Remote Computer-Aided Breast Cancer Detection and Diagnosis System Based on Cytological Images

Yasmeen Mourice George, Hala Helmy Zayed, Mohamed Ismail Roushdy, and Bassant Mohamed Elbagoury

***Abstract*—The purpose of this study is to develop an intelligent remote detection and diagnosis system for breast cancer based on cytological images. First, this paper presents a fully automated method for cell nuclei detection and segmentation in breast cytological images. The locations of the cell nuclei in the image were detected with circular Hough transform. The elimination of false-positive (FP) findings (noisy circles and blood cells) was achieved using Otsu's thresholding method and fuzzy c-means clustering technique. The segmentation of the nuclei boundaries was accomplished with the application of the marker-controlled watershed transform. Next, an intelligent breast cancer classification system was developed. Twelve features were presented to several neural network architectures to investigate the most suitable network model for classifying the tumor effectively. Four classification models were used, namely, multilayer perceptron using back-propagation algorithm, probabilistic neural network (PNN), learning vector quantization, and support vector machine (SVM). The classification results were obtained using tenfold cross validation. The performance of the networks was compared based on resulted error rate, correct rate, sensitivity, and specificity. Finally, we have merged the proposed computer-aided detection and diagnosis system with the telemedicine platform. This is to provide an intelligent, remote detection, and diagnosis system for breast cancer patients based on the Web service. The proposed system was evaluated using 92 breast cytological images containing 11 502 cell nuclei. Experimental evidence shows that the proposed method has very effective results even in the case of images with high degree of blood cells and noisy circles. In addition, two benchmark data sets were evaluated for comparison. The results showed that the predictive ability of PNN and SVM is stronger than the others in all evaluated data sets.**

***Index Terms*—Circular Hough transform (CHT), computer-aided detection and diagnosis (CADx), fine-needle aspiration cytology (FNAC), fuzzy c-means (FCM) clustering, learning vector quantization (LVQ), marker-controlled watershed transform, multilayer perceptron (MLP), Otsu's thresholding method, probabilistic neural network (PNN), support vector machine (SVM).**

Manuscript received July 23, 2012; revised February 12, 2013; accepted August 9, 2013. Date of publication October 22, 2013; date of current version August 21, 2014.

Y. M. George is with the Computer Science Department, Faculty of Computer and Informatics, Benha University, Qalyubiyah, Egypt (e-mail: yasmeen.mourice@fci.bu.edu.eg).

H. H. Zayed is with the Computer Science Department, Faculty of Computer and Informatics, Benha University, Banha 13511, Egypt (e-mail: hala.zayed@fci.bu.edu.eg).

M. I. Roushdy and B. M. Elbagoury are with the Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo 11566, Egypt (e-mail: mroushdy@cis.asu.edu.eg; Bassantai@yahoo.com).

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

Digital Object Identifier 10.1109/JSYST.2013.2279415

I. Introduction

BREAST cancer is the top cancer in women both in the developed and the developing world [1]. Early detection of cancer greatly increases the chances for successful treatment. The number of victims of this deadly cancer is a constant reminder that new approaches capable of improving patient survival are still desperately needed. The integration of computer models into the radiological imaging interpretation process can increase the accuracy of image interpretation. Palpable breast lesions can be accurately diagnosed by preoperative tests (such as physical examination, mammography, fine-needle aspiration cytology (FNAC), and core needle biopsy) [2] and [3]. Mammography is most often used for screening purposes rather than for precise diagnosis. It allows a physician to find possible locations of microcalcifications and other indicators in breast tissue. When a suspicious region is found, the patient is sent to a pathologist for a more precise diagnosis. This is when the FNA is taken. FNA provides a way to examine a small sample of the questionable breast tissue that allows the pathologist to describe the type of the cancer in detail. It has gained popularity due to its fast and easy approach, being inexpensive [4] and [5].

Computer-aided detection and diagnosis (CADx) is becoming an increasingly important tool to assist radiologists in the breast cancer detection and diagnosis. Current CADx systems in clinical use serve as a second reader for breast cancer detection. Numerous CADx models are being developed to help in breast US and MRI interpretation [6]. However, developing CADx systems for classification of malignant and benign lesions based on FNAC are under development by a number of research groups [7]. Any CADx system consists of mainly two phases, namely, the segmentation phase and the classification phase. Implementing a complete CADx system based on FNAC image has not finished yet; however, each phase has been individually carried out by many researchers.

For the first phase that is related to the detection and segmentation systems based on breast FNAC images, several approaches have been proposed. The literature of microscopic cellular image detection and segmentation approaches is shown in Table I. As it can be observed, some methods usually work in practice under given assumptions and/or need the end-user's interaction/cooperation [8]–[10]. However, the user interaction hinders the automated cell image analysis. Other methods do not deal with poor quality and noisy images [11] and [12]. This is not acceptable as most of cytological images have low quality. In addition, the overlapped cells problem is not

TABLE I
ADVANTAGES AND LIMITATIONS OF STATE-OF-THE-ART METHODS FOR BREAST CELL NUCLEI DETECTION AND SEGMENTATION

Method [ref] Year	Advantages	Limitations
Chang Wen Chen et al. [8] 1998	Uses a combination of adaptive k-means clustering and knowledge based morphological operations for 3D data segmentation. Closed boundaries.	Does not remove false findings (cytoplasm cells). Does not handle overlapped cells. Grayscale images Requires a-prior knowledge about the object for accurate segmentation
William N. Street et al. [9] 2000	Cell nuclei detection based on generalized Hough transform. Remote Cytological Diagnosis and Prognosis of Breast Cancer. Accurate closed boundaries.	Does not remove false findings (cytoplasm cells). Does not handle overlapped cells. Grayscale images. Requires user intervention.
Lin Yang et al. [15] 2005	Based on color gradient vector flow (GVF) active contour model for segmentation. High rate of active segmentation.	One nucleus per slide. Does not handle overlapped cells.
Anne E Carpenter et al. [13] 2006	Closed boundaries. Measures a number of accurate cell features. The first free open source system for cell image analysis.	Does not remove false findings (cytoplasm cells). Grayscale images.
Metin N. Gurcan et al. [14] 2006	Automated cell nuclei segmentation method based on morphological operations. Closed boundaries.	Does not remove false findings (cytoplasm cells). Does not handle overlapped cells.
Xiaodong Yang et al. [18] 2006	Automated cell nuclei segmentation based on marker controlled watershed algorithm. Prevents over segmentation.	One nucleus per slide. Does not handle overlapped cells.
Maciej Hrebien et al. [12] 2008	Hough transform adopted for circle detection in pre-segmentation stage. Automatic nuclei localization based on (1+1) search strategy. Cell segmentation using watershed, active contours and cellular automata GrowCut algorithms separately.	Does not remove false findings (cytoplasm cells). Does not handle overlapped cells. Does not deal with poor quality and noisy images.
Xiaobo Zhou et al. [16] 2009	Automated analysis system for cell nuclei segmentation using adaptive thresholding and watershed algorithm. Closed boundaries.	Over segmentation. Does not remove false findings (cytoplasm cells). Does not handle overlapped cells. Grayscale images
Jierong Cheng et al. [10] 2009	Clustered nuclei segmentation using shape markers and marking function in a watershed-like algorithm. Prevents over segmentation.	Does not remove false findings (cytoplasm cells). Does not handle overlapped cells. Grayscale images Requires user intervention in case of very irregular boundary.
Marina E. Plissiti et al. [11] 2011	Cell nuclei segmentation based on morphological reconstruction and watershed transform algorithm. Elimination of false findings based on shape and texture features. Prevents over segmentation.	Does not deal with noisy images. The dataset images do not contain blood cells.

considered in many methods [11]–[16]. These methods deal with one cell image or isolated cells to avoid the problem of identifying two merged cells as one object. In addition, the FP findings elimination problem is not performed by many methods. This problem deals with removing the blood cells and cytoplasm boundaries. However, handling both the overlapped cells and also the elimination of FP findings are essentially for effective cell nuclei diagnosis. Some methods deal with cytoplasm markers elimination but do not deal with blood cell markers elimination [11] and [17]. Another problem, missing color information in many methods [8]–[10], [13], [16] and [18] by converting the color image into gray scale image. This neglected information affects the segmentation results.

For the second phase that is related to the breast FNAC diagnosis systems, many researchers have carried out intelligent diagnostic systems specifically to provide "second opinion" for pathologists in making diagnosis based on FNAC images [19]–[21], as shown in Table II. One can find approaches to breast cancer classification, namely, k-nearest neighbors, support vector machines (SVMs), multilayered perceptron, radial basis network, general regression neural network, and probabilistic neural network (PNN). Some of the mentioned approaches are concentrated on classifying FNA slides based on an existing benchmark data set [22]. Other techniques involve images containing isolated cells in the diagnosis system [9], [23] and [24], which are not applicable in real life. The classification techniques are based on a number of cell features for the characterization of a cell as normal or abnormal [25] and [26]. However, some techniques provide a small number of features that affects the classification results [27].

The purpose of this study is to develop a CADx system for breast cancer patients. Fig. 1 shows the phases of the

TABLE II
ADVANTAGES AND LIMITATIONS OF STATE-OF-THE-ART METHODS FOR BREAST CELL NUCLEI DIAGNOSIS

Method [ref] Year	Advantages	Limitations
Marek Kowal et al. [23] 2011	Tumors are classified using four different classification methods: k-nearest neighbors, naive Bayes, decision trees and classifiers ensemble.	does not detect and split the overlapped cells More sophisticated features need to be extracted Does not deal with high resolution images Does not be compared with benchmark datasets
Lukasz Jelen et al. [27] 2008	Classification of fine needle aspiration biopsy tissue based on support vector machines.	Only five features are extracted for each slide. Only SVM classification model is performed Does not be compared with benchmark datasets
Shekhar Singh et al. [24] 2011	Classifies input features into benign, malignant and also classifies malignant tumor in three types based on feed forward back propagation neural network.	Features are extracted from images containing isolated single cell. Only feed forward back propagation neural network is performed. Does not be compared with benchmark datasets.
Isa et al. [21] 2007	Uses the hybrid multilayered perceptron for classification of FNAC images based on thirteen extracted feature. The training is based on the modified recursive prediction error.	Only one classification model is performed. Does not be compared with benchmark datasets.
Tüba Kiyan et al. [22] 2004	Wisconsin breast cancer data (WBCD) classification based on radial basis network, general regression neural network and probabilistic neural network.	Uses benchmark dataset for classification.
William N. Street [9] 2000	provides remote predictive analysis for breast cancer diagnosis and prognosis	Features are extracted from images containing one single cell.

