DE GRUYTER OPEN

JAISCR, 2014, Vol. 4, No. 2, pp. 149 – 162
10.1515/jaiscr-2015-0005

WEB–BASED FRAMEWORK FOR BREAST CANCER CLASSIFICATION

Tomasz Bruździński[1], Adam Krzyżak [2], Thomas Fevens [2] and Łukasz Jeleń[3]

[1]*Institute of Computer Engineering, Control and Robotics, Wrocław University of Technology, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland*

[2]*Department of Computer Science and Software Engineering, Concordia University, 1455 De Maisonneuve Blvd. West, Montreal, Quebec, Canada H3G 1M8*

[3]*Institute of Computer Engineering, Control and Robotics, Wrocław University of Technology, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland*

Abstract

The aim of this work is to create a web-based system that will assist its users in the cancer diagnosis process by means of automatic classification of cytological images obtained during fine needle aspiration biopsy. This paper contains a description of the study on the quality of the various algorithms used for the segmentation and classification of breast cancer malignancy. The object of the study is to classify the degree of malignancy of breast cancer cases from fine needle aspiration biopsy images into one of the two classes of malignancy, high or intermediate. For that purpose we have compared 3 segmentation methods: k-means, fuzzy c-means and watershed, and based on these segmentations we have constructed a 25–element feature vector. The feature vector was introduced as an input to 8 classifiers and their accuracy was checked.

The results show that the highest classification accuracy of 89.02 % was recorded for the multilayer perceptron. Fuzzy c–means proved to be the most accurate segmentation algorithm, but at the same time it is the most computationally intensive among the three studied segmentation methods.

1 Introduction

Nowadays the mammary gland cancer is one the most common cancers present in the world [10]. In Poland alone the number of diagnosed cases based on data delivered by the National Cancer Registry for both male and female breast cancer for year 2011 was 16643 [2]. Diagnosing cancer before it starts to produce symptoms is an important matter. Mostly because cancers that are found when they are already causing symptoms tend to be larger and are more likely to have already spread beyond the breast. Therefore the treatment options are limited since such cancers are less responsive to any kind of therapy. In contrast, breast cancers which are diagnosed earlier are more likely to be smaller, still confined to the breast with many efficient treatment options available.

The size and spread range are some of the most important factors in predicting the outlook of a patient's survival. Nowadays there aren't any fully reliable, inexpensive nor non–invasive diagnostic methods for the identification of breast pathology.

The most common diagnostic methods include: self-examination (palpation), mammography or ultrasound imaging and fine needle aspiration biopsy (FNA), each of them differs in a degree of sensitivity and invasiveness. FNA, being the most invasive and the most accurate method, requires collecting a tissue material directly from a tumor for microscopic verification and examination in order to ex-

clude or confirm the presence of cancerous cells [2].

To propose a proper treatment there is a need for an estimation of cancer stage and its malignancy grade. Cancer staging is a process of determining the size and metastasis of the cancer that associates a stage to a case. The most commonly used staging system for breast cancer nowadays is TNM (Tumor, Nodes, Metastasis) [3]. Besides staging, when predicting the progression of the cancer, it is essential to estimate its malignancy grade. In this paper the scale proposed by Bloom and Richardson in 1957 [7] is used to determine the malignancy grade. In this system tumor is assigned a low, intermediate or high malignancy grade. In order to obtain the resulting virulence the cells polymorphy, ability to reform histoformative structures and mitotic index needs to be evaluated. The evaluation process proposed by the Bloom-Richardson scheme utilizes three factors that use a point based scale for assessing previously mentioned features. The malignancy grade is then assigned based on the value calculated by summation of all points awarded for each factor. This is a very difficult procedure that requires extensive knowledge and experience of the cytologist making a diagnosis. It is well known that usually the human is the weakest link of any process as he tends to make mistakes, so the diagnosis is only as good as the pathologist making it. In order to minimize the human factor an automatic computer framework can be introduced which can assist doctors during the diagnostic process. Due to the importance of a proper and accurate determination of the breast cancer diagnosis many approaches can be found in the literature that tackle this problem. One of them includes a firefly method for nuclei detection [22] or even an approach that involves the analysis of thermograms [20]. In this paper we deal with the classification of breast cancer based on the fine needle aspiration biopsy. To the best of our knowledge, the computerized breast cytology classification problem was first investigated by Wolberg *et al.* in 1990 [32]. The authors described an application of a multi-surface pattern separation method to cancer diagnosis. The proposed algorithm was able to distinguish between a 169 malignant and 201 benign cases with 6.5% and 4.1% error rates, respectively depending on the size of the training set. When 50% of samples were used for training, the method returned a larger error. Using 67% of sample images reduced the error to 4.1%. The same authors introduced a widely used database of pre-extracted features of breast cancer nuclei obtained from fine needle aspiration biopsy images [24] (available as the Wisconsin Breast Cancer Database (WBCD) at the UCI Machine Learning Repository [1]). Later, in 1993, Street et al. [31] used an active contour algorithm, called 'snake' for precise nuclei shape representation. The authors also described 10 features of a nucleus used for classification. They achieved a 97.3% classification rate using a multi-surface method for classification.

