	71.00	71.13
	IJ	66.88	74.63	**74.75**	73.13
	MCW	75.25	78.00	76.13	**78.88**
SVM	2UNETW	77.38	82.38	**83.00**	82.38
	CPD	78.50	81.00	**83.13**	80.63
	CPRGB	66.75	69.25	69.00	**72.13**
	IJ	72.00	**78.38**	77.63	77.88
	MCW	74.13	**80.25**	**80.25**	78.63

different test sets.

We can observe that the accuracy of nuclei segmentation has some influence on classification accuracy because 2UNETW has generally better accuracy than other segmentation results. However, we expected that the segmentation accuracy would have a greater impact on the accuracy of the classification.

In the next experiment, we checked how the number of features selected influences classification accuracy. For this purpose, we were sequentially increasing the number of selected features starting from a single one. The procedure was repeated 16 times. In each iteration, it is limited to find only a predefined number of features. The feature selection algorithm was choosing features using training images and next test their goodness using test images. At the end we get 80 results because the procedure was repeated for all combinations of segmentation methods, classification methods and test sets. These results are aggregated and presented in the form of box-plots. The central sign in the box indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points (not outliers), and the outliers are plotted using the plus symbol.

We repeated such an experiment for every feature selection method. The results obtained for training data and test data using the FS method are presented in Figs. 5 and 6 respectively. We can observe that, by increasing the number of features up to 16, we can fit models closely to

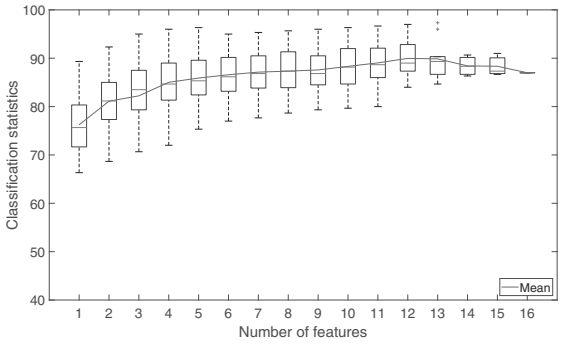

Fig. 5. Forward selection (FS): training.

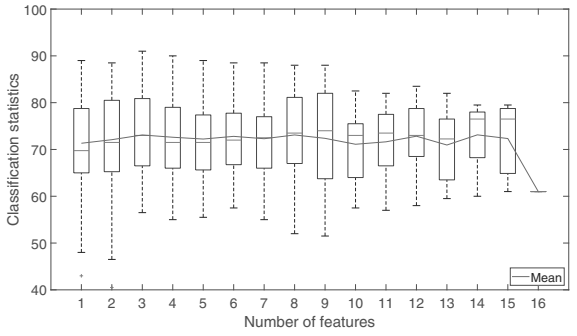

Fig. 6. Forward selection (FS): testing.

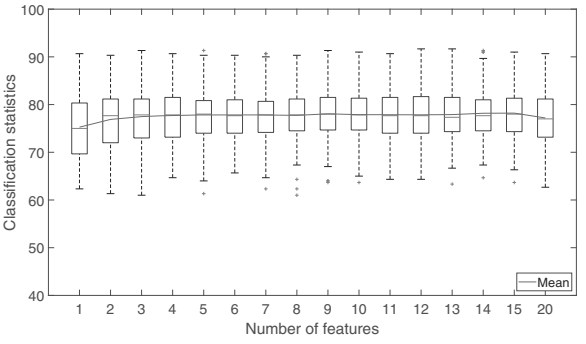

Fig. 7. Correlation feature selection (CFS): training.