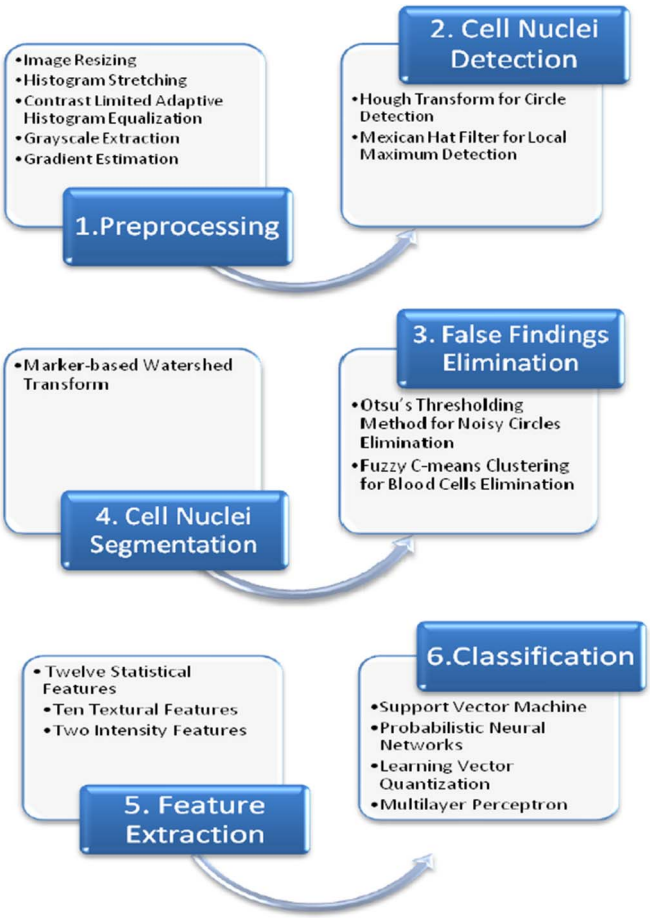


Fig. 1. Proposed CADx system phases.

overall proposed CADx system. Moreover, this system will be merged with the telemedicine platform. This is to provide an intelligent remote CADx system. The proposed detection and segmentation method deal with most of the mentioned problems related to cell image analysis. In addition, the cell features are extracted from our data set images. Moreover, the images contain overlapped cells and cell clusters. The slides are classified using different classification networks to find the optimum classification model. In addition, the classifiers are evaluated using six different data sets to find a general classification model for all these data sets.

The rest of this paper is divided into six sections. Section II describes the process of acquisition of images used in our system. Section III presents the methods performed in the detection and segmentation stage, including the preprocessing techniques and false findings elimination. Section IV shows the feature extraction and classification phase. Section V proposes the design and implementation of the remote diagnostic system. Section VI shows the experimental results and discussion. The last part of the work includes a conclusion, future work and references. These sections are described in detail in the following paragraphs.

II. DATA SET DESCRIPTION

The studied data set was based on microscopic images of breast FNAC specimens obtained in cooperation with specialists from the archive of Early Cancer Detection Unit-Obstetrics and Gynecology Department, Ain Shams University Hospitals. Samples were taken from breast lumps using 23-22G needle and spread on glass slides, stained with May Grunwald Giemsa stain or Diff. Quick stain. The data set consists of 92 FNAC images, including 45 images of benign tumors, and 47 of malignant tumors. The total number of cell nuclei in the images was 11 502, including 2474 benign nuclei, 2634 malignant nuclei, and 6394 blood cells. The images were acquired through an Olympus digital camera adapted to a trinocular optical microscope. Images were captured using $10\times$ and $40\times$ magnification lens. The size of the acquired images was 2560 $\times$ 1920. The images were stored in JPG format. The image itself was coded using the RGB color space and was not subject to any kind of lossy compression.

III. Preprocessing and Segmentation Phase

A. Preprocessing

As the processing time is a very essential factor in image processing, we resized the images from a resolution of 2560 × 1920 pixels to 640 × 480 pixels. Then, a contrast enhancement and edge sharpening technique is applied as a great deal of images has a low contrast. In this paper, we use simple histogram processing with a linear transform of the image levels of intensities, namely, the cumulated sum approach [28] with 1% saturation at low and high intensities of the input image. After applying histogram stretching, we apply the contrast-limited adaptive histogram equalization (CLAHE) [29] to improve the quality of images. CLAHE operates on small regions in the image, which are called tiles, rather than the entire image. Each tile's contrast is enhanced. The contrast correction is conducted for each color channel separately resulting in an image being better defined for later stages of the presented detection and segmentation methods.

After enhancing the image, a gray scale was extracted from the colored image to be used in the next steps of the proposed method. As the color components of an image do not carry important information as the luminosity does, they can be removed to reduce processing complexity in stages that require only, e.g., gradient estimations. An RGB color image can be converted to gray scale by removing blue and red chrominance components from the image defined in the YCbCr color space, leaving only the luminosity one. The luminosity component can be determined using [30], i.e.,

$$Y = 0.299R + 0.587G + 0.114B. \tag{1}$$

Finally, the gradient image is estimated as it will be used in the nuclei detection and segmentation stages. We use a gradient image as the feature indicating nucleus occurrence or absence in a given fragment of the cytological image. The gradient image is a saturated sum of gradients estimated in eight directions on the gray-scale image prepared in the previous step. The base gradients are calculated using Roberts and Sobel mask methods [28]. Exemplary results of the preprocessing stage are illustrated in Fig. 2.

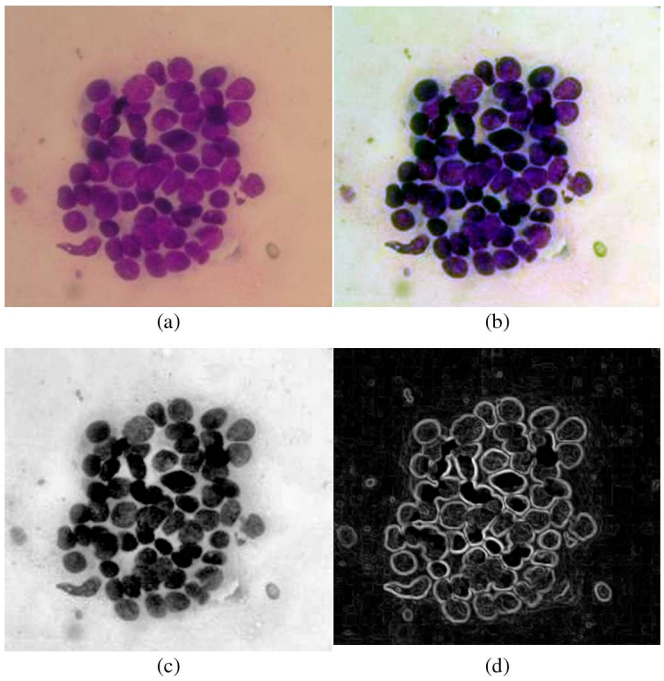


Fig. 2. Results of the preprocessing stage. (a) Original image after the resizing. (b) Enhanced image using histogram stretching and CLAHE technique. (c) Y level image. (d) Feature image g (gradient image).

B. Cell Nuclei Detection

The prerequisite for the avoidance of oversegmentation that the watershed transform [31] produces is the detection of nuclei markers, which will be used as starting points in the flooding process, for the determination of the boundaries of the nuclei. For the detection of nuclei locations, we perform circular Hough transform (CHT) and Mexican hat filter.

1) Circle Detection: The nuclei we have to segment have an elliptical shape. Most of them resemble an ellipse but, unfortunately, the detection of the ellipse is computationally expensive as the ellipse has the parametric equations $x = a\cos(\alpha)$, $y = b\sin(\alpha)$ where a and b can be additionally rotated. On the other hand, the shape of the ellipse can be approximated by a given number of circles. The detection of circles is much simpler in the sense of the required computations because we have only one parameter, which is the radius R. These observations and simplifications form a basis for a nucleus detection algorithm. In our approach, we try to find such circles with different radii in a given feature space.

The Hough transform [32] can be easily adopted to determine the parameters of a circle when a number of points that fall on the perimeter are known. A circle with radius R and center (a, b) can be described with the parametric equations $x = a + R\cos(\theta)$, $y = b + R\sin(\theta)$. The task of the search program is to produce parameter triplets (a, b, R) to describe each circle.

The Hough transform in a 2-D discrete space can be written as

$$\mathrm{HT}_{\mathrm{discr}}(R, i_0, j_0) = \sum_{i=i_0-R}^{i_0+R} \sum_{j=j_0-R}^{j_0+R} g(i,j)\partial\left((i-i_0)^2 + j - j02 - R2\right) \tag{2}$$

where g is a 2-D feature image. In this paper the feature image is the gradient image prepared in the preprocessing stage, and ∂ is Kronecker delta (equal to unity at zero). $\mathrm{HT}_{\mathrm{discr}}$ plays the role of an accumulator that accumulates the similarity levels of g to the circle placed at (i_0, j_0) and defined by the radius R.

2) Finding Local Maximum Points (Nuclei Markers): After finding the accumulator array in the circle detection step, it is necessary to concentrate the hot spots in the accumulator array as it corresponds to the centers of the nuclei. To achieve this, at first, we threshold the values in the accumulator by a given T value to obtain each cell nucleus as a connected component. Next, we calculate the bounding box containing each connected component, and we define the corresponding subimage in the accumulator array. Finally, each subimage is filtered with

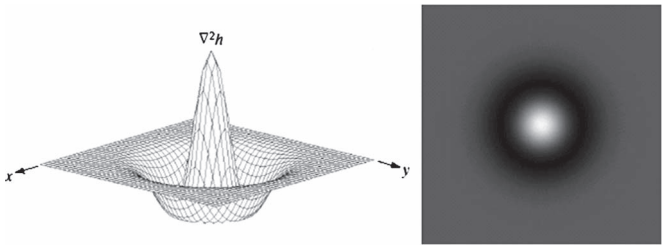
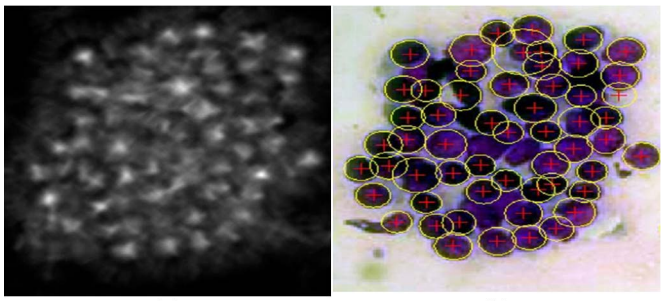


Fig. 3. Mexican hat graph.



(a) (b)

Fig. 4. Cell nuclei detection results. (a) Accumulator array. (b) Detected positions resulted from local maximum filtering using 17 × 17 Mexican hat mask.

17 × 17 Mexican hat mask (see Fig. 3) to get the local maximum point in each nucleus. Sample results of the cell nuclei detection method are shown in Fig. 4.

C. False Findings Elimination

In the FNAC image, there exists some blood cells that appeared through the staining process. Moreover, there are some arcs between cell nuclei that are approximated by CHT as a circle. In the last stage, we have detected the candidate nuclei markers in the FNAC image. These markers are either a tumor cell or a blood cell or a noisy circle. We need only the tumor markers to be used as starting points in the flooding process when applying the watershed transform. Hence, it is essential to remove the unwanted markers, which correspond to blood cells and noisy circles. For this purpose, we perform Otsu's thresholding method to eliminate the noisy circles then we proceed with the application of classification algorithms for the separation between the points of true nuclei and blood cells.

