Localization-Based Methodology for a Fully Automatic Breast Cancer Image Diagnosis System Using Convolutional Neural network

Ali Fawzi Mohammed Ali[1†] and Mehdi G. duaimi[2††]
University of Baghdad, college of science, Dept. computer science, Baghdad, Iraq

Summary
Digital histopathology images represent the major evolutions in a modern medicine diagnosis. Pathological examinations consider the standard medical diagnosis protocols and play a critical and important role in the diagnosis process. Early tumor detection step in the diagnosis stage is obtained by the cytological testing of the breast image mainly based on the cell morphology and architecture distribution. In this paper, we present a localization methodology to predict and diagnosis the breast cancer type. The proposal relies on a fully automatic analysis of the fine needle biopsies for cytological images. A fully automatic isolation (cells detection) and filtration steps of the nuclei cells images are designed and implemented as a fully automatic computer-aid system to classify the nuclei cells type of the breast cancer by using such a powerful classification approach. Instead of relying on supervised classifiers such as an SVM and Neural Network, we proposed such robustness framework for the breast cancer cell image diagnosis system using a Convolutional Neural Network (CNN). Our approach in this paper relies on design and implementation such a fully supervised learning system to diagnosis the nuclei cells images by two main steps. The first one is the cells detection and filtration approach using Circular Hough Transform (CHT) and Support Vector Machine (SVM). In this step, we rely on fully automatic cells images selection that we have designed and implemented to train the SVM classifier that is used later for the cells filtration. This approach allows to automatically selection of a set of nuclei images to train the SVM classifier which is proposed for cell images filtration and isolation. Secondly, the new set of nuclei cell images is used by the CNN to predict and classify the breast cancer types. The fully automatic diagnostic system achieves about (99.41%) diagnosis accuracy results which illustrate a higher performance diagnosis system based on our implementation which is effective, valuable by providing an accurate diagnostic accuracy result as either benign or malignant.

Key words:
Breast cancer, histopathological images, SVM classification, Convolutional Neural Network (CNN).

1. Introduction

The breast cancer Computerized diagnosis system has represented a significant procedure in the early detection and diagnosis of cancer and increases the successful treatment cases since the last two decades, the intervention of computerized treatment strategies reduced patients' death ratio International Agency for Research on Cancer (IARC) defines the breast cancer as the most cancer disease among women especially those between 40 and 55 years of age. In the recent study that was in 2008, 1,384,155 diagnosed cases of breast cancer were discovered and about 458,503 cases deaths which are caused by the disease worldwide which is about 22.9% [1], [2]. Since the 1980s, the number of the cases deaths have been increased by about (3% to 4%) a year. However, the effeteness of the treatment disease is mainly relying on the earlier stage of cancer tumor detection [3] [4].

Fine Needle Biopsy (FBN) is defined as an examination technique to removes cells from a suspicious lump in the breast [5] [6]. In this approach, an automatic morphometric testing and diagnosis can improve the diagnosis which allows for screening and testing on a large scale of medical materials. In some cases, they are difficult and uncertain cases which would require furthering testing and examination [7], [8], [9].

In Portugal, 4 500 out of the 5 million female population are diagnosed with breast cancer every year, meaning that approximately 10% of Portuguese women will develop breast cancer at some stage of their lives [10] [11] [12]. Each day 11 new cases are detected and 4 women will die [13]. Fig. 1 illustrates the incidence rates of breast cancer around the world [14] [15]. Machine learning and datamining approaches can be used to design a computer Aid-system that is using to assist the physicians in a diagnosing stage of the breast cancer disease. Computer Aid-system provides a necessary treatment and prevents the impact which includes the possibility of death reasons [14].

Fig.1. Incidence rates of breast cancer worldwide (pink being the highest per capita rate) [14]

A. Background and motivation

Manuscript received July 5, 2017
Manuscript revised July 20, 2017

Breast cancer is an abnormal growth of breast cells that caused by from the inner lining of milk ducts or by the lobules which supply and support the ducts with milk [15]. Usually, breast cancer either begins in the cells of the lobules, so as it shown in Fig. 2, the ducts or the milk producing glands which drain the milk from the lobules to the nipple. In this case, at the beginning, the breast cancer can be constructed in the stromal tissues, which include the fibrous and fatty connective tissues of the breast [16].