Xiong *et al.* [33] used partial least squares regression was used to classify that WBCD database with 699 (241 malignant, 458 benign) cases with a 96.57% classification rate. Numerous other researchers have worked with the WBCD database (see [25] and reference therein) with resulting classification rates ranging from 94.74% to 99.54%.

Malek *et al.* [23] used active contours to segment nuclei and classified 200 (80 malignant, 120 benign) cases using a fuzzy c-means classifier achieving a 95% classification rate.

Niwas *et al.* [27] presented a feature extraction method based on the analysis of nuclei Chromatin texture using a complex wavelet transform. These features were used with a k-nearest neighbor classifier where using a data set of 20 malignant and 25 benign cases they achieved a classification rate of 93.33%. Filipczuk *et al.* [11] used a circular Hough transform to detect cell nuclei, which are subsequently classified as correct or not by an SVM. Using a k-nearest neighbor, naive Bayes, or an SVM classifier on selected features sets, using 67 (42 malignant, 25 benign) cases, a classification rate of 98.51% was achieved. George *et al.* [14]) used a circular Hough transform to detect cell nuclei, confirming these nuclei using thresholding and fuzzy c-means clustering. Twelve features were then passed to several neural network architectures using 92 (47 malignant, 45 benign) cases (and the WBCD database for comparison) with the best result being the probabilistic neural network with sensitivity of 95.49% and specificity of 83.16%.

It is important to notice that the above mentioned approaches have concentrated on classifying FNA slides as benign or malignant and are also called malignancy diagnosis. The system presented in the current study classifies a malignancy stage of cancer, called malignancy grading. The biopsy

being classified is nearly always malignant due to the pre-screening process before taking an FNA. Henceforth, in this paper, we are studying malignancy grading, not malignancy diagnosis.

2 Proposed Framework

In Fig. 1 the general concept of the proposed system is presented. It can be divided into two parts, Browser part and server part.

– **Browser part** – Its main task is to provide user with a set of operations which allow him to upload images to be processed and review results of classification. Its secondary task is to send data to a server in form of unprocessed images and retrieve processed results. It may be seen as a presentation layer of the system. The idea is that it is user friendly and intuitive.

– **Server part** – It is a core layer of the system. It performs all computational tasks including necessary calculations and features extraction. It handles all data structures essential for proper classification process. It is also easily extendable by other algorithms with visible separation between presentation, business and delegate layers.

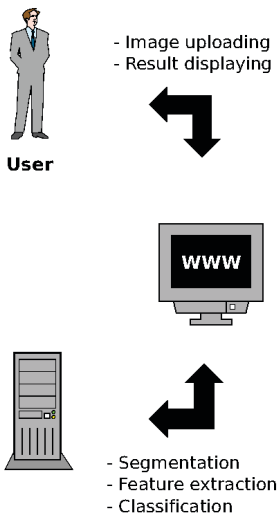

Figure 1. General concept of the automatic web–based classification system.

The proposed web–based framework is divided into three stages. The first is FNA cytological images segmentation followed by the feature extraction of the meaningful and indispensable features describing segmented nuclei. The output vector of extracted features is then transferred to the last part, a classifier which classifies an image into one of the two possible malignancy classes.

3 Segmentation

In this paper the focus is put on two image clustering segmentation algorithms and one region growing technique supported by histogram thresholding. The algorithms that were applied for the malignancy classification include a fuzzy c–means and k–means clustering as well as a watershed segmentation.

3.1 K–means clustering

One of the simplest unsupervised learning algorithms that solves clustering problem. The algorithm's input parameter is only a number of input clusters k which needs to be known before clustering process can begin. The main idea is to define k centroids, one for each cluster, which should be placed in cunning way since k means is a heuristic algorithm and there is no guarantee that it will converge to global optimum. Next step is to take each point belonging to a given data set and associate it to the nearest centroid. When all of the points are assigned, the first step of the algorithm is completed and an initial grouping is done. Following procedure is to re-calculate k new centroids as centers of the groups calculated initially. After that we have new k centroids and the association procedure for all of the data needs to be repeated [18]. The consecutive steps of a generated loop make the k centroids change their location until convergence is reached. The aim of this algorithm is to minimize an objective function which is a squared error function:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i^j - c_j||^2 \qquad (1)$$

where $||x_i^j - c_j||^2$ is a chosen distance measure between a data point x_i^j and the cluster centre, c_j is an indicator of the distance of the n data points from their respective cluster centers.

Here we use the RGB color distance between pixel and a mean cluster RGB color as a measure of distance. The initial clusters centroids are picked in random fashion. The segmentation result is picked as a cluster whose mean RGB value is the highest. After empirically testing different setting of clusters the conclusion was reached that the optimal number of clusters is 3. Higher value of k resulted in dispatching of a meaningful data, sometimes even creating holes inside nuclei or jagged groups. When using only 2 clusters, too much meaningless data was introduced into a segmented image and the result was not satisfactory. Three clusters is a trade off between processing meaningless data and discarding potentially important information.