1) Otsu's Thresholding Method: We use Otsu's thresholding method [33] for removing the markers that belong to noisy circles. We produce a binary mask BW with the regions of interest in the image by thresholding the gray-scale image prepared in the preprocessing stage. We compute a threshold (level) that can be used to convert an intensity image to a binary image using Otsu's method, which chooses the threshold to minimize the intra class variance of the black and white pixels. Then, for each detected marker, we get the corresponding point in the binary mask BW and remove the marker if its corresponding point belongs to the background area.

2) Fuzzy C-Means Clustering:

a) Background: Fuzzy c-means (FCM) is a method of clustering, which allows one piece of data to belong to two or more clusters. It is based on minimization of the following objective function [17] and [34]:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^m \|X_i - C_j\|^2, \qquad 1 \le m < \infty \tag{3}$$

where m is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j, x_i is the ith of d-dimensional measured data, c_j is the d-dimension center of the cluster, and $\|*\|$ is any norm expressing the similarity between any measured data and the center. We have randomly generated a fuzzy partition matrix $U[\text{Clusters\_Num} \times \text{Data\_Num}]$, where Clusters_Num is the number of clusters, and Data_Num is the number of data points. The summation of each column of the generated U is equal to unity, as required by FCM clustering. Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} and the cluster centers c_j by

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\|X_i - C_j\|}{\|X_i - C_k\|}\right)^{\frac{2}{m-1}}}, \qquad C_j = \frac{\sum_{i=1}^{N} u_{ij}^m \cdot x_i}{\sum_{i=1}^{N} u_{ij}^m}. \tag{4}$$

This iteration will stop when $\max_{ij}\{|u_{ij}^{(k+1)} - u_{ij}^{(k)}|\} < \varepsilon$, where ε is a termination criterion between 0 and 1, whereas k are the iteration steps. This procedure converges to a local minimum or a saddle point of J_m.

b) Blood cells elimination: Now, the final set of the candidate nuclei markers contains true findings corresponding to nuclei cells and false findings corresponding to blood cells, we need to remove the blood cell markers to be ready for the segmentation stage. For this purpose, we apply FCM clustering algorithm. Given the fact that the FCM algorithm is unsupervised, the clustering technique, i.e., does not require any training. Moreover, it is independently applied in each image. The detected areas are separated into two classes, namely, the true nuclei class and the rest of the findings. For the definition of the set of nuclei feature vectors, each pattern was centered at each centroid in the initial color image. Then, the feature vector is the intensity information of the neighborhood of the centroid with a 5 × 5 × 3 pattern size (the third dimension corresponds to the color). Representative results of Otsu's thresholding method and FCM clustering algorithm in the real image are shown in Fig. 5.

D. Cell Nuclei Segmentation

In order to separate attached cancer cells into individual objects, we further process the result from last step with marker-controlled watershed.

1) Marker-Controlled Watershed Transform: The watershed transform can be classified as a region-based segmentation approach. The intuitive idea underlying this method comes from geography. Since any gray-scale image can be considered as a topographic surface (we regard the intensity of a pixel as altitude of the point). Let us imagine the surface of this relief being immersed in still water, with holes created in local minima. Water fills up the dark areas "the basins," starting at

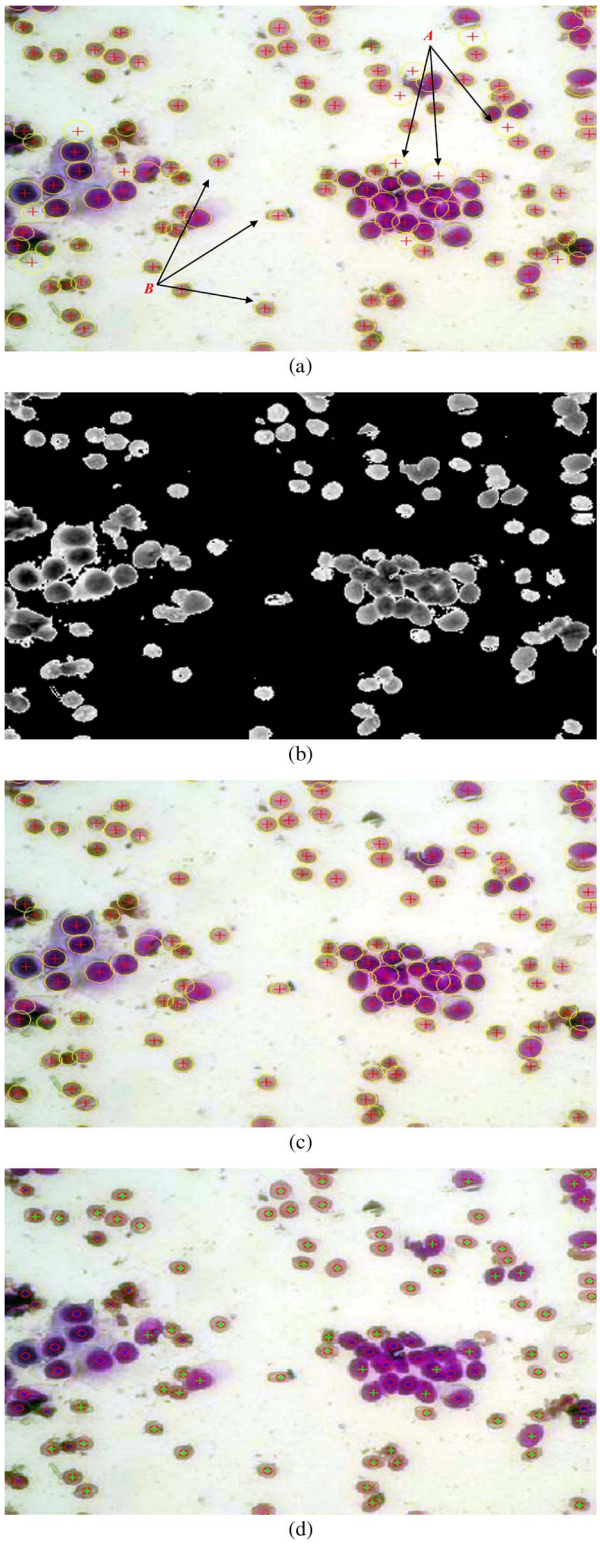


Fig. 5. Representative results for FP findings elimination. (a) Cell nuclei detection result using CHT for a cytological image. Notice that there are noisy circles such as region A and blood cells such as region B. (b) Otsu's thresholding on the Y level. (c) Cell nuclei positions after applying Otsu's thresholding method. Notice that the positions corresponding to noisy circles are removed. (d) FCM clustering for detected markers where green markers represent blood cells locations, and blue markers represent cancer cells locations.

these local minima. Where waters coming from different basins meet, we will build dams. When the water level has reached the highest peak in the landscape, the process is stopped. As a result, the landscape is partitioned into regions or basins separated by dams, which are called watershed lines or simply watersheds.

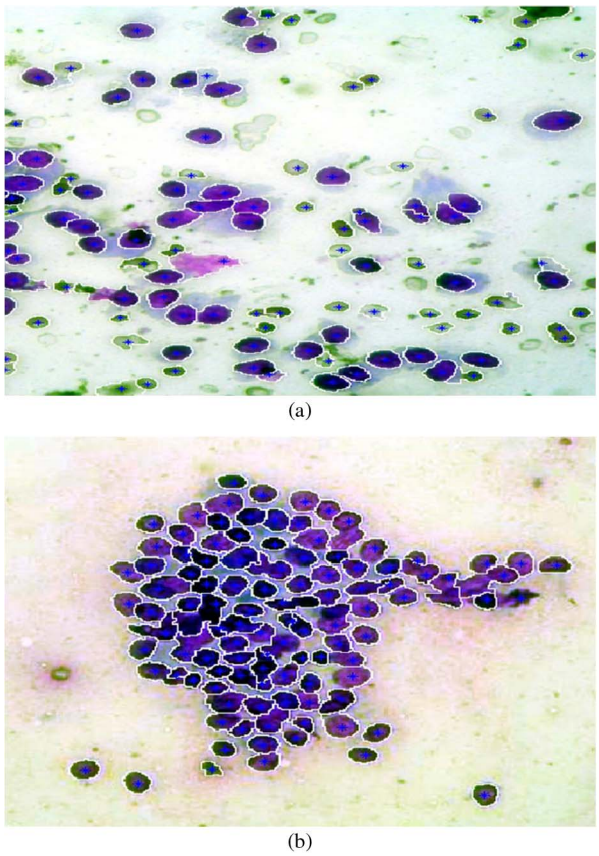


Fig. 6. Watershed algorithm results for cell nuclei segmentation where (a) benign FNAC image and (b) malignant FNAC image.

The application of the watershed transform in the default form usually results in oversegmentation of the image because of the presence of artifacts and noise. To avoid this undesirable effect, the watersheds are applied in edge images with markers [35] and [36]. They are used as starting points of the flooding process. In our approach, the nuclei markers are automatically determined as illustrated in the previous sections. The result of the marker-controlled watershed transform in an image with nuclei markers is depicted in Fig. 6.

IV. Classification Phase

A. Feature Extraction

The efficient classification of nuclei cells from the total segmented regions requires the generation of meaningful features of very good discriminative ability. Having found the areas of the nuclei enclosed by the detected boundaries, features concerning the shape and the texture of the detected regions can be easily determined. In this paper, we use ten shape-based features and two textural features [37]. The values obtained for these features yield a good differentiation between cancerous and healthy cells. These features are proposed as input data for the classification phase.

TABLE III
SHAPE FEATURES USED IN THIS PAPER

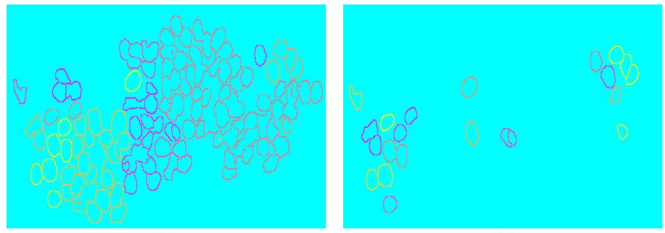
Feature	Definition
Perimeter	This is the total number of the nucleus boundary points.
Area	This is measured simply by counting the number of points inside the nucleus region.
Compactness	Perimeter and area are combined to give a measure of the cell nucleus compactness using the formula $C = 4\pi\ area/perimeter^2$. Compactness measures the efficiency with which a boundary encloses an area. For a circular region we have that C ≈ 1. This represents the maximum compactness value.
Smoothness	The smoothness of a nuclear contour is quantified by measuring the difference between the length of a radial line and the mean length of the lines surrounding it.
Eccentricity	Eccentricity specifies the eccentricity of the ellipse that has the same second-moments as the region. It allows us to track how much a segmented nucleus differs from a healthy nucleus. Healthy nuclei will assume circular shapes while cancerous nuclei can assume arbitrary shapes. We calculate eccentricity as the ratio of the distance between the foci of an ellipse and its major axis length. The values of this feature vary between 0 and 1. These are degenerate cases because an ellipse whose eccentricity is 0 is actually a circle, while a shape whose eccentricity is 1 is a line segment.
Solidity	Solidity specifies the proportion of the pixels in the convex hull that are also in the region. Computed as $Area/Convexarea$.
Equivalent Diameter	This specifies the diameter of a circle with the same area as the region. Computed as $\sqrt{4\ area\ /\ \pi}$.
Extent	Extent specifies the ratio of pixels in the region to pixels in the total bounding box. Computed as the Area divided by the area of the bounding box.
Major Axis Length	This specifies the length (in pixels) of the major axis of the ellipse that has the same normalized second central moments as the segmented nucleus.
Minor Axis Length	This specifies the length (in pixels) of the minor axis of the ellipse that has the same normalized second central moments as the segmented nucleus.