B. Breast Cancer Diagnosis

Breast cancer detection by using the triple-test comprise self-examination (palpation), mammography or ultrasonography imaging, and fine needle (FNB). Mammograms screening which is a specialist technique that can be used to check and diagnosis for breast cancer in women who have no symptoms or signs of the disease, however, a mammogram can localize the suspicion of malignancy breast tissue but cannot take a specific information about the detected calcification. This study focusses on (FNB) as an important role for examining the abnormality of breast tissue cells that consists of obtaining material directly from the tumor. The collected material is then tested using a microscope to conclude the prevalence of abnormality nuclei cancer cells [17].

C. Dataset

The dataset that is used for training, cross validation and testing of the proposed system consists of total 130 patient cases. 65 of them are malignant cases and 65 are benign. Each case was represented by tested area that is selected from its virtual slide. The breast cancer images are 24-bit RGB color space (8 bits for each channel) of (JPEG) Some images of breast cancer are chosen from the dataset that is used in this work [18]. Two types of breast cancer have been used in this thesis (benign and malignant images) which are shown in Fig. 3.

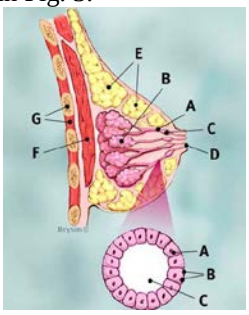
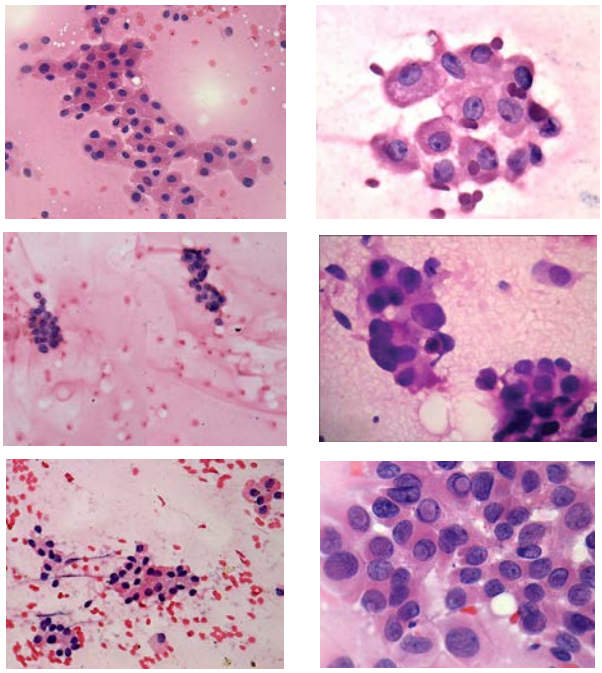


Fig. 2. Breast Anatomy [16]



(a) (b)

Fig. 3. Maligned Breast Cancer image samples (a) is a benign case, (b) a malignant case

2. Related Works

There were many types of research that recently have been interested in using the computer-aided methodology for cytology and digital pathology imaging system. Some of them dealing with the breast cancer tumor diagnosis by analysis of the cytological images.

Jeleń et al. [19] have presented a prediction and classification approach for breast cancer based on using the level sets segmentation method. level sets segmentation method used as a method for breast cancer cell detection. Then they used the whole segmented image to extract the features which were used for the classification model. Although in this work, the classification effectiveness approaches were tested on total 110 cases, 44 were malignant, and 66 were benign. The highest accuracy for classification and prediction results that have been satisfied in this approach was 82.6%. In Niwas et al. [20] have presented another method which based on the analysis of nuclei cells texture using another domain by wavelet transform. As we can see in this work there is no segmentation and detection approach for breast cancer detection, so they depend on the wavelet transform a texture domain to extract the tissue features. In this work, for classification approach effectiveness proposed a k-nearest neighbor algorithm, and it has been tested on 45 (20 malignant, 25 benign) images. the highest accuracy for this approach was reached to 93.33%. Malek et al. [21]