3.2 Fuzzy c–means clustering

Similarly to k–means, a fuzzy c–means is a method of clustering but it allows one piece of data to belong to two or more clusters in a fuzzy logic fashion. In this algorithm each point has a degree of cluster membership rather than completely belonging to just one cluster as in the k-means segmentation. Because of that, it is possible that the points on the edge of the cluster belongs to the cluster in a lesser degree than those in the centre of it [4]. The method was developed by J. C. Dunn [8] and improved by J. C. Bezdek [5] and frequently used in pattern recognition. The objective of this algorithm is to minimize the following objective function:

$$J = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^{m} ||x_i - c_j||^2 \qquad (2)$$

where m is any real number greater than 1, u_{ij} is the degree of membership of x_i in cluster j, x_i is the i–th dimension of the d–dimensional measured data, c_j is the d–dimension centre of the cluster and $||*||$ is any norm expressing the similarity between any measured data and the centre.

Fuzzy partitioning is carried out by iterative optimization of the objective function with update of membership u_{ij} and the cluster centers c_j. The iterations stops when the error of the result is lower than set accuracy, or the number of iterations already computed is higher than maximum number of iterations set. The parameters of the algorithm are: desired accuracy, maximum number of iterations, number of clusters, and m fuzzy parameter which controls how much weight is given to the closest centre and must be greater or equal to 1.

3.3 Watershed segmentation

The watershed algorithm exploits the properties of grey-level images in a way that they may be seen as topographic relief. The grey level of a pixel is interpreted as an altitude in the relief. High intensity denotes peaks and hills while low intensity denotes valleys. The main idea is that each isolated valley (local minima) of the image is filled with different water color (label). When the water level rises connecting nearby peaks (gradients) it will merge with water of other color. In order to prevent that, the barriers are created in those locations where water merges. The process of flooding and constructing of barriers continues until all of the peaks are under water. Finally, the barriers created with the algorithm are a result of the segmentation process. Due to a noise or local irregularities in the gradient images it is common to over-segment an image [30].

In case of over-segmentation, which is a very common problem with watershed algorithm, results would not be very meaningful for a problem of nuclei segmentation. For this purpose a variation of the watershed method, called marker-controlled watershed, was applied. The principle here is the same but instead of flooding from local minima, a set of markers which will most certainly belong to a foreground is used as points of origin. That way the over-segmentation is prevented. The input markers for marker–controlled watershed are calculated according to the following:

1 Regions which will most certainly belong to a foreground are specified and labelled

2 Regions which will most certainly belong to a foreground or non-objects are specified and labelled

3 Remaining regions which we are uncertain are labelled

A process starts with the RGB image converted to a greyscale using Otsu's binarization. From the result of Otsu's binarization two images are created. The first is an image to which erosion was applied in order to remove the boundary pixels. After that, in order to isolate a foreground region, the distance

transform with a proper threshold was applied. The second image created from Otsu's thresholding output is a result of image dilation. With these two images, the points 1) and 2) above are completed and we can calculate remaining regions that cannot be associated to foreground nor background. The watershed algorithm is a solution to find them. These areas are normally around the boundaries of a foreground and background where the images meet. It can be obtained by subtracting these two areas. When the region labeling is done, the marker image is ready and can be passed to the watershed algorithm along with original image for segmentation.

4 Classification

Data classification is a process of identifying to which set of categories (sub-populations or classes) new observation belongs to. In image analysis it would be an operation of assigning one set of categories to a new image based on classification of a feature vector that was previously extracted based on the segmentation results [9].

Classification is a process based on a training set of data containing observations (or instances) whose category membership are known [15]. This means that in order to properly classify a new instance the classifier has to have some set of previously made observations (for instances in form of a database or a flat file) with instances that are a base of prediction. These individual observations are analyzed into a set of quantifiable properties. The properties can be variously categorized, for example as *ŇAÓ*, *ÓBÓ*, *ŇABÓ* or *Ň0Ó* for blood type. Depending on the application, more common types like integer-values or real-valued can also be assigned.

An algorithm that implements classification, especially in a concrete implementation, is known as a classifier. The term "classifier" sometimes also refers to the mathematical function, implemented by a classification algorithm that maps input data to a category.

In this study we have compared several classifiers which include Naïve Bayes classifier, Logistic regression, Decision Trees, Decision table and neural networks.

4.1 Naïve Bayes classifier

Naïve Bayes classifier belongs to a family of simple probabilistic classifiers that are based on a Bayes theorem assuming that there is a very strong (naive) independence between features. In other words, a naïve Bayes classifier assumes that a value of one of the features is unrelated to the presence or absence of any other feature, in scope of a class variable. For instance, an orange may be considered to be an orange if its color is orange, it is round, and about 3" in diameter. However this classifier considers each of these features to contribute independently to the probability that this fruit is an orange, not taking into consideration the presence or absence of other features. The advantage of this approach is that used in this way it only requires a small amount of training data to estimate the parameters necessary for classification. This is related to independency of variables that the algorithm assumes and we need to determine only the variances of the variables for each class and not the entire covariance matrix [26].