Shape features. The detected boundaries for the nuclei are expected to present an ellipse-like shape and several features to describe this characteristic are chosen. Ten features are calculated from the extracted shape of the detected region boundary, namely, perimeter, compactness, smoothness, eccentricity, solidity, equivalent diameter, extent, major axis length, and minor axis length. Table III shows detailed explanation for these features.

Textural features. Two features are calculated from the texture of the cell nucleus that is the standard deviation for the intensities of the region in both the gray scale of the RGB color model and the gray scale of YCbCr color model, which is Y level.

B. Classification

Classification is a task of assigning an item to a certain category, which is called a class, based on the characteristic features of that item. This task in any classification system is performed by a classifier that takes a feature vector as an input and responds with a category to which the object belongs. A feature vector is a set of features extracted from the input data. In this paper, the feature vector represents the 12 features extracted for each nucleus as illustrated above in the feature extraction phase. The classification step was realized using four well known supervised classification algorithms, namely, SVM [38], learning vector quantization (LVQ) [39], PNN, and multilayer perceptron (MLP) using back-propagation algorithm [28]. We have chosen these classifiers as they have achieved higher accuracy with many researchers for breast FNAC classification [40].



(a) (b)

Fig. 7. (a) Cell nuclei before removing large connected components. (b) Cell nuclei after removing large connected components.

C. Training and Testing

Training and testing data sets are constructed using tenfold cross validation [41]. The performance of the classifier is calculated using the unknown nuclei features of the testing data set. Some researches provide one feature vector for each FNAC image by extracting N features for each nucleus and then calculate some measures for each feature for all nuclei in this image [21], [24], and [27]. Wisconsin diagnostic breast cancer is the most commonly used data set by many researchers [22]. Wisconsin data set computes three measures, which are the mean, standard error $Se = Std\_dev/\sqrt{length}$, and "worst" or largest (mean of the three largest values). In this case, the number of patterns will be equal to the number of data set images. Other researchers generate one feature vector for each nucleus [9] and [23]. This is performed when the number of data set images is fairly small. The generated data set features will contain a number of patterns that are equal to the number of nuclei in all data set images.

In this paper, we have constructed four different input data matrices from our data set images. As mentioned before, our data set images contain 92 FNAC image. The first input data matrix (data set 1) contains one feature vector for each FNAC image with a dimension of 92 × 60. Each FNAC slide is represented by 60 features, 12 features are extracted for each nucleus, and then, five measures are computed for each feature. The five calculated measures in this paper are min, max, mean, standard error, and "worst" (mean of the three largest values). The second input data matrix (data set 2) is the same as data set 1, but we apply filtering for removing large connected components in the segmented image before feature extraction. We have found that many contiguous cells are extracted as one connected component, as illustrated in Fig. 7(a). The values of the features will be largely dependent on the size of the connected component. Thus, we have filtered the segmented image by removing the connected component with an area greater than some threshold. Fig. 7(b) shows the segmented image after the removal of large components.

The third input data matrix (data set 3) contains one feature vector for each nucleus with a dimension of 3260 × 12. Each nucleus is represented by 12 features. The dimension of 3260 is the number of connected components in all images in our data set. Finally, the fourth input data matrix (data set 4) is the

TABLE IV
DETAILS OF THE DATA SETS USED IN TRAINING AND TESTING

	No. of patterns	Features	Description
Dataset 1	92 Image 45 Benign 47 Malignant	60 features: five measures (min , max, mean, standard error and worst) and twelve features (Perimeter, Area, Compactness, Smoothness, Eccentricity, Solidity, Equivalent Diameter, Extent, Major Axis Length, Minor Axis Length, Standard deviation for grayscale intensities and Standard deviation for Y-level intensities).	Nuclei statistical features Before large components removing Feature vector for each image
Dataset 2	92 Image 45 Benign 47 Malignant	Same as dataset 1.	Nuclei statistical features After large components removing Feature vector for each nucleus
Dataset 3	3260 Image 1135 Benign 2125 Malignant	Twelve features (Perimeter, Area, Compactness, Smoothness, Eccentricity, Solidity, Equivalent Diameter, Extent, Major Axis Length, Minor Axis Length, Standard deviation for grayscale intensities and Standard deviation for Y-level intensities).	Nuclei statistical features Before large components removing Feature vector for each image
Dataset 4	2561 Image 933 Benign 1628 Malignant	Same as dataset 3.	Nuclei statistical features After large components removing Feature vector for each nucleus
Dataset 5	569 Image 357 Benign 212 Malignant	30 features: three measures (mean, standard error and worst) and ten features (radius, texture, standard deviation of gray-scale values, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimension).	Nuclei statistical features Feature vector for each image
Dataset 6	699 Image 458 Benign 241 Malignant	9 features: Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli	Breast tissue features Feature vector for each image

same as data set 3 but after removing the large components. The dimension of data set 4 is 2561×12. This means that we have removed 699 connected components.

We have trained the four constructed data sets. In addition, we have used two benchmark data sets with different features to compare the classification results with our data set results. Breast cancer Wisconsin (diagnostic) data set [42] (data set 5) with a dimension of 569×30 where 569 represents the number of data set images, and 30 is the feature vector length. The last trained data set is the breast cancer Wisconsin (Original) data set [43] (data set 6). Table IV shows detailed a description about all six data sets. We have trained each data set using the four illustrated classifiers.

The parameters needed for training and testing the data sets based on the four classifiers are shown in Table V. We have gotten the values of these parameters after several experiments.

TABLE V
TRAINING PARAMETERS USED IN SVM, LVQ, PNN, AND MLP

SVM	Number of iterations =1000
	Kernel function : Linear
	C parameter = 1
MLP	Number of hidden layers = 1
	Hidden layer neurons = 24
	Hidden activation function = sigmoid function
	Output layer neurons = 2
	Output activation function = linear function
	Max number of iterations = 200
	Goal (error limit) = 1e-05
	Learning algorithm : Levenberg-Marquardt back-propagation
LVQ	Number of hidden neurons =30
	Learning rate =0.01
	Max number of iterations = 100
	Goal (error limit) = 1e-04
	Learning function = 'learnlv1'
PNN	Goal (error limit) = 1e-05

TABLE VI
TRAINING TIMES FOR THE CLASSIFIERS USED IN THIS PAPER

Classifiers	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5	Dataset 6
	Time in sec (mean ± std)					
SVM	0.44 ± 0.98	0.09 ± 0.01	41.4 ± 6.61	33.54 ± 10.18	47.19 ± 2.44	87.77 ± 2.49
LVQ	71.84 ± 3.97	69.6 ± 0.66	2145.22 ± 21.16	1681.82 ± 25.64	387.95 ± 7.86	463.45 ± 4.37
PNN	0.43 ± 0.8	0.18 ± 0.0	0.19 ± 0.0	0.18 ± 0.0	0.18 ± 0.0	0.18 ± 0.0
MLP	133.65 ± 52.42	30.17 ± 12.97	115.8 ± 6.89	95.52 ± 2.06	144.39 ± 53.43	19.56 ± 6.8

The training and testing sets were chosen using tenfold cross validation. We took into consideration four neural networks classifiers, namely, SVMs, MLP, LVQ, and PNN. We applied these classifiers on the six data sets that contain our study data sets and benchmark data sets as illustrated above. The training times for the classifiers presented in this paper are provided in Table VI. As shown in the table, the PNN classifier takes the lowest training time. This is followed by SVM, LVQ, and MLP. For the training results, Fig. 8 shows the MLP training performance for the six data sets. The training performance shows that the first data set has achieved the training goal (error) in 87 epochs, and the second data set has reached the training goal in 26 epochs, whereas the other data sets have finished the training after 200 epochs without reaching the training goal. This means that the number of training samples greatly affects the performance convergence.

V. REMOTE SYSTEM

Telemedicine is the use of medical information exchanged from one site to another via electronic communications for the health and education of the patient or healthcare provider and for the purpose of improving patient care. Telemedicine includes consultative, diagnostic, and treatment services. One of the telemedicine applications is the remote diagnosis systems. We have provided remote diagnostic system for the detection and classification of breast cancer based on FNAC radiological images. Fig. 9 shows the components of the proposed remote diagnosis system.

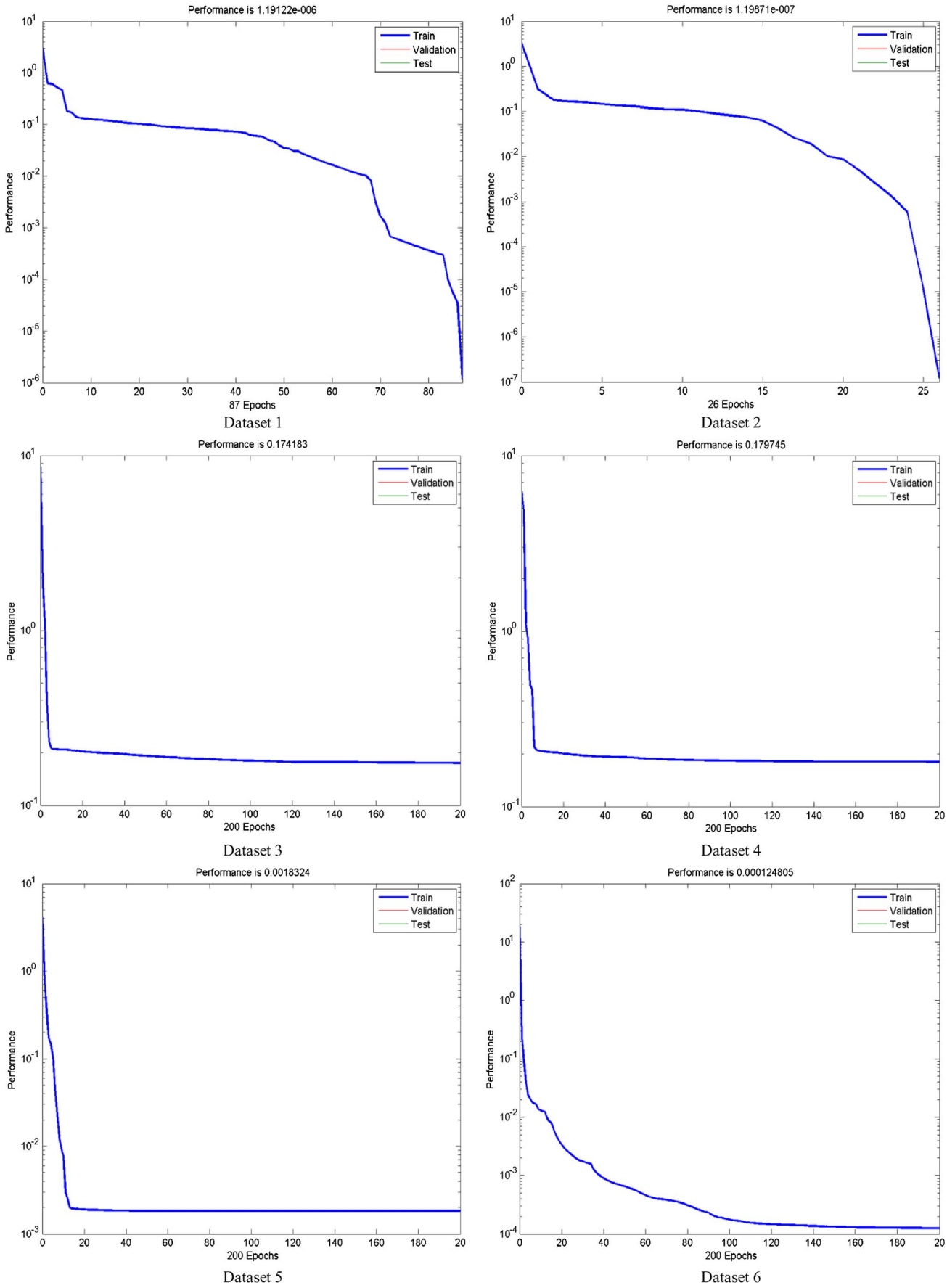


Fig. 8. MLP training performance for six data sets.