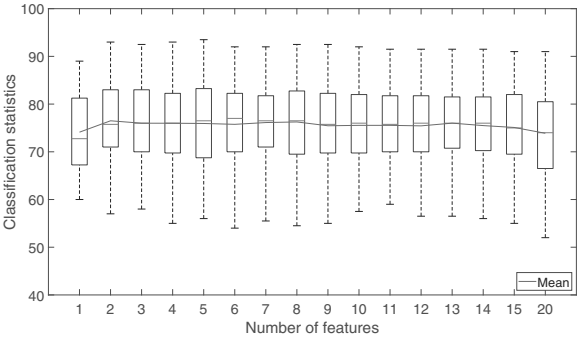

Fig. 8. Correlation feature selection (CFS): testing.

training data (average accuracy above 80.00%). However, the accuracy for test data at the beginning increases but then stabilizes and at the end it decreases. This is a symptom of overfitting. The average accuracy for test data ranges between 70–73% and is achievable even by 2 or 3 features. The best accuracy for the test images is equal to 91.00%; unfortunately, the worst result is equal to 40.50%. It can be concluded that the FS method accurately adjusts the set of features to the training set, however, these features do not provide good results for the test set.

The same experiment was conducted for the CFS method. The results for training images are presented in Fig. 7 and for test images in Fig. 8. The method was not able to achieve as high accuracy for training data as the FS approach (the average accuracy does not exceed 80.00%). Nonetheless, for test data it gave better results than FS (the accuracy range from 52.00% to 93.50% and mean accuracy oscillating from 73.88% to 76.26%). Like for the FS method, we can observe symptoms of overfitting. Moreover, we can see that 2 features ensure a similar accuracy as models with more features.

The results obtained for the CFS method using non-redundant features are presented in Fig. 9 for training data and in Fig. 10 for test data. Despite the large reduction in the number of features by eliminating redundant ones, the results for the training set have not been improved. Training accuracy ranges from 60.67% to 93.67%, while the average values from all classifiers do not exceed 80%. The same situation is for test data (accuracy is from 49.00% to 92.50%, whereas average accuracy oscillates between 72.40% to 76.12%)

The results of feature selection based on the chi-square test are presented for the training set in Fig. 11 and for the test set in Fig. 12. The accuracy for training data ranges from 60.33% to 91.67% and, as with the other methods, average values do not exceed 80%. The results obtained for test data are slightly better than for the other methods and range from 55.00% to 93.50%, with average values in the range from 73.43% to 76.58%.

In Fig. 13, we compared the mean values of test accuracy obtained for all feature selection methods. We can clearly observe that the FS method performs much worse than the others. On the other hand, we cannot see a significant difference between CFS methods and that based on the chi-squared test. Moreover, we can conclude that the highest accuracy is obtained for models which use 4–6 features.

In Table 5 we showed 10 top features, which were most frequently chosen by the feature selection

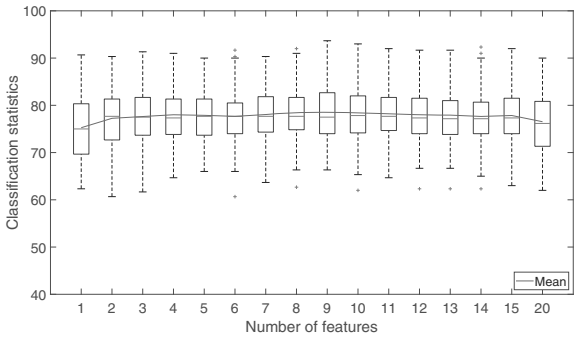

Fig. 9. Correlation feature selection using non-redundant features (CFSNR): training.

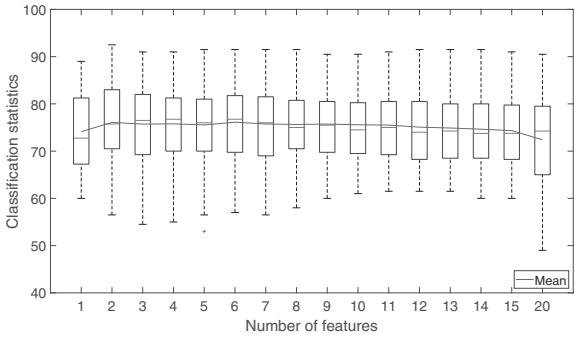

Fig. 10. Correlation feature selection using non-redundant features (CFSNR): testing.