4.2 Logistic regression

This method is otherwise known as a logit regression which is a type of probabilistic statistical classification model used when categorical dependent variable (for instance a class label) can be only one of the two values (on dichotomous scale). Usually values of features describing some observation are based on occurrence or absence of some event that is the topic of prediction. In such a case by using logistic regression it is possible to calculate a probability of such an event. Formally, logistic regression model is a general case of a linear model in which the logit was used as a bounding function [26].

4.2.1 Logistic model trees

Logistic model trees are a type of a classification trees with logistic regression functions at the leaves. The only changed parameter, during generation of the LMT, is a minimal number of instances at which a node is considered for splitting and is set to 15.

4.2.2 Multinomial logistic regression

This is another classifier that uses logistic regression, a multinomial logistic regression in particular, with a ridge estimator. It is it allowed to perform an unlimited number of iterations and the log-likelihood value of ridge is set to 10^{-8}.

4.3 Decision trees

Decision trees are a type of a predictive model which maps observations of an item to conclusions about the item's target value. In this approach for the classification process a tree structure is used where leaves represent class labels and branches represent conjunctions of features that lead to those class labels [19].

This tool allows for visual and explicit representation of decisions and decision making process (see fig. 2). An insight into what features are taken into account during classification process and in what way can be made. Opposed to neural networks the decision trees are human readable and the process of classification can be understood without any problems and neural networks are more like black boxes.

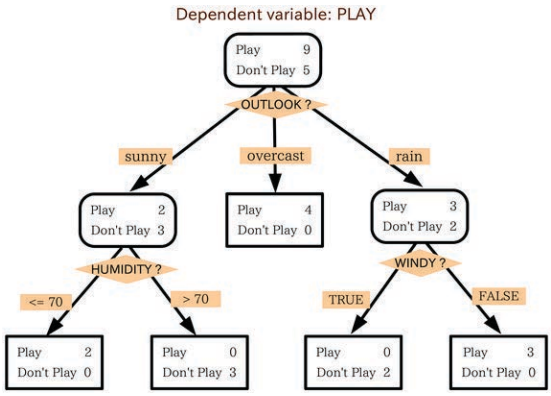

Figure 2. Example of a decision tree. Taken from Wikipedia.

Here we applied a C4.5, PART and a decision stump variants of a decision trees.

C4.5

This is an algorithm used to generate a decision tree developed by Ross Quinlan [28]. The C4.5 builds trees using the concept of information entropy. At each node of the tree, algorithm chooses the attribute of the data that most effectively splits its set of samples into subsets of each class. The criterion of splitting is the difference of entropy. The attribute with the highest entropy is chosen to make a decision. The C4.5 then recurs on the smaller subsets of data. In this work the following parameters were used for this algorithm:

- Confidence factor for pruning is equal to 0.25,
- Minimum number of instances per leaf is equal to 2,
- Number of data for reduced error pruning is equal to 3.

PART

PART is a decision list algorithm which uses divide-and-conquer technique. Builds a partial C4.5 decision tree in each iteration and makes the best leaf into a rule. The decision list is a representation of Boolean functions [29]. The parameters for generating the component decision trees of PART is the same as for the C4.5 algorithm.

Decision stump

This ia a machine learning model consisting of one-level decision tree. It makes predictions based on the value of a single input feature [17].

Depending on the input feature there are two possibilities for creating the stump:

- Creating a leaf for each possible feature value,
- Creating a leaf that corresponds to the one chosen category and the other leaf to all other categories.

4.4 Decision table

Decision tables are a precise and compact way of modelling complicated logic. They are similar to flowcharts and if-then-else set of statements which associate conditions with actions to be performed. Each decision corresponds to a variable, relation or predicate whose possible values are listed among the condition alternatives. Decision table is a hierarchical breakdown of the data with two attributes at each level of hierarchy. Decisions are made by the inducer the same way as in the decision tree, but the attributes are evaluated across the entire level of the tree rather than on a specific sub-tree. The result of

course is presented as a hierarchical table instead of a tree [13]. Parameters used for this algorithm are as follows:

– Number of folds for cross validation is equal to 10,

– The method used for finding good attribute combinations is BestFirst (greedy hillclimbing augmented with backtracking facility).

4.5 Neural networks

Here we have implemented a multilayer perceptron which is a feedforward artificial neural network model mapping sets of input data into a set of appropriate outputs. Perceptron is a function that maps input feature vector (real–values) to an output value f(x) (binary–values) [6]:

$$f(x) = \begin{cases} 1 & \text{if } w*x+b>0 \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where w is a vector of real–valued weights same size as the input features vector, $w*x$ is the dot product (weighted sum) and b is a bias – constant term independent from any input value.

Multilayer perceptron utilizes a supervised learning technique called backpropagation for training the network. Moreover it is a modification of a standard linear perceptron which is able to distinguish data that is not linearly separable [16].

The parameters for the applied MLP are following:

– 3 hidden layers,

– Learning rate, which is the amount the weights are updated, equals to 0.3,

– The momentum parameter, which is applied to the weights during the update, equals to 0.2.