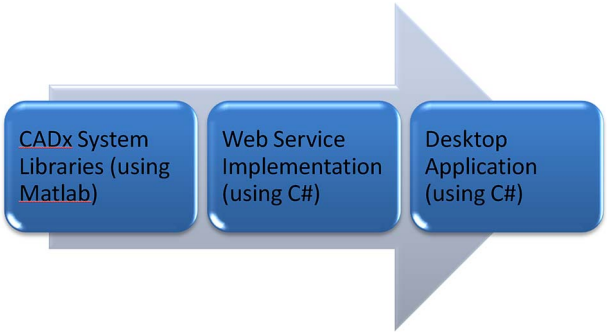


Fig. 9. Remote diagnostic system components.

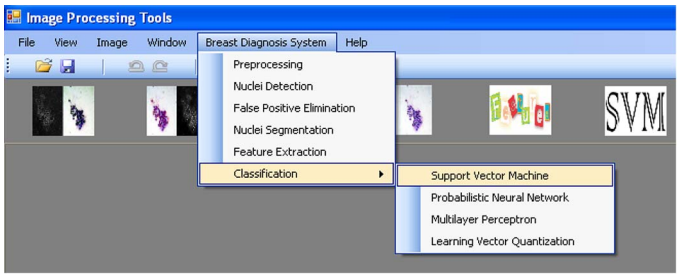


Fig. 10. Main remote diagnosis system functionalities.

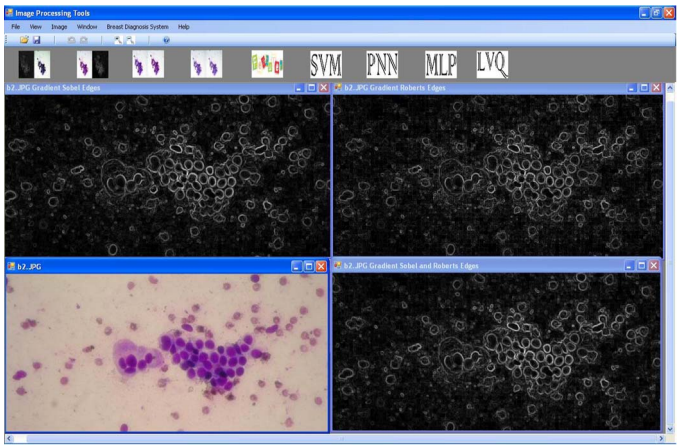


Fig. 11. Preprocessing stage results for the proposed remote system. Gradient images.

First, all phases of the proposed CADx system were implemented using Matlab softwarem, and the library file for the system was generated using a deployment tool. Next, a Web service [44] that used this library file was performed to be our remote server. Finally, a desktop application based on C# windows application was implemented as an interface for a physician and a patient at any place and at any time.

Fig. 10 shows the main functions proposed in the developed remote system. As illustrated, we have six main functions to help the physician in the screening and diagnosis of any FNAC image. The first function is the image preprocessing. Through this function, the physician can view the edges of the opened FNAC slide, as shown in Fig. 11. In addition, the physician can view the adjusted and equalized version of the FNAC slide, i.e., gray level and Y level, as shown in Fig. 12.

The second function is the nuclei detection, which results in the accumulator image and the detected nuclei markers for the

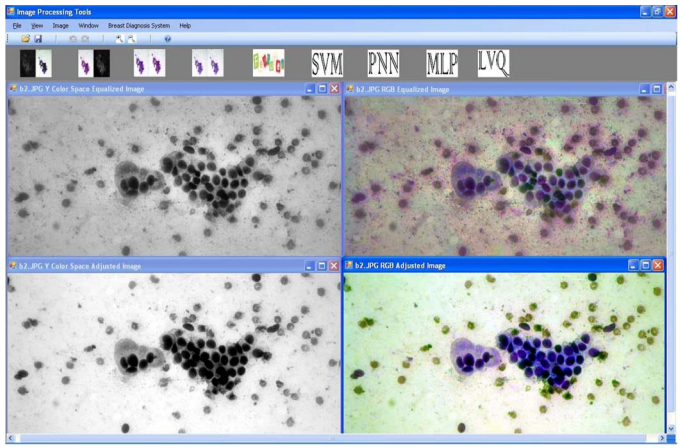


Fig. 12. Preprocessing stage results for the proposed remote system. Enhanced images.

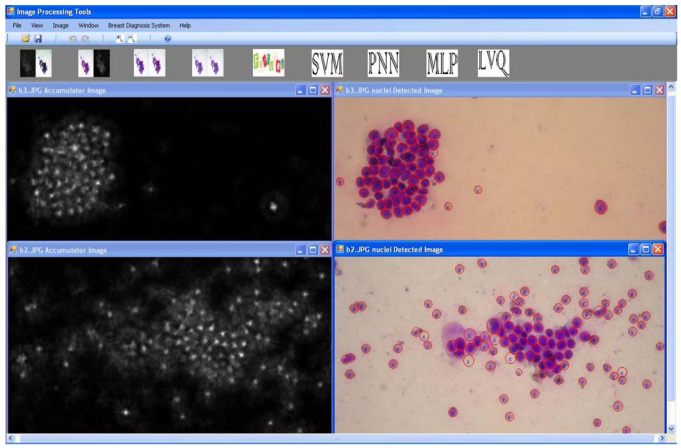


Fig. 13. Nuclei detection sample results for the proposed remote diagnosis system.

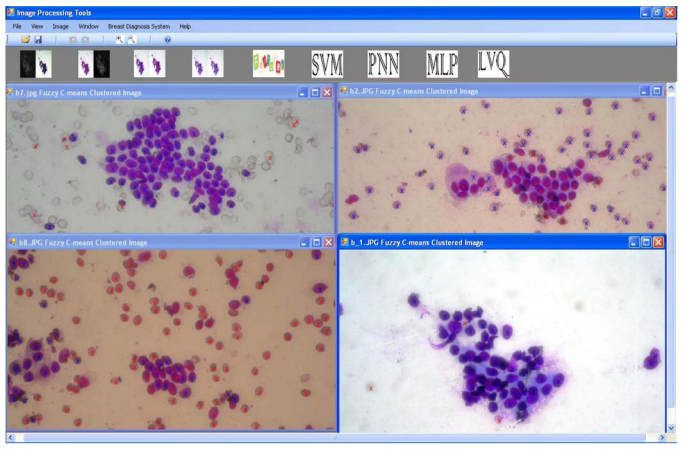


Fig. 14. FP elimination sample results for the proposed remote diagnosis system.

FNAC slide, as shown in Fig. 13. The third function is the FP elimination that outputs the clustered nuclei markers, as shown in Fig. 14. The fourth function is the nuclei segmentation, which results in the boundary of the nuclei in the FNAC image, as shown in Fig. 15.

The fifth function is the feature extraction for the segmented image. As illustrated before, we extract for each slide the four

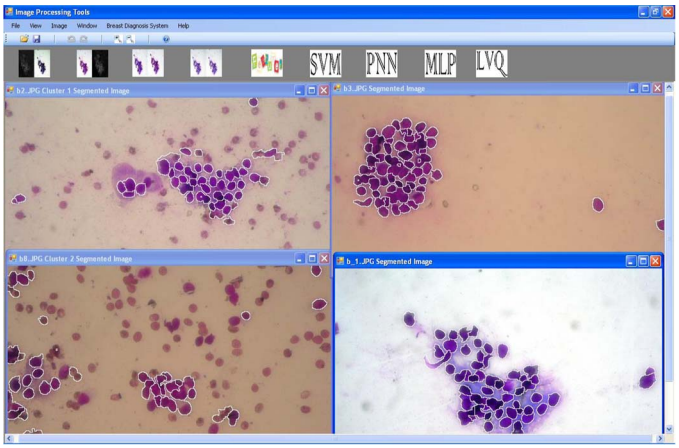


Fig. 15. Segmentation results sample for the proposed remote diagnosis system.

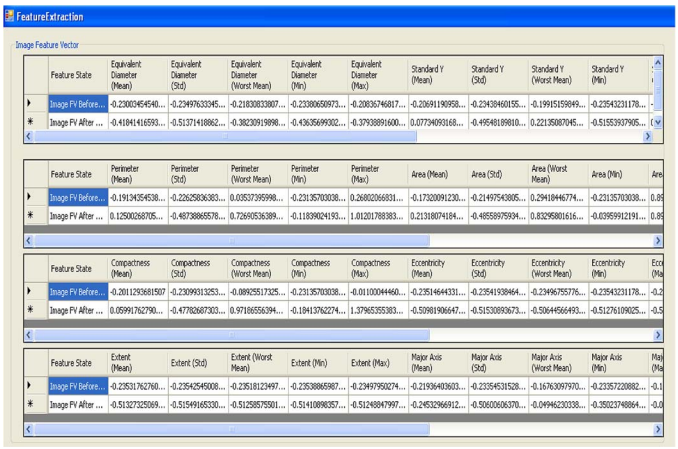


Fig. 16. Feature extraction sample results for the proposed remote diagnosis system.