have proposed an active contour as segmentation to segment nuclei cell. They used the whole segmented cell images to extract the features which were used for the classification model. Their model has been used to classify 200 cases, 80 of them were malignant, and 120 were benign. The main classification algorithm that has been proposed in this approach was fuzzy c-means algorithm. The highest accuracy that has been achieved in this approach was 95%. Xiong et al. [22] have presented for breast a Partial least squares regression. This approach did not describe the segmentation approach that used to extract nuclei, as well as the feature extraction approach that used for the classification model. This approach was used to classify 699 cases, 241 of them were malignant, and 458 were benign images. The higher accuracy of breast cancer prediction and classification approach that has been achieved in this approach was 96.57% effectiveness. Pawel et al [23] have presented an approach for breast cancer diagnosis by relying on the Circular Hough Transform (CHT) for nuclei cells detection and using SVM classifier for cells filtration. This approach depends on a semi-supervised learning approach by extracting a set of 25 features were extracted from the nuclei cells images that been extracted by using CHT and SVM. After the SVM has been trained after a 300 cells images were manually selected. SVM classifier is used to classify the complete diagnostic procedure. This approach was tested on 737 cases and achieved 98.51%.

3. Proposed system

In this paper, we propose a comprehensive fully automatic supervised breast cancer diagnostic system. This approach is based on the localization of cells images of FNB slides images. The main task of this approach is to predict and classify a case of FNB image as a benign or malignant case. This is done by first detecting and isolating the nuclei cells image by applying morphometric, textural and topological features. Then we used the filtered image as a new dataset for the final classification approach by using the Convolutional Neural Network (CNN). The proposed approach depends on three main steps to predict and classify the breast cancer. Eventually, in the first step, we settled on the fast and robust approach which determines the localization approach for nuclei cell detection and isolation. The localization methodology based on perfect circular cells detection using the CHT creates a new dataset containing 12350 cells images records. The second step is the nuclei cell filtration using a fully automatic SVM approach which automatically selects a set of training cells images to train the SVM classifier. This step is designed to retain only those circular nuclei cells images. The third and final step is the final classification. In this step, we use such a powerful classification framework by proposing a Convolutional Neural Network (CNN) as a feature extraction and final classification approach. The code used to process these steps was developed in MATLAB environment. Figure 4 demonstrates the flowchart of our proposed method, and the details of each module are described below.

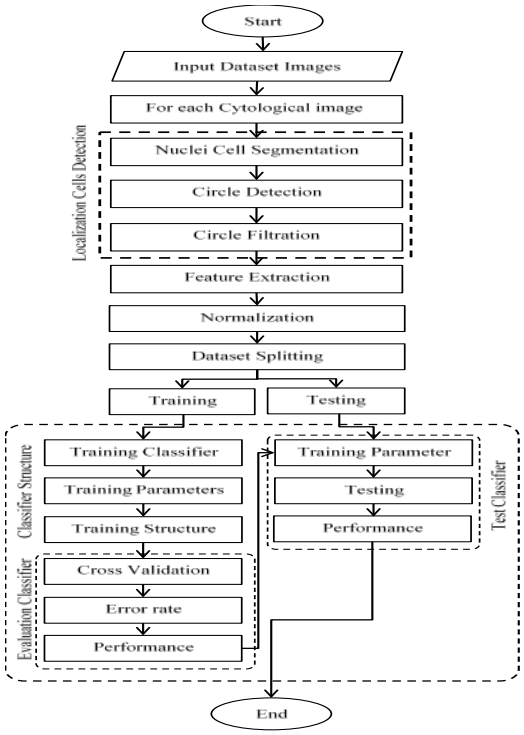


Fig. 4. Block digram of the proposed system

3.1 A fully Automatic Localization Nuclei Cells Selection Approach

Localization approach is a method that we proposed to determine the circular nuclei cells in the slide images. The cell nuclei need to be filtered and isolated from the background and from other objects on the tested slide image. Those images have such uncorrelated objects such as red blood cells, and cytoplasm. This task in this approach is fully automatically done by detecting the cells images first using Circular Hough Transform (CHT), then another approach has been proposed to select automatically a set of perfect nuclei cells images to train the SVM classifier that will automatically classify the whole cells of the original dataset to correct nuclei cells and incorrect ones. In our case, we set the automatic perfect cells selection to 500 cells images as a training set to train the SVM classifier and test the whole 12350 detected circle cell images. By experimental result, we found that number (500) is the perfect one to reach the higher accuracy of the training and the test of the localization approach. The whole flowchart of the

localization step is illustrated in Fig. 5. The main steps of the localization approach are described below:

A. Circularity Nuclei Cells Detection

As a first step on the localization approach, the CHT is used depending on the previous [23] approach which approved that the CHT was more accurate than the other approaches for cells detection in medical images.