5 Data set and feature set

5.1 Data set

The database used in this paper consists of 346 FNA images used for the breast cancer diagnosis with known malignancy grade. All of the images are stained with the HE technique (Haematoxylin and Eosin) which stains nuclei with purple and black color, cytoplasm with shades of pink and uses orange and red colors for red blood cells. At this point it has to be mentioned that the focus of this study was to classify the malignancy of the breast cancer. This is due to the fact that tissue collected during the FNA examination is always cancerous. Therefore, there is no need to check if the case is benign or malignant. It is more important to determine cancer's malignancy.

All of the images were digitalized with Olympus BX 50 microscope with mounted CCD-IRIS camera. The digitalization process was conducted at the Department of Pathology of the Wrocław Medical University, Poland with a help of a PC class computer with MultiScan Base 08.98 software. The images are recorded with a resolution of 764x571 pixels with a printing density of 96 dpi. Because at the Wrocław Medical University, Poland there was no low malignancy cases recorded since 2004, our database consist only of images with high (G3) and medium (G2) malignancy samples, therefore the classification is considering only these two cases.

5.2 Feature set

In order to obtain meaningful classification results, a set of features needs to be calculated from the segmented images. In this section a list of extracted parameters is discussed. To assure that the process of malignancy classification is performed only on important and necessary features, a vector of 25 features was built. The vector consists of both low and high magnification features (based on low and high magnification images). In this study the extracted features chosen for classification process are a mixture of features introduced by authors of [21] and [12] as an attempt to create larger vector utilizing advantages of both sets. In the end the following set of features was extracted:

1 **Low magnification features:**

– **Area of groups** – average number of nuclei pixels. This feature provides representation of the tendency of groups to create large groups. When this feature is large there is one couple of big groups present in the image.

- **Number of groups** – feature which determines the number of groups that were not discarded during the image segmentation process. High value of this feature suggests a large number of small groups in the image.
- **Dispersion** – statistical variation of cluster areas. Small values of this feature represents groups of similar size present in the image.

2 **High magnification features:**

- **Nuclei area** – same as area of groups but for high magnification images.
- **Perimeter of a nucleus** –length of a nuclear envelope. Computed as an average number of pixels in the group which has at least one neighboring pixel which is not a part of that group.
- **Convexity** – ratio of the nucleus area and its convex hull (minimal area of the convex polygon containing the nucleus).
- **X–centroid** – alias major axis length. The average of longest diameter of a nuclei. The length of the nuclei along the x axis.
- **Y–centroid** – alias minor axis length. The average of shortest diameter of a nuclei. The length of the nuclei along the y axis.
- **Orientation** – calculated from the binary representation of the nucleus and image momentum.
- **Vertical projection** – average sum of all segmented pixels on y axis in horizontal direction.
- **Horizontal projection** – the average sum of all segmented pixels on x axis in vertical direction.
- **Luminance mean** – average luminance of all segmented nuclei groups in the image.
- **Luminance variance** – statistical variation of luminance for each group.
- **Eccentricity** – measure of how much the nuclei deviates from the circle. Calculated from image moments.
- **Distance from weight centroid** – for the need of this feature a segmented image binary centroid coordinates are calculated. Using the coordinated a distance to each nuclei is calculated as na average Euclidean distance between the centroid and the nucleus.
- **Distance from color centroid** – calculated as an average of the distance between the color cluster meant used during the segmentation and the subsequent groups of average color.

3 **Original image features:**

- **Histogram mean** – set of three features extracted as a histogram mean of a Red, Green and Blue channels.
- **Histogram energy** – set of three features where the histogram energy for each RGB channel is calculated.
- **Histogram variance** – statistical variation of histogram mean. Calculated for each channel separately.

6 Results

In this section we will present the results obtained in this study. The first set of results is devoted to segmentation. There are a couple of things that can be noticed based on segmentation results (see Fig. 3 and 4). First observation is that the watershed segmentation algorithm, even with the markers approach, is an algorithm with the least precision for the task at hand. It discards whole meaningful elements in the image and even makes holes in the properly detected nuclei. Its usefulness for segmenting low magnification images is questionable, however the output for the high magnification image is better but not ideal.

Another observation is that k–means and fuzzy c–means algorithms provided similar results. The reason behind that is that they are based on the same principle. However when taking a closer look at the results it became obvious that fuzzy c–means is slightly better and more accurate. It provides a less jagged borders than in the other two methods. It is also better at recognizing similar parts in the original image where other two algorithms tend to classify background data as nuclei. This is mostly visible when low magnification results are compared. K–means is classifying some parts of the image which are a little darker than the surrounding pixels (but not being nuclei) while fuzzy c–means properly recognizes them as background and discards them.

What is also worth noticing is that the fuzzy c–means, despite being the best algorithm for segmentation out of the three investigated methods, is also the one which has the highest computational times. Somewhere around 45 seconds for an image of a size 764x571 pixels is really high compared to the other algorithms which take no longer than 5 seconds. For watershed segmentation we noted a time not longer than 1 second. The difference in quality between k–means and fuzzy c–means is really small, but the gain on the performance when using the k-means is really high. Here we checked if classification results support this conclusion. This is why we constructed a feature vector based on the achieved segmentations. The feature that where used to build the feature vector was described in section 5.2. In table 1 an example of such a vector is presented for all the segmentation algorithms. In this case the same segmentation algorithm was used to calculate both the low and high magnification features. From that table we can notice that watershed algorithm is not the best choice for the task of automatic segmentation. In comparison with the remaining two algorithms is significant. Another fact is that the fuzzy c–means is better at picking clusters because the average distance from the RGB centroid of particular groups in the segmented image and luminance variance of those groups is less than 10^{-3}.