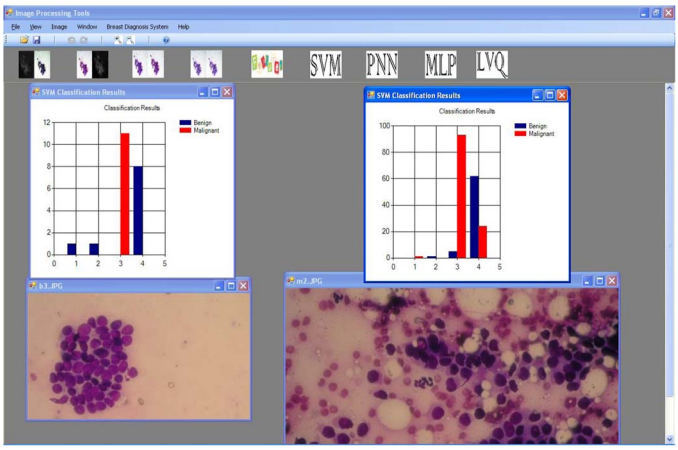


Fig. 17. Classification sample results for the proposed remote diagnosis system.

feature vectors, as shown in Fig. 16. We classify the tumor based on the four feature vectors extracted from the FNAC slide.

The last function is the classification. Through this function, the physician can see chart for the diagnosis result (i.e., benign or malignant), as shown in Fig. 17. Both charts in this figure show the diagnosis result based on SVM. The first chart shows the classification results for a benign slide, and the second chart represents the classification results for a malignant slide. In the first chart, the first column is related to the classification results based on the image feature vector before the large connected components elimination. The second column is same as the first one but after the elimination of large connected components. The third column represents the classification results based on the nucleus feature vector before the large connected components elimination. The fourth column is the same as the third one but after the removal of large connected components.

TABLE VII
EXECUTION TIME OF THE INDIVIDUAL STAGES OF THE PROPOSED SEGMENTATION METHOD FOR IMAGES OF SIZE 640 × 480

Stages of the proposed method	Time in sec (mean ± std)
Pre-processing step	1.0861 ± 0.0160
Cell nuclei detection step	8.2707 ± 0.8749
False positive elimination step	2.9656 ± 0.9676
Cell nuclei segmentation step	0.8923 ± 0.0480

TABLE VIII
VALUES OF THE PARAMETERS INCLUDED IN THE PROPOSED SEGMENTATION METHOD

Stage of the method	Parameter	Value
Pre-processing stage – CLAHE	Clip_Limit	0.02
Pre-processing stage – CLAHE	Num_Tiles	[8 8]
Detection stage – CHT	Grd_Thresh	10
Detection stage – CHT	Rad_Range	[4 21]
Detection stage – Local Maximum Detection	Filter_Size	[17 17]
False positive findings elimination stage – FCM	Pattern_Size	5 x 5 x 3
False positive findings elimination stage – FCM	Stop_Criteria (ε)	1e-5
False positive findings elimination stage – FCM	Max_No_iteration	100

VI. RESULTS AND DISCUSSION

A. Detection and Segmentation Phase

The proposed method for the detection and segmentation of breast cytological images is fully automated, and its application was performed without any end-user interference. This phase was developed in Matlab version 10b using a dual-core PC with a 2.0-GHz processor and 2 GB of RAM. The processing times of the individual steps of the method are provided in Table VII. As shown in the table, nuclei detection step takes much time than the other stages of the proposed method. This stage takes a long time because of the creation of the accumulator in CHT algorithm. It must be noted that the Hough transform for ellipse detection takes much time than circle detection because the required parameters needed for accumulator creation in circle detection is less than those needed for ellipse detection. That is the reason why we have performed circle detection instead of ellipse detection. In addition, the table shows that the execution time of the overall proposed method is very good despite of including many stages and approaches.

The parameters of the several steps of the proposed method are computed after several experiments on randomly selected images from our data set. These parameters are shown in Table VIII. In this table, Grd_Thres parameter was used to

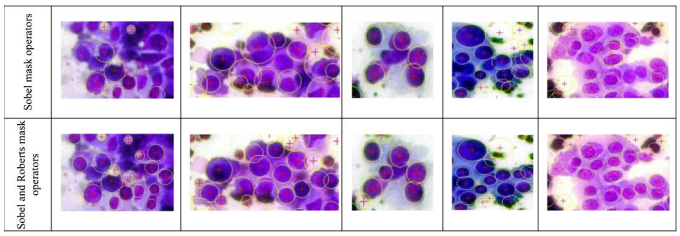


Fig. 18. Sample CHT results based on the gradients from Sobel operators and both Roberts and Sobel operators.

TABLE IX
NUCLEI DETECTION RESULTS

	Tumor nuclei	Blood cell nuclei and noisy markers (false positive findings)	Total
Our dataset	5108	6394 blood cell nuclei marker	11,502
Detected markers	4881	7624 blood cell and noisy markers	12,505

ignore weak edges with an intensity value less than or equal to 10. In addition, CHT will search for circles with radices from 4 to 21. Finally, the feature vector size for each marker point will be of length 75. As a window of size 5×5 is centered on this marker point, and then R, G, and B for each pixel in this window will be added in the feature vector related to this marker point. For the evaluation of the method, we have examined the performance of the different steps of the proposed method, as illustrated in the following sections.

1) Preprocessing Stage: The preprocessing step results an enhanced image, which also produces the best gray level and prepares the gradient image. In this step, the CLAHE is performed on [8, 8] tiles, and the clip limit is set to 0.02. Experimental results show that the best gray scale level is the Y level of YCbCr color space image. We have found that using the green level in the nuclei detection stage has given better results than the other gray levels in 22 specimens, whereas the gray level has produced best results for 32 specimens, and the Y level has achieved the best results for 38 specimens. In addition, our experiments show that the gradient image estimated from a saturated sum of both Sobel and Roberts mask operators is better than the gradient produced from each one individually. As in the first gradient, more edges are determined that result in more detected circles when applying CHT, as shown in Fig. 18.

2) Cell Nuclei Detection Stage: The detection step successfully identifies most of the nuclei markers in the image. We have used CHT and local maximum filtering for the detection of these markers. Several tests have been performed to find the best values for circles' radii range and the size of Mexican filter. The range of circles' radii used in CHT was set to (4–21). When the range of circles' radii was smaller than this range, then some tumor nuclei were missed. However, when the range of circles' radii was greater than the above range, then more than one marker is detected for each cell nucleus. Hence, this is the best range to avoid both problems. In addition, the size of Mexican hat filter used in the local maximum filtering was 17×17.

The detection step results are shown in Table IX. As shown in the table, the number of the overall detected markers is 12 505 markers, including 4881 tumor nuclei markers and 7624 blood cells and noisy markers. As it can be observed, there are a large number of FP detected markers, which correspond to blood cells and noisy markers. Hence, for effective segmentation results, it is essential to eliminate these false markers. In addition, in the detection step, 227 from 5108 true nuclei are missed, and the true nuclei detection rate was 97.08%. Table X shows detailed results for tumor cell nuclei detection performance.

TABLE X
DETAILED ACCURACY FOR TUMOR CELL NUCLEI DETECTION

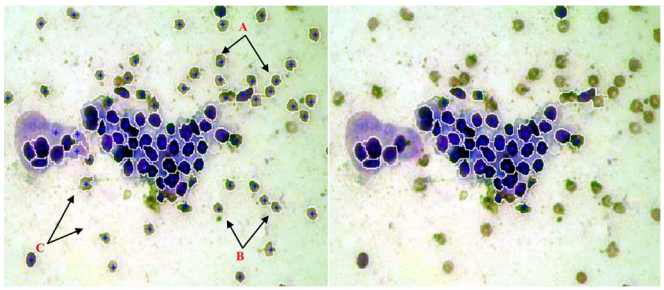
	True Detected Nuclei	Missed Nuclei	Total Cells	Detection Rate
Benign Cells	2364	110	2474	95.55 %
Malignant Cells	2517	117	2634	95.56%
Tumor Cells	4881	227	5108	95.56%

TABLE XI
FCM CLUSTERING SPECIFICITY, SENSITIVITY, AND FP RATE

Total Detected Nuclei Centers	4881
True Positive (TP)	4661
True Negative (TN)	6174
False Positive (FP)	1450
False Negative (FN)	220
Sensitivity (True Positive Rate)	95.49 %
Specificity (True Negative Rate)	83.16 %
False Positive Rate (100 – Specificity)	16.84

3) FP Findings Elimination Stage: The FP findings elimination step includes Otsu's thresholding and FCM clustering technique. The Otsu's thresholding method yields in the reduction of FP findings at the rate of 23.03% when applied on Y level, whereas the reduction rate was 23.35% when applied on the green level. This means that the green level yields more noisy circles than the Y level and this ensures that Y level is better than the green level in the proposed method. On the other hand, the noisy circles elimination has been achieved, whereas we have no loss of true nuclei markers. In addition, it should be determined that if we omit Otsu's thresholding method, there will be some markers, which belong to the noisy circles, and they will introduce interference in the clustering step as there will be three clusters in the images that contain noisy markers, whereas two clusters in the images without noisy markers. It must be noted that the images in our data set contain severe noise and large number of blood cells; thus, it is essential to eliminate the markers related to them.

For the FCM clustering performance, we have manually calculated the number of true-positive (TP), true-negative (TN), FP, and false-negative (FN) findings in all images in our data set. Two widely used statistical measures for the performance of the classification are calculated, namely, sensitivity and specificity. The sensitivity is defined as sensitivity $= \mathrm{TP}/\mathrm{TP} + \mathrm{FN}$. It measures the proportion of actual nuclei that is correctly identified. The specificity is defined as specificity $= \mathrm{TN}/\mathrm{TN} + \mathrm{FP}$. It measures the proportion of candidate markers that is not nuclei and is correctly characterized. In the clustering step, 4661 from 4881 detected nuclei are TP, whereas 1450 detected nuclei locations are FP. In addition, 6174 detected locations are TN, which means that our data set contains a lot of noisy markers and blood cells, whereas 220 nuclei locations are FN. Table XI shows detailed results for FCM clustering algorithm performance.



(a) (b)

Fig. 19. (a) Watershed transform for markers extracted from local maximum filter (before clustering step). Notice that there are many FP findings such as regions A, B, and C due to blood cells that do not correspond to the true nuclei. (b) The final segmentation results after the FCM clustering step.

As it can be observed, we consider the FP findings problem effectively. However, we do not care about FN findings. Hence, is it mandatory to eliminate FN or not? If we get a deep look at the table of results, we can observe that many reasons show that handling the problem of FP findings is more essential than handling the one related to FN findings in our data set. From one hand, the number of FN findings is 220. This means that each image in our data set contains in average 2 FN nuclei and 51 TP nuclei locations. If the classification step does not take into consideration the features of FN findings, then the features of TP findings are enough for good diagnosis results as the number of FN nuclei are very small with respect to the overall TP nuclei. On the other hand, the number of FP findings is 1450. This means that each image in our data set includes in average 15 FP nuclei. If the classification step takes into consideration the features of 15 nuclei, which are not tumor nuclei, aside the 51 TP nuclei, then the diagnosis will be greatly affected as the number of FP findings is larger with respect to the overall TP locations. Hence, this study does not handle the problem of FN findings elimination. This problem should be solved to generalize the application of the proposed method for all breast FNAC data sets. Solving this problem may be included in our future work when using more data sets.

4) Cell Nuclei Segmentation Stage: Finally, the nuclei segmentation step produces the boundary of cell nuclei based on the application of the marker-based watershed transform. Fig. 19 shows the results of segmentation before and after FP elimination step. This figure shows the effectiveness of the marker-based watershed transform for nucleus boundary extraction. In addition, the figure shows that the elimination step is very essential for effective segmentation.