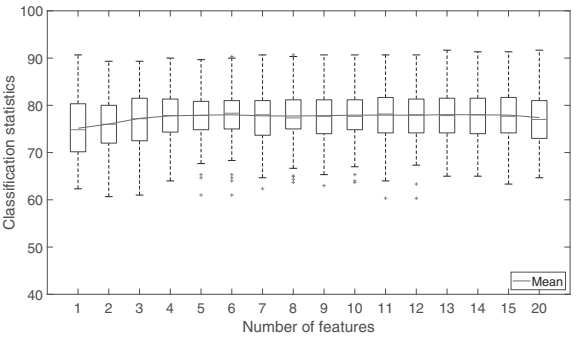

Fig. 11. χ^2 based feature selection: training.

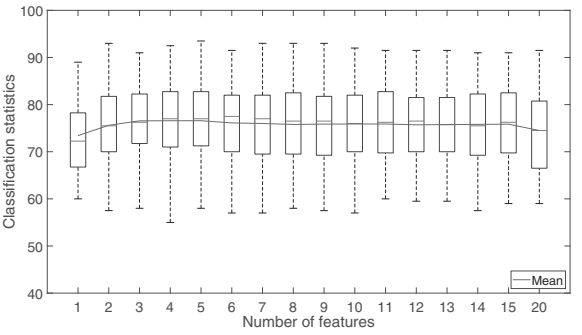

Fig. 12. χ^2 based feature selection: testing.

algorithms. We can observe that all methods chose similar features.

5. Conclusions

Breast cancer diagnosis using cytological images is a very difficult challenge. The content of such images is highly complex and their analysis in an automated way is difficult. This contribution concentrated on the problem of feature selection for automatic breast cancer classification. To test the methods of feature selection we had to build a processing pipeline which contains nuclei segmentation, feature extraction, feature selection and classification. We proposed a hybrid nuclei segmentation method based on two U-Net neural networks and a marker-controlled watershed. We also used well-known segmentation methods based on the watershed transform and implemented in ImageJ and CellProfiler. We tested 4 feature selection methods in terms of classification accuracy. Four classification techniques were used in the final step to measure the goodness of feature sets. The obtained results for test sets range from 40.50% to 93.50% and strongly depend on the selection of the training and test sets.

Based on the results of all conducted experiments we can conclude that the feature selection process is prone to over-fitting. We can observe this especially for wrapper based techniques. We found that changing test data may have huge impact on the accuracy of classification. If we did not repeat the classification experiments many times for different training and test sets, we could report the accuracy either very low or very high.

When it comes to choosing a classifier, the best results are obtained by the SVM; however,

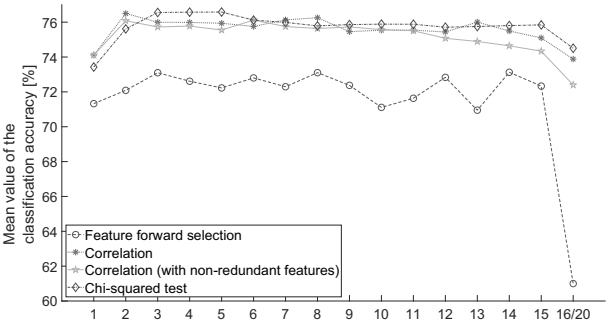

Fig. 13. Mean accuracies for all feature selection methods.

Table 5. Most frequently chosen features (M: mean, V: variance, D: median, T: standard deviation, S: skewness, K: kurtosis, I: interquartile range).

	FS	CFS	CFSNR	CHI2
1	BE_M	D2NNN_I	D2NNN_I	D2NNN_I
2	D2NNN_M	BE_M	D2NNN_V	BE_M
3	D2NNN_I	D2NNN_M	BE_I	D2NNN_T
4	MR_M	D2NNN_T	D2NNN_M	BE_D
5	A_V	D2NNN_V	BE_M	D2NNN_M
6	GLRLM_07_V	BE_D	MnAL_I	P_D
7	BE_D	MnAL_I	BE_V	MjAL_D
8	MjAL_I	A_I	BE_D	MnAL_I
9	D2NNN_V	BE_I	CD_D	A_I
10	MjAL_K	D2NNN_D	GLRLM_06_V	D2NNN_D

the improvement is not significant. Generally one classification techniques and feature selection methods have little influence on the classification accuracy.