The last set of results to be presented in this work is the comparison of the accuracy of applied classification algorithms. Table 2 contains performance results of used classifiers for different combinations of segmentation algorithms. The results were obtained by using the 10–fold cross validation technique which assesses how the results of statistical analysis will generalize to new and independent data set. Its purpose is to check the model against the overfitting problem which occurs when model is too dependent on the training data set.

7 Conclusion

In this work three segmentation algorithms and classifiers were compared for the problem of creation of a web-based decision supporting system for automatic breast cancer malignancy grading. Whole process of decision making starting with image acquisition moving into image segmentation and feature extraction, ending with classification step was described. Algorithms chosen for comparison are a result of scrupulous literature review and showed to be very precise in the described application. The suggested feature vector obtained from segmented images allows for a high quality classification of FNA breast cancer images.

The main conclusion is that the described approach provided promising results. The error rate around 15 % for most of the cases indicates that the problem of automatic classification of breast cancer FNA images can be resolved by the proposed solution but it still needs some improvements to be more accurate. Not all of the combinations and classifiers however are as good as the others. The best two are the multilayer perceptron and logistic regression. Both are in a scope of minimizing the error rate and providing the best prediction for G3 cases being classified as G3. Other algorithms were good for minimizing the error rate but due to the fact that the data was severely unbalanced (136 samples of G2 and only 37 samples of G3) they mostly failed with correct classification of G3 samples. The optimistic results were also obtained with the C4.5 decision tree algorithm which shows a room for future improvement.

For the segmentation task, the following combinations of algorithms showed the best nuclei representation:

- Watershed for low magnification and k–means for high magnification (MLP - 89.02 %; Logistic regression 83.81 %)
- Fuzzy c–means for low magnification and k–means for high magnification (MLP - 88.44 %; Logistic regression 83.24 %)
- K–means for low magnification and fuzzy c–means for high magnification (MLP - 87.28 %; Logistic regression 84.97 %)
- Only k–means (MLP - 88.44 %; Logistic regression 83.81 %)
- Only fuzzy c–means (MLP - 88.44 %; Logistic regression 86.70 %)

Another conclusion is that the increase of recognition rate of minority class in almost all cases leads to decreased accuracy for majority class objects. However as stated before the early detection of high

malignancy breast cancer is vital for the life of patients and provides some means of efficient treatment. Therefore that trade-of is worth its cost.

The accuracy of the obtained results is very promising despite the fact that the proposed methods are not able to handle properly all of the test cases. The detection of overlapping nuclei in the images could be improved as well as the brightness of the images. The segmentation quality could also improve by the introduction of pre-processing methods to the input images which should resolve most of the mentioned problems. Also a recognition rate could be improved by the introduction of more attributes with high classification power to the feature vector that were not tested in this study.

References

[1] UCI machine learning repository.

[2] National Cancer Registry, The Maria Skłodowska – Curie memorial Cancer Center, Department of Epidemiology and Cancer Prevetion, December 2013.

[3] TNM breast cancer staging, December 2014.

[4] M.N. Ahmed, S.M. Yamany, N. Mohamed, A.A. Farag, and T. Moriarty. A modified fuzzy c-means algorithm for bias field estimation and segmentation of mri data. *IEEE Transactions on Medical Imaging*, 21:193–199, 2002.

[5] J.C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.

[6] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[7] H.J.G. Bloom and W.W. Richardson. Histological grading and prognosis in breast cancer. *British Journal of Cancer*, 11:359–377, 1957.

[8] J.C. Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3:32–57, 1973.

[9] A. Ethem. *Introduction to Machine Learning*. MIT Press, Boston, 2010.

[10] J. Ferlay, I. Soerjomataram, M. Ervik, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D.M. Parkin, D. Forman, and F. Bray. Cancer incidence and mortality worldwide. *IARC Cancer Base*, No. 11, 2012.

[11] P. Filipczuk, T. Fevens, A. Krzyzak, and R. Monczak. Computer-aided breast cancer diagnosis based on the analysis of cytological images of fine needle biopsies. *IEEE Transactions on Medical Imaging*, PP(99):1–1, 2013.

[12] P. Filipczuk, M. Kowal, and A. Obuchowicz. Fuzzy clustering and adaptive thresholding based segmentation method for breast cancer diagnosis. *Computer Recognition Systems*, 4(5):613–622, 2011.

[13] D.L. Fisher. Data, documentation and decision tables. *Comm ACM*, 9(1):26–31, 1966.

[14] Y.M. George, H.H. Zayed, M.I. Roushdy, and B.M. Elbagoury. Remote computer-aided breast cancer detection and diagnosis system based on cytological images. *IEEE Systems Journal*, PP(99):1–16, 2013.