Hough transform is a method that can be applied to detect such a circular shape in a given image [24]. Circle Hough Transform (CHT) has been designed to find a circle shape characterized by a detect the center point (x_0, y_0) of the circle in addition to the radius r. However, Ellipse Hough Transform (EHT) also applied to find the elliptical formations coded by detecting the center (x_0, y_0) and the orientation of the ellipse Θ of its semi axes a and b. CHT algorithms are mainly used to detect circles and ellipses which are computationally more expensive than line detection algorithms in any tested image according to a large number of parameters involved in describing the shapes. The main procedure to determine a circle in any image, it is necessary to compute the accumulate votes in the three-dimensional parameter space which is (x_0, y_0, r). Although, detecting an ellipse in the image the search must be performed in the five-dimensional parameter space which is (x_0, y_0, Θ, a, b) [24]. The circle or ellipse are simply presented in parameter space, by compared to the line, since the parameters of the circle can be directly transferred to the parameter space. The circle detection equation is defined by Eq. (1) [24]:

$$r^2 = (x - a)^2 + (y - b)^2 \quad (1)$$

Where a and b are the center of the circle in the x and y direction, and r is the radius of the detected circle. The parametric representation of the detected circle in any image is defined by Eq. (2) and (3) [24]:

Therefore, the role of the CHT for circle detection is to search for the triplet parameters in each image which are (a, b, r) that determines the points of (x_i, y_i) [24].

$$x = a + rcos(q) \quad (2)$$
$$y = b + rsin(q) \quad (1)$$

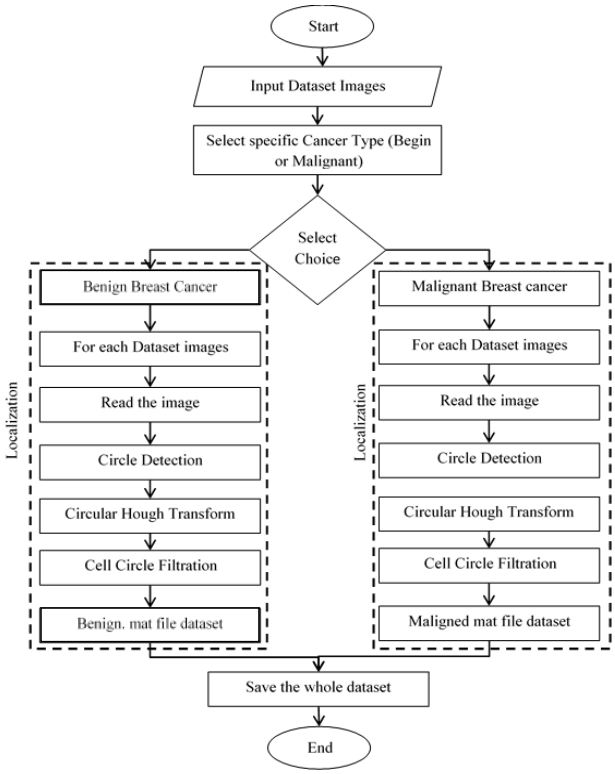


Fig. 5. Localization neculei cells detection and filteration approach

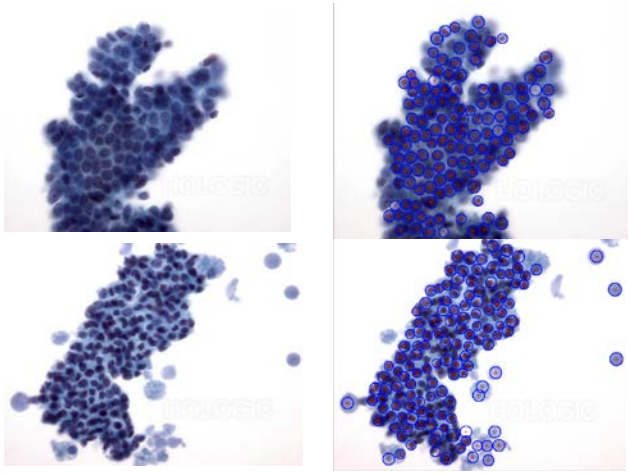


Fig. 6. Another example of cell detection using (CHT) Circular Hough transform (a) is the original image, (b) is the CHT image