The proposed method has very good results despite containing a high degree of noise and a huge number of blood cells in our data set. To ensure the effective results of the proposed method, detailed comparison between the proposed method and the other methods maintained in the state of the art are shown in Table XII. We have compared our method with the segmentation method proposed in the literature by Hrebien *et al.* [12]. Hrebien uses CHT as a presegmentation, $(1+1)$ evolutionary search strategy for cell nuclei localization, and finally, the watershed algorithm, active contouring, and a cellular automata GrowCut method for cell nuclei segmentation. In this paper, we localize the nuclei by local maximum filtering, which is very fast and has better results comparing to the $(1+1)$ evolutionary search strategy. In addition, Hrebien's method does not handle the elimination of blood cells and noisy markers detected by the Hough transform. In this paper, we handle this elimination by using Ostu's thresholding method and the unsupervised FCM clustering. Finally, our method takes very small time processing comparing with Hrebien's method.

We have also compared our method with the method proposed by Plissiti *et al.* [17]. Marina's method is based on the morphological reconstruction for cell nuclei detection, FCM for false finding elimination, and finally, the application of watershed transform for cell nuclei boundary extraction. As shown in the table, our sensitivity and specificity is larger than Marina's method. This means that the Hough transform for cell nuclei detection is much better than the morphological reconstruction. It must be noted that Marina's method does not handle the elimination of blood cells markers. Her method deals with the removal of cytoplasm markers only.

B. Feature Extraction and Classification Phase

Here, we will demonstrate the performance of the four classifiers presented in Section II along with results obtained for applying these classifiers on the six illustrated data sets for comparison. The method was developed in Matlab version 10b using a dual core PC with a 2.0-GHz processor and 2 GB of RAM.

Two widely used statistical measures for the performance of the classification are calculated, namely, sensitivity and specificity. The sensitivity measures the proportion of malignant tumors that are correctly identified, whereas the specificity measures the proportion of benign tumors that are correctly characterized. We have used these measures for the performance analysis and for providing tools to select a possibly optimal classification model. Table XIII shows the sensitivity measure of the four tested classifiers for the six data sets, and Table XIV shows the specificity measure.

First, both tables show that the sensitivity and specificity measures for data set 1 are better than data set 2, and in addition, both measures for data set 3 are better than data set 4 for all classifiers. As mentioned before, data set 1 and data set 3 contain features for all connected components, whereas data set 2 and data set 4 contain features for connected components after removing the large ones. This means that removing large connected components from the segmented nuclei does not improve any classifier performance. As all classifiers gives higher sensitivity and specificity in data sets with features without removing large connected components (data set 1 and data set 3).

Next, both tables show that the sensitivity and specificity measures for data set 1 and data set 2 are lower than data set 3 and data set 4 for all classifiers. As mentioned before, data set 1 and data set 2 include one feature vector for each FNAC image, whereas data set 3 and data set 4 include one feature vector for each nucleus. This means that extracting feature vector for each nucleus is better than extracting feature vector for each FNAC image.

TABLE XII
COMPARISON BETWEEN THE PROPOSED METHOD AND THE OTHER METHODS APPEARED IN THE LITERATURE

Method [ref] Year	No. of images	Image size	No. of Tumor Cells	Performance Criteria	Quantitative Results	Time in sec (mean)
Lin Yang et al. [15] 2005	31	180 x 180	933	Estimated accuracy	85 %	-
Xiaodong Yang et al. [18] 2006	-	-	1178	Correctly segmented Over segmented Under segmented	98.8 % , 0.5 %, 0.7 % Respectively	-
Maciej Hrebien et al. [12] 2008	-	704 x 576	-	Visual Inspections	Watershed Accuracy 68.74 %	4-5 min
Jierong Cheng et al. [10] 2009	Dataset1 (53)	1392 x 1040	383	Correctly segmented Over segmented Under segmented	97.39 % , 1.83 %, 0.78 % Respectively	-
	Dataset2 (4)	450 x 450	432		96.30 %, 3.24 %, 0.46 % Respectively	
Marina E. Plissiti et al. [11] 2011	90	1536 x 2048	10,248	Hausdroff distance for the ground truth	Watershed (1.71 ± 0.54) ACM (2.48 ± 2.30) GVF (2.65 ± 3.23)	2-5 min
Marina E. Plissiti et al. [17] 2011	38	1536 x 2048	5,617	Sensitivity (Se) Specificity (Sp)	90.57 % (Se) 75.28 % (Sp)	83.04
This work 2012	92	640 x 480	11,502	Sensitivity (Se) Specificity (Sp)	95.49 % (Se) 83.16 % (Sp)	13.22

TABLE XIII
SENSITIVITY OF THE TESTED CLASSIFIERS FOR THE SIX DIFFERENT DATA SETS

Classifier	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5	Dataset 6	Average sensitivity
SVM	78.82	59.44	99.6	98.12	98.77	96.1	88.48
LVQ	83.23	58.6	98.86	98.01	92.34	96.65	87.95
PNN	97.02	94.84	93.78	92.48	100	99.8	96.32
MLP	74.85	57.7	81.3	75.41	97.93	96.74	80.66

TABLE XIV
SPECIFICITY OF THE TESTED CLASSIFIERS FOR THE SIX DIFFERENT DATA SETS

Classifier	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5	Dataset 6	Average specificity
SVM	92.23	64.09	99.64	98.65	94.63	96.68	90.99
LVQ	79.44	60.99	99.03	97.76	84.23	92.69	85.69
PNN	97.28	95.17	93.67	92.22	89.87	99.23	94.57
MLP	68.86	57.23	82.46	77	94.94	77.19	76.28

Next, the average measure for each classifier across all six data sets shows that the best classifier is PNN. This is followed in order by SVM, LVQ, and MLP. As shown in the tables, the PNN classifier gives the best measures for four data sets (data set 1, data set 2, data set 5, and data set 6); whereas SVM gives the best measures for two data sets (data set 3 and data set 4). It must be noted that PNN and SVM take very small training times when comparing with LVQ and MLP training times, as shown in Table VI.

Finally, we have trained the benchmark Wisconsin data sets (data set 5 and data set 6) for comparison with our study data sets. The results show that our study data sets have closed results comparing with the Wisconsin data set results; this means that our data set features are accurate and can be compared with benchmark data set features.

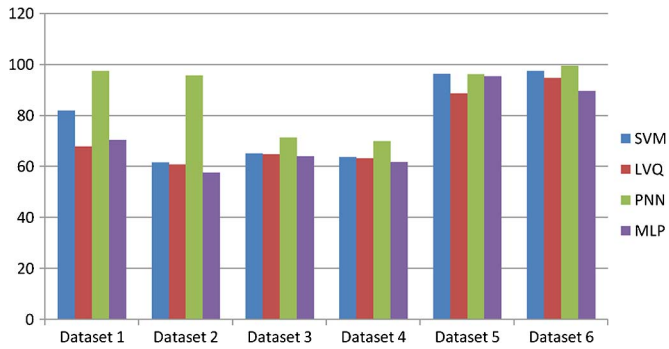


Fig. 20. Correct rates of the tested classifiers for each training data set.

Fig. 20 includes a chart for the correct rates of the tested classifiers for each training data set. All the achieved results show that the predictive ability of both PNN and SVM are stronger than the LVQ and MLP for the evaluated data sets.

VII. CONCLUSION AND FUTURE WORK

FNAC is an essential component in the preoperative management of breast lesions. Its accuracy, ease of use, and affordability are factors that cause its popularity. The advent of imaging technology together with the clinical expertise of the clinician contributed to its increased sensitivity.

In this paper, we have developed a fully automated method for the segmentation of cell nuclei in breast FNAC images. Through the work, we have effectively overcome the problem related to the detection of nuclei locations. Moreover, we have successfully eliminated the FP nuclei markers, which result in the efficient segmentation of nuclei boundaries.

The conducted experiments show that the Hough transform adopted for circle detection can be effectively used for the presegmentation of cell nuclei in FNAC images. In addition, it has been verified by the results that the best gray level for the processing of the FNAC image is the Y level of the YCbCr color space. Otsu's thresholding method eliminates all of the noisy circles, whereas we have no loss of true nuclei markers.

In addition, the FCM algorithm produces high accuracy for the clustering of the markers corresponding to true nuclei and blood cells. The proposed method shows that the outcome of the marker-controlled watershed transform is accurate nuclei boundaries. The main advantage of the proposed detection and segmentation method is that it is fully automated and it is suitable for images with a high degree of noise and blood cells and cell overlapping, as it can successfully detect not only the nuclei of isolated cells, but also the nuclei in cell clusters.

In addition, we have developed a computer-aided diagnosis system for breast FNAC cancer. For the feature extraction, we used ten shape-based features and two textural features. The values obtained for these features yield a good differentiation between benign and malignant cells. For the classification phase, we have performed four different classification models, namely, multilayer perceptron, PNN, LVQ, and SVM. The classification results were obtained using tenfold cross validation.

Six data sets were used to examine the efficiency of the proposed classifiers. Four data sets were constructed from our material, and the other two data sets are benchmark data sets and were evaluated for the comparison. The conducted experiments show that the classification results of the data sets contain one feature vector for each FNAC image (data sets 1 and 2) are not as good as the data sets contain one feature vector for each nucleus (data sets 3 and 4). In addition, the classification results of the data sets with features without removing large connected components (data sets 1 and 3) are better than those with features after removing large connected components (data sets 2 and 4). The classification performance was capable of producing up to 99.7 % sensitivity and specificity for our data sets. The results showed that the predictive ability of both PNN and SVM was stronger than the LVQ and multilayer perceptron for the evaluated data sets.

Finally, we have merged the proposed CADx system with a telemedicine platform. This is to provide an intelligent remote detection and diagnosis system for the breast cancer patients. For this purpose, we have implemented Web service, which includes all the phases of our CADx system. In addition, we have performed a breast diagnosis desktop application that invokes the Web service functions. This application can be used by physicians from their homes or any other place.

Although the results obtained so far are encouraging, more investigations are needed to further improve the performance of the false finding elimination with the implementation of other clustering algorithms. In addition, other segmentation techniques need to be performed to find the best segmentation technique for nuclei segmentation. In addition, the future work includes improving the clustering algorithms results (FCM, SVM, LVQ, MLP, and PNN) with the selection of different nuclei features and performing hybrid clustering algorithms. In addition, the implementation of other clustering algorithms needs to be investigated. Furthermore, the CADx system is being tested in a larger data set in order to evaluate the robustness of the proposed system. More work need to be done in the classification of cell nuclei malignancy to know the degree of malignancy of the FNAC image. Finally, the proposed remote CADx system can be used as the basis for further applications, such as mobile application for remote diagnosis for breast FNAC images.

ACKNOWLEDGMENT

The authors would like to thank Prof. Z. Shehab El-Din, Professor of Pathology, Early Cancer Detection Unit-Obstetrics and Gynecology Department, Faculty of Medicine; Ain Shams University for providing FNA breast cytological slides for the construction of the image data set.