[15] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning, 2nd. edition*. Springer, New York, 2009.

[16] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall, 1998.

[17] R.C. Holte. Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, 11(1):63–90, 1993.

[18] T. Kanungo, D. M. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Y. Wu. An efficient k-means clustering algorithm: Analysis and implementation. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 881–892, 2002.

[19] S.B. Kotsiantis. Supervised machine learning: A review of classification techniques. *Informatica*, pages 249–268, 2007.

[20] B. Krawczyk and P. Filipczuk. Cytological image analysis with firefly nuclei detection and hybrid one–class classification decomposition. *Engineering Applications of Artificial Intelligence*, 31:126–135, 2014.

[21] B. Krawczyk, Ł. Jeleń, A. Krzyżak, and T. Fevens. Oversampling methods for classification of imbalanced breast cancer malignancy data. *Lecture Notes in Computer Science (LNCS)*, 7594:483–490, 2012.

[22] B. Krawczyk and G. Schaefer. A hybrid classifier committee for analysing asymmetry features in breast thermograms. *Applied Soft Computing*, 20:112–118, 2014.

[23] Jihene Malek, Abderrahim Sebri, Souhir Mabrouk, Kholdoun Torki, and Rached Tourki. Automated breast cancer diagnosis based on gvf-snake segmentation, wavelet features extraction and fuzzy classification. *Journal of Signal Processing Systems*, 55(1-3):49–66, 2009.

[24] O.L. Mangasarian, R. Setiono, and W.H. Wolberg. Pattern Recognition via Linear Programming: Theory and Application to Medical Diagnosis. *Large-Scale Num. Opt., Philadelphia: SIAM*, pages 22–31, 1990.

[25] A. Marcano-Cedeño, J. Quintanilla-Domínguez, and D. Andina. WBCD breast cancer database classification applying artificial metaplasticity neural network. *Expert Systems with Applications*, 38(8):9573 – 9579, 2011.

[26] T. Mitchell. *Machine Learning, Generative and Discriminative Classifiers: Naive Bayes and Logistic Regression (Draft Version)*. McGraw Hill, 2005.

[27] S.I. Niwas, P. Palanisamy, and K. Sujathan. Wavelet based feature extraction method for breast cancer cytology images. In *IEEE Symposium on Industrial Electronics Applications (ISIEA)*, pages 686–690, Oct 2010.

[28] J.R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993.

[29] R.L. Rivest. Learning decision lists. *Machine Learning*, 2:229–246, 1987.

[30] J.B.T.M Roerdink and A. Meijster. The watershed transform: definitions, algorithms, and parallelization strategies. *Fundamenta Informaticae*, 41:187–228, 2000.

[31] W.N. Street, W.H. Wolberg, and O.L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In *IS&T/SPIE Inter. Symp. on Electronic Imaging: Science and Technology*, volume 1905, pages 861–870, 1993.

[32] W.H Wolberg and O.L. Mangasarian. Multisurface Method of Pattern Separation for Medical Diagnosis Applied to Breast Cytology. *Proceedings of National Academy of Science, USA*, 87:9193–9196, 1990.

[33] Xiangchun Xiong, Yangon Kim, Yuncheol Baek, Dae Wong Rhee, and Soo-Hong Kim. Analysis of breast cancer using data mining & statistical techniques. In *Proc. 6th Int. Conf. on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing and 1st ACIS Int. Worksh. on Self-Assembling Wireless Networks*, pages 82–87, 2005.

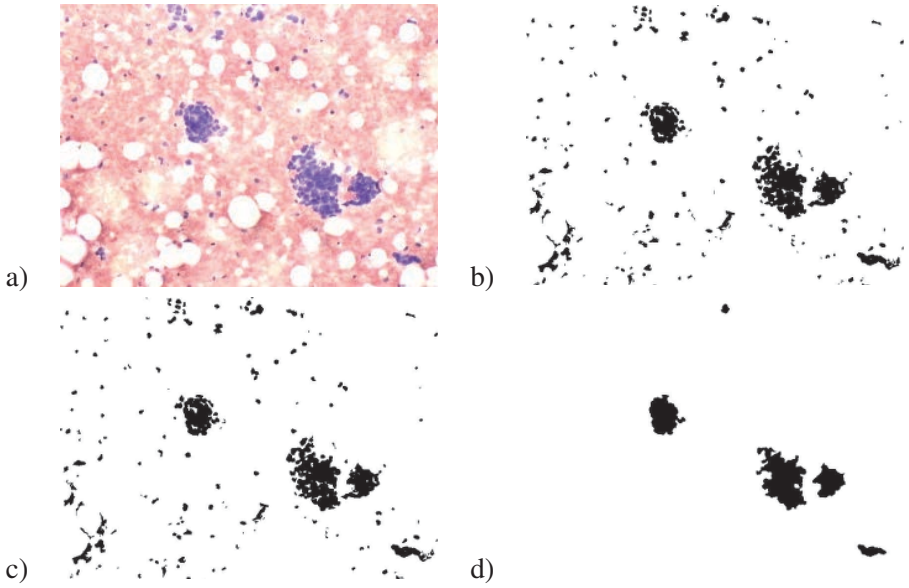

Figure 3. Low magnification segmentation results. a) Original image, b) K–means segmentation, c) Fuzzy c–means segmentation, d) Watershed segmentation.