The first step of the localization algorithm is to detect edges in the image by letting I to be a grayscale tested image. We define edge differently on the previous approach [23] by selecting the indicator by the following equation (4) [24]:

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

where ∇ is the gradient operator of the threshold value. In the processing step, we select the red channel as it proposed in the previous approach [23] where the difference in values between nuclei and red blood cells is the greatest. Also, the cytoplasm, which surrounds the nuclei, is barely visible. After choosing the optimal and an appropriate channel (Red) which the nuclei cells are more visible in. We applied the proposed algorithm for cell detection which is the Circular Hough Transform. Examples of detected edges that has been applied during the Circular Hough Transform. To attempt the first step of the localization approach, we applied the circle detection algorithm depending on the (CHT) that summarized in Algorithm (1) [24], and some example results of cells detection using (CHT) is shown in Fig.6.

Algorithm (1) Circular Hough Transformation (CHT)

Input: Tested Image

Output: Number of circles and dimensions (x, y) for each one

1. Find Edges
2. **//HOUGH BEGIN**
3. **For** each edge point
4. Draw a circle with center in the edge point with r
5. Increment all coordinates that the perimeter of the circle passes through in the accumulator
6. Find one or several maxima in the accumulator
7. **//HOUGH END**
8. Map the found parameters (r, a, b) corresponding to the maxima back to the original image

B. Circle Cells Filtration and Isolation using SVM

The variation of nuclear sizes is relatively high. so, sometimes a circle which has been detected by using the Circular Hough Transform (CHT) comprises two or more nuclei simultaneously. There are also other objects present in the images such as red blood cells as it shown before in Fig.6. Although much brighter in the red channel than the nuclei, they are occasionally detected by the Hough transform (CHT). Another issue is false positives caused by some cases, for example, geometric arrangements in the background being incorrectly identified as the boundary of a nucleus. To select the perfect nuclei cells, we suggest doing the circular cells filtration and isolation. In this case, we need to remove all these nonnuclear objects that we have detected during the cells detection step such as a cytoplasm.

C. Automatic Training Images Selection for Cells Filtration using Variation Image Histogram

In term of training the SVM for cell filtration in some images which consist of the nuclei cell images, we proposed an automatic approach for cell image selection as it shown in Fig.7 to train the SVM classifier to do the cells filtration. In this case, the automatic cell images selection depends on three categories to prepare the selection:

1) Variation of the Color Intensity Feature:

The automatic nuclei cells selection approach depends on the blue channel to perform the cell image segmentation. In this case, selection approach ensures accepting just the infection nuclei cell that has a high variation, This is done by computing the histogram of three color variation and accepting just the cell that has a high variation on the blue channel compared with the other channels, such as the blood cell images as it shown in equation (5), and the incorrect cell detection images.

$$Var(x) = \frac{1}{n^2}\sum_{i}\sum_{j>1}\left(x_i - x_j\right)^2 \quad (5)$$

Fig. 8 shows an example of the accepted nuclei cell and the rejected one. We can notice that in Fig.8, (a) the distribution of the blue channel is higher than the other color, therefore, this cell has been selected. In contrast, in Fig.8, (b) we can see that the distribution of the red color is higher than the other color which in this case this cell has been rejected.

2) Geometric Features using Self-Organization Maps Neural Network (SOM)

Self-Organizing Maps (SOMs) is described by Teuvo Kohonen [25] which is an unsupervised neural network technique that can be used to visualize the high-dimensional data sets in lower dimensional representations. SOM is made up of multiple nodes. Each node vector has a fixed position on the SOM grid, a weight vector of the same dimension as the input space, and an associated sample from the input data. Each sample in the input space is "mapped" or "linked" to a node on the map grid. One node can represent several input samples.The algorithm to produce a SOM from a sample data set can be summarized as Algorithm (2) shows:

Algorithm (2) Self-Self-Organizing Map (SOM)

1. **Select** the size and type of the map. Typically, hexagonal grids are preferred since each node then has 6 immediate neighbors.
2. **Initialize** all node weight vectors randomly.
3. **Choose** a random data point from training data and present it to the SOM.
4. **Find** the "Best Matching Unit" (BMU) in the map – the most similar node. Similarity is calculated using the Euclidean distance formula.
5. **Determine** the nodes within the "neighborhood" of the BMU such as the size of the neighborhood decreases with each iteration.
6. Adjust weights of nodes in the BMU neighborhood towards the chosen data point.
7. Repeat **Steps 2-5** for N iterations until convergence.