The results also indicate that the segmentation accuracy has some impact on the classification accuracy, but not as strong as we thought. Our experiments showed that, using state-of-the-art methods of nuclei segmentation, feature selection and classification, we can expect on the average a 76% accuracy in breast cancer classification.

Acknowledgment

The authors express sincere thanks and appreciation to Dr. Roman Monczak, University Hospital in Zielona Góra, Poland, for preparing test images and Dr. Michał Żejmo, University of Zielona Góra, Poland, for preparing results of semantic segmentation. This work was supported by the National Science Center in Poland (2015/17/B/ST7/03704).

References

Araújo, T., Aresta, G., Castro, E., Rouco, J., Aguiar, P., Eloy, C., Polónia, A. and Campilho, A. (2017). Classification of breast cancer histology images using convolutional neural networks, *PLOS ONE* **12**(6): 1–14.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees*, Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA.

Cheng, J. and Rajapakse, J.C. (2009). Segmentation of clustered nuclei with shape markers and marking function, *IEEE Transactions on Biomedical Engineering* **56**(3): 741–748.

Cortes, C. and Vapnik, V. (1995). Support-vector networks, *Machine Learning* **20**(3): 273–297.

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification, *IEEE Transactions on Information Theory* **13**(1): 21–27.

Filipczuk, P., Fevens, T., Krzyżak, A. and Monczak, R. (2013). Computer-aided breast cancer diagnosis based on the analysis of cytological images of fine needle biopsies, *IEEE Transactions on Medical Imaging* **32**(12): 2169–2178.

Haralick, R., Shanmugam, K. and Dinstein, I. (1973). Textural features for image classification, *IEEE Transactions on Systems, Man, and Cybernetics* **3**(6): 610–621.

ImageJ (2015). Nuclei watershed separation, `https://imagej.net/Nuclei_Watershed_Separation`.

Irshad, H., Veillard, A., Roux, L. and Racoceanu, D. (2014). Methods for nuclei detection, segmentation, and classification in digital histopathology: A review—current status and future potential, *IEEE Reviews in Biomedical Engineering* **7**: 97–114.

Jeleń, L., Fevens, T. and Krzyżak, A. (2008). Classification of breast cancer malignancy using cytological images of fine needle aspiration biopsies, *International Journal of Applied Mathematics and Computer Science* **18**(1): 75–83, DOI: 10.2478/v10006-008-0007-x.

Jung, C. and Kim, C. (2010). Segmenting clustered nuclei using H-minima transform-based marker extraction and contour parameterization, *IEEE Transactions on Biomedical Engineering* **57**(10): 2600–2604.

Khoshdeli, M., Cong, R. and Parvin, B. (2017). Detection of nuclei in H&E stained sections using convolutional neural networks, *2017 IEEE EMBS International Conference on Biomedical Health Informatics, Orlando, FL, USA*, pp. 105–108.

Kłeczek, P., Dyduch, G., Jaworek-Korjakowska, J. and Tadeusiewicz, R. (2017). Automated epidermis segmentation in histopathological images of human skin stained with hematoxylin and eosin, *Proceedings of SPIE: Medical Imaging* **10140**: 10140–10140–19.

Kowal, M. and Filipczuk, P. (2014). Nuclei segmentation for computer-aided diagnosis of breast cancer, *International Journal of Applied Mathematics and Computer Science* **24**(1): 19–31, DOI: 10.2478/amcs-2014-0002.

Koyuncu, C.F., Akhan, E., Ersahin, T., Cetin-Atalay, R. and Gunduz-Demir, C. (2016). Iterative h-minima-based

marker-controlled watershed for cell nucleus segmentation, *Cytometry A* **89**(4): 338–349.