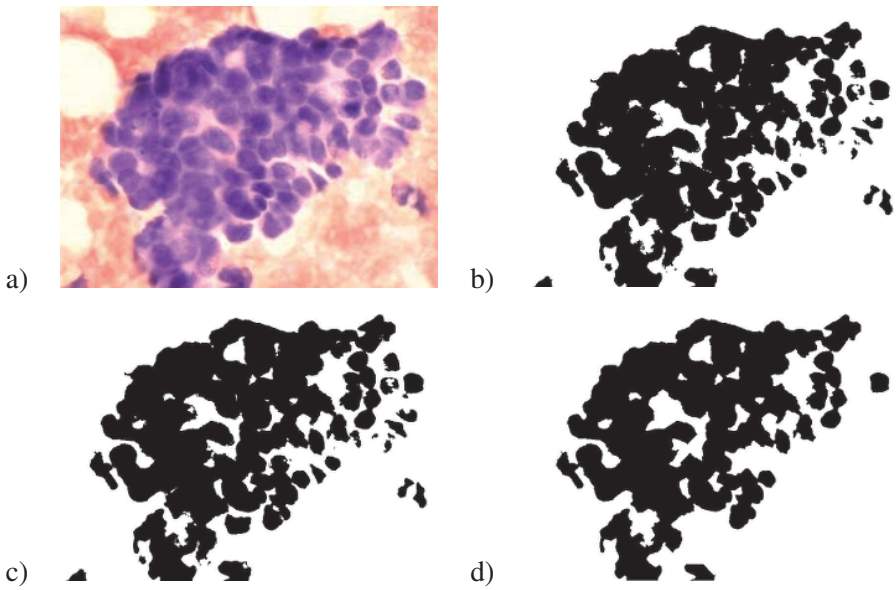

Figure 4. High magnification segmentation results. a) Original image, b) K–means segmentation, c) Fuzzy c–means segmentation, d) Watershed segmentation

Table 1. Sample results of feature extraction for three segmentation algorithms.

Feature	K–means	Fuzzy c–means	Watershed
Groups area [px]	360.9	387.8	6893.2
Number of groups	118	110	5
Dispersion	2137	2226	8010
Nuclei area [px]	1959.3	1836.8	3079.8
Perimeter [px]	236.0	220.5	323.3
Convexity	0.918	0.927	0.873
X–centroid	52.00	49.36	70.64
Y–centroid	46.75	44.60	64.51
Orientation	0.526	0.527	0.417
Vertical projection	82.64	80.70	178.62
Horizontal projection	61.55	60.11	133.03
Luminance mean	153.42	146.95	171.17
Luminance variance	10.54	0.00	20.36
Eccentricity	0.039	0.039	0.152
Distance from centroid	363	369	359
Histogram mean for:			
R channel	246.2	246.2	246.2
G channel	209.83	209.83	209.83
B channel	199.18	199.18	199.18
Histogram variance for:			
R channel	21.93	21.93	21.93
G channel	40.22	40.22	40.22
B channel	28.57	28.57	28.57
Histogram energy for:			
R channel	0.22	0.22	0.22
G channel	0.037	0.037	0.037
B channel	0.014	0.014	0.014
Distance from centroid RGB	7	0	–
Computing time [ms]	1220	38593	156

Table 2. Error rates for different segmentation set–ups.

Segmentation set–up	C4.5	PART	Decision table	Decision table	Multilayer perceptron	LMT	Logistic	Naïve Bayes
Low magnification watershed, High magnification Fuzzy c–means	16.18 %	15.61%	15.03 %	16.18 %	15.61 %	12.14 %	14.45 %	17.92 %
Low magnification watershed, High magnification K–means	18.50 %	18.50%	16.19 %	17.34 %	10.98 %	11.56 %	16.18 %	17.34 %
Low magnification Fuzzy c–means, High magnification watershed	16.76 %	20.23%	17.92 %	25.43 %	15.03 %	15.61 %	15.61 %	35.84 %
Low magnification Fuzzy c–means, High magnification K–means	23.12 %	15.61%	20.23 %	23.12 %	11.56 %	18.50 %	16.76 %	19.07 %
Low magnification K–means, High magnification watershed	16.76 %	17.91%	17.91 %	25.43 %	14.45 %	15.02 %	16.18 %	36.42 %
Low magnification K–means, High magnification Fuzzy c–means	20.23 %	20.81%	20.23 %	21.39 %	12.72 %	19.07 %	15.03 %	18.50 %
Low magnification and High magnification watershed	15.03 %	15.03%	16.18 %	18.50 %	16.76 %	14.45 %	16.76 %	38.15 %
Low magnification and High magnification Fuzzy c–meanss	17.92 %	18.50%	18.50 %	21.39 %	11.56 %	16.76 %	13.29 %	19.07 %
Low magnification and High magnification K–means	19.65 %	16.18%	17.34 %	17.92 %	11.56 %	15.61 %	16.18 %	17.92 %