REFERENCES

[1] CancerNet, A service of the National Cancer Institute. [Online]. Available: http://www.cancer.gov/cancertopics/types/breast

[2] T. Ishikawa, Y. Hamaguchi, M. Tanabe, N. Momiyama, T. Chishima, Y. Nakatani, A. Nozawa, T. Sasaki, H. Kitamura, and H. Shimada, "False-positive and false-negative cases of fine-needle aspiration cytology for palpable breast lesions," in *Breast Cancer*, Tokyo, Japan, 2007, vol. 14, pp. 388–392.

[3] P. Mendoza, M. Lacambra, P.-H. Tan, and G. M. Tse, "Fine needle aspiration cytology of the breast: The nonmalignant categories," *Pathol. Res. Int.*, vol. 2011, pp. 547580-18–547580-8, 2011.

[4] M. Auger and I. Huttner, "Fine-needle aspiration cytology of pleomorphic lobular carcinoma of the breast: Comparison with the classic type," *Cancer Cytopathol.*, vol. 81, no. 1, pp. 29–32, Feb. 1997.

[5] M. H. Bukhari and Z. M. Akhtar, "Comparison of accuracy of diagnostic modalities for evaluation of breast cancer with review of literature," *Diagn. Cytopathol.*, vol. 37, no. 6, pp. 416–424, Jun. 2009.

[6] T. Ayer, M. U. Ayvaci, Z. X. Liu, O. Alagoz, and E. S. Burnside, "Computer-aided diagnostic models in breast cancer screening," *Imag. Med.*, vol. 2, no. 3, pp. 313–323, Jun. 2010.

[7] L. Hadjiiski, B. Sahiner, and H.-P. Chan, "Advances in CAD for diagnosis of breast cancer," *Current Opinion in Obstetrics & Gynecology*, vol. 18, no. 1, pp. 64–70, Feb. 2006.

[8] C. W. Chen, J. Luo, and K. J. Parker, "Image segmentation via adaptive K-mean clustering and knowledge-based morphological operations with biomedical applications," *IEEE Trans. Image Process.*, vol. 7, no. 12, pp. 1673–1683, Dec. 1998.

[9] W. N. Street, "Xcyt: A system for remote cytological diagnosis and prognosis of breast cancer," in *Proc. Soft Comput. Tech. Breast Cancer Prog. Diag.*, 2000, pp. 297–322.

[10] J. Cheng and J. C. Rajapakse, "Segmentation of clustered nuclei with shape markers and marking function," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 3, pp. 741–748, Mar. 2009.

[11] M. E. Plissiti, H. Nikou, and A. Charchanti, "Combining shape, texture and intensity features for cell nuclei extraction in Pap smear images," *Pattern Recog. Lett.*, vol. 23, no. 6, pp. 838–853, Apr. 2011.

[12] M. Hrebien, P. Stec, T. Nieczkowski, and A. Obuchowicz, "Segmentation of breast cancer fine needle biopsy cytological images," *Int. J. Appl. Math. Comput. Sci.*, vol. 18, no. 2, pp. 159–170, Jun. 2008.

[13] A. E. Carpenter, T. R. Jones, M. R. Lamprecht, C. Clarke, I. H. Kang, O. Friman, A. Guertin, J. H. Chang, R. A. Lindquist, J. Moffat, P. Golland, and D. M. Sabatini, "Cell profiler: Image analysis software for identifying and quantifying cell phenotypes," *Genome Biol.*, vol. 7, no. 10, p. R100, 2006.

[14] M. N. Gurcan, T. Pan, H. Shimada, and J. Saltz, "Image analysis for neuroblastoma classification: Segmentation of cell nuclei," in *Proc. IEEE 28th Annu Conf. Eng. Med. Biol. Soc.*, 2006, vol. 1, pp. 4844–4847.

[15] L. Yang, P. Meer, and D. J. Foran, "Unsupervised segmentation based on robust estimation and color active contour models," *IEEE Trans. Inf. Technol. Biomed.*, vol. 9, no. 3, pp. 475–486, Sep. 2005.

[16] X. Zhou, F. Li, J. Yan, and S. T. C. Wong, "A novel cell segmentation method and cell phase identification using Markov model," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 2, pp. 152–157, Mar. 2009.

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

[18] X. Yang, H. Li, and X. Zhou, "Nuclei segmentation using marker controlled watershed, tracking using mean-shift, and Kalman filter in time-lapse microscopy," *IEEE Trans. Circuits Syst.*, vol. 53, no. 11, pp. 2405–2414, Nov. 2006.

[19] M. M. Beg and M. Jain, "An analysis of the methods employed for breast cancer diagnosis," *Int. J. Res. Comput. Sci.*, vol. 2, no. 3, pp. 25–29, Jun. 2012.
[20] N. Salleh, M. Arshad, and N. O. H. Sakim, "Evaluation of morphological features for breast cells classification using neural networks," in *Tools and Applications With Artificial Intelligence*. New York, NY, USA: Springer, 2009, pp. 1–9.
[21] N. A. M. Isa, E. Subramaniam, M. Y. Mashor, and N. H. Othman, "Fine needle aspiration cytology evaluation for classifying breast cancer using artificial neural network," *Amer. J. Appl. Sci.*, vol. 4, no. 12, pp. 999–1008, 2007.
[22] T. Kiyan and T. Yildirim, "Breast cancer diagnosis using statistical neural networks," *J. Elect. Electron. Eng.*, vol. 4, 2004.
[23] M. Kowal, P. Filipczuk, A. Obuchowicz, and J. Korbicz, "Computer-aided diagnosis of breast cancer using Gaussian mixture cytological image segmentation," *J. Med.l Inf. Technol.*, vol. 17, pp. 257–262, 2011.
[24] S. Singh, P. R. Gupta, and M. K. Sharma, "Breast cancer detection and classification of histopathological images," *Int. J. Eng. Sci. Technol.*, vol. 3, no. 5, pp. 4228–4232, May 2010.
[25] P. Filipczuk, M. Kowal, and A. Marciniak, "Feature selection for breast cancer malignancy classification problem," *J. Med. Inf.Technol.*, vol. 15, pp. 193–199, 2010.
[26] R. Nithya and B. Santhi, "Comparative study on feature extraction method for breast cancer classification," *J. Theor. Appl. Inf. Technol.*, vol. 33, no. 2, pp. 220–226, Nov. 2011.
[27] L. Jelen, T. Fevens, and A. Krzyzak, "Classification of breast cancer malignancy using cytological images of fine needle aspiration biopsies," *Int. J. Appl. Math. Comput. Sci.*, vol. 18, no. 1, pp. 75–83, Mar. 2008.
[28] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 2nd ed. Boston, MA, USA: Addison-Wesley, 2001.
[29] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics Gems IV*. San Diego, CA, USA: Academic, 1994, pp. 474–485.
[30] J. C. Russ, *The Image Processing Handbook*. Boca Raton, FL, USA: CRC Press, 1999.
[31] S. Beucher and C. Lantuejoul, Use of Watersheds in Contour Detection, Sep. 1979.
[32] M. Roushdy, "Detecting coins with different radii based on Hough transform in noisy and deformed image," *ICGST Int. J. Graphics, Vision Image Process.*, vol. 7, no. 1, pp. 25–29, Apr. 2007.
[33] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst. Man Cybern.*, vol. 9, no. 1, pp. 62–66, Jan. 1979.
[34] J. C. Bezdek, *Fuzzy Models for Pattern Recognition*, S. K. Pal, Ed. New York, NY, USA: IEEE Press, 1992.
[35] M. E. Plissiti, H. Nikou, and A. Charchanti, "Watershed-based segmentation of cell nuclei boundaries in Pap smear images," in *Proc. IEEE Int. Conf. ITAB*, 2010, pp. 1–4.
[36] M. Wang, X. Zhou, F. Li, J. Huckins, R. W. King, and S. T. C. Wong, "Novel cell segmentation and online SVM for cell cycle phase identification in automated microscopy," *Bioinformatics*, vol. 24, no. 1, pp. 94–101, 2008.
[37] W. H. Wolberg, W. N. Street, and O. L. Mangasarian, "Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates," *Comput. Appl. Early Detect. Staging Cancer*, vol. 77, no. 2/3, pp. 163–171, Mar. 1994.
[38] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.
[39] T. Kohonen, "The self-organizing map," *Proc. IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.
[40] *Datasets Used for Classification: Comparison of Results*, Department of Informatics, Nicolaus Copernicus University Computational Intelligence Laboratory. [Online]. Available: http://www.is.umk.pl/projects/datasets.html
[41] S. Geisser, *Predictive Inference*, 1st ed. New York, NY, USA: Chapman and Hall, 1993.
[42] W. H. Wolberg, Breast Cancer Wisconsin (Diagnostic) Data Set. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29
[43] W. H. Wolberg, Breast Cancer Wisconsin (Original) Dataset. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29
[44] A. Freeman and A. Jones, *Microsoft.NET XML Web Services step by Step*. Redmond, WA, USA: Microsoft Press, 2003.

Yasmeen Mourice George was born in Cairo, Egypt, in 1987. She received the B.Sc. and M.Sc. degrees in computer science from Faculty of Computer and Information Sciences, Ain Shams University, Cairo, in 2008 and 2013 respectively.

From August 2008 to May 2009, she worked as a Dot Net Developer (Windows application and PDA applications). From May 2009 to March 2013, she was a Demonstrator with the Computer Science Department, Faculty of Computer and Informatics, Benha University, Qalyubiyah, Egypt. She is currently a Teacher Assistant with the Computer Science Department, Faculty of Computer and Informatics, Benha University. Her current research interests include image processing, biomedical image processing, computer vision, machine learning, robotics, and artificial intelligence.

Hala Helmy Zayed received the B.Sc. degree in electrical engineering (with honors), the M.Sc. degree, and the Ph.D. degree from Zagazig University, Banha, Eqypt, in 1985, 1995, and 1989, respectively, all in electrical and communication engineering.

She is now a Professor of computer science with the Faculty of Computer and Informatics, Benha University, Banha. Her areas of research are pattern recognition, content based image retrieval, biometrics, and image processing.

Mohamed Ismail Roushdy received the Ph.D., M.Sc., and B.Sc. degrees from Faculty of Science, Ain Shams University, Cairo, Egypt, in 1993, 1984, and 1979, respectively.

From 1989 to 1991, his experimental doctoral research work was conducted at Bochum University, Bochum, Germany. Since 2007, he is currently a Professor of computer science. Since 2010, he has been a Dean with the Faculty of Computer and Information Sciences, Ain Shams University. His areas of research are artificial intelligence, image processing, data mining, and expert systems.

Bassant Mohamed Elbagoury received the Ph.D. and M.Sc. degrees in computer science from the Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt, in 2009 and 2005, respectively.

She was a Ph.D. Researcher with the NAO Robot team Humboldt, Germany. She finished her Ph.D. thesis in only two and half years. Since 2003, she participated in many international conferences in Paris, Poland, Germany, Jordan, and USA. Since 2005, she has been a Reviewer in AAAI USA conferences. She is currently an Assistant Professor of computer science with the Faculty of Computer and Information Sciences, Ain Shams University. Her areas of research are robotics, artificial intelligence, mobile computing, cloud computing, and image processing.