Nurzynska, K. (2018). Optimal parameter search for colour normalization aiding cell nuclei segmentation, *in* S. Kozielski *et al.* (Eds.), *Beyond Databases, Architectures and Structures. Facing the Challenges of Data Proliferation and Growing Variety*, Springer International Publishing, Cham, pp. 349–360.

Otsu, N. (1979). A threshold selection method from gray-level histograms, *IEEE Transactions on Systems, Man, and Cybernetics* **9**(1): 62–66.

Paramanandam, M., O'Byrne, M., Ghosh, B., Mammen, J.J., Manipadam, M.T., Thamburaj, R. and Pakrashi, V. (2016). Automated segmentation of nuclei in breast cancer histopathology images, *PLOS ONE* **11**(9): 1–15.

Piórkowski, A. (2016). A statistical dominance algorithm for edge detection and segmentation of medical images, *in* E. Piętka *et al.* (Eds.), *Information Technologies in Medicine*, Advances in Intelligent Systems and Computing, Vol. 471, Springer, Cham, pp. 3–14.

Roffo, G. (2016). Feature selection library (Matlab toolbox), *arXiv*: 1607.01327.

Ronneberger, O., Fischer, P. and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation, *CoRR*: abs/1505.04597.

Ruifrok, A.C. and Johnston, D.A. (2001). Quantification of histochemical staining by color deconvolution, *Analytical and Quantitative Cytology and Histology* **23**(4): 291–299.

Sadanandan, S.K., Ranefall, P., Guyader, S.L. and Wählby, C. (2017). Automated training of deep convolutional neural networks for cell segmentation, *Scientific Report* **7**: 7860, DOI: 10.1038/s41598-017-07599-6.

Spearman, C. (1904). The proof and measurement of association between two things, *The American Journal of Psychology* **15**(1): 72–101.

Szemenyei, M. and Vajda, F. (2017). Dimension reduction for objects composed of vector sets, *International Journal of Applied Mathematics and Computer Science* **27**(1): 169–180, DOI: 10.1515/amcs-2017-0012.

Tang, X. (1998). Texture information in run-length matrices, *IEEE Transactions on Image Processing* **7**(11): 1602–1609.

Vincent, L. (1993). Morphological grayscale reconstruction in image analysis: Applications and Efficient Algorithms, *IEEE Transactions on Image Processing* **2**(2): 176–201.

Więcławek, W. and Piętka, E. (2015). Watershed based intelligent scissors, *Computerized Medical Imaging and Graphics* **43**: 122 – 129.

Yang, X., Li, H. and Zhou, X. (2006). Nuclei segmentation using marker-controlled watershed, tracking using mean-shift, and Kalman filter in time-lapse microscopy, *IEEE Transactions on Circuits and Systems I: Regular Papers* **53**(11): 2405–2414.

Marek Kowal received his MSc and PhD degrees in electrical engineering from the University of Zielona Góra, Poland, in 2000 and 2004, respectively. Currently, he is an assistant professor in the Institute of Control and Computation Engineering at the University of Zielona Góra. He has published about 60 papers in refereed journal and conference papers. His current interests include stochastic geometry, image analysis, and machine learning.

Marcin Skobel received an MSc degree in geodesy, surveying and cartography (2013) from the AGH University of Science and Technology in Kraków and an MSc degree in computer science (2018) from the University of Zielona Góra. Currently, he is a PhD student at the Faculty of Computer, Electrical and Control Engineering at the University of Zielona Góra. His main research interests are focused on image processing and machine learning in computer vision.

Norbert Nowicki received an MSc degree in computer science from the University of Zielona Góra, Poland, in 2011, and an MSc degree in biomedical engineering in 2014 at the same university. Since 2016 he has been a PhD student at the Institute of Control and Computation Engineering, University of Zielona Góra. Also, he works as a medical physicist at the University Hospital, in the Department of Medical Physics. His current interests include computer-aided diagnosis, medical imaging processing and analysis.

Received: 23 October 2018
Accepted: 10 December 2018