In this case, we use unsupervised learning neural network through the SOM to accomplish the localization segmentation by segmenting each pixel in the cell images to get perfect segmentation, after that we undertake the segmentation tracking to compute the boundary of each

cell image in blue channel to select the perfect boundary or nuclei image. This produces a binary mask through the SOM. The Morphological image was used to remove some binary noise to get normalized edges using image closing. Filling the holes in each cell is achieved to get perfect map, the automatic cell images selection will test every single cell image that has passed through the intensity color feature in the blue channel. During this stage, the proposed approach for automatic cell image selection will be tested via computing the circular arteries by computing the perimeter of all area as it shown in equation (6) [26].

$$circularities = \frac{Perimeters^2}{(4\text{xpixallAreas})} \quad (6)$$

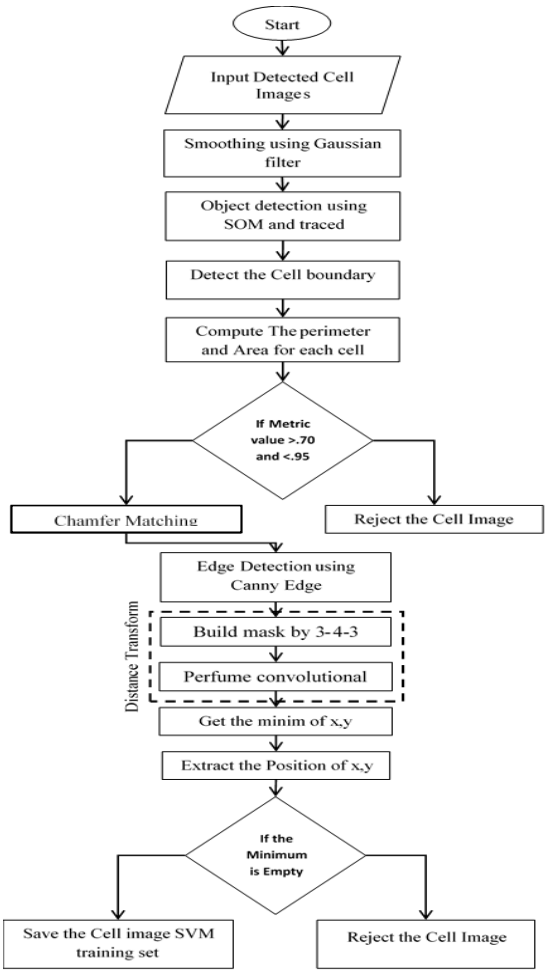


Fig. 7. Block diagram of the full automatic cell images filteration using chamfer matching for SVM training cell images selection.

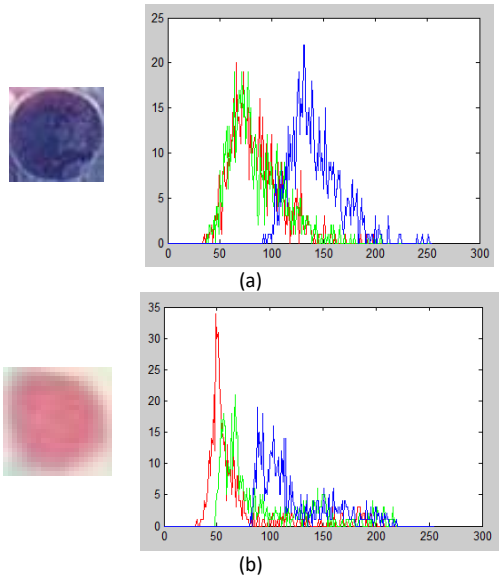


Fig. 8. Histogram of the variation of color intensity feature of correct necli cell selection, (a) Correct neculei cell, (b) incorrect neculei cell.

3) Every cell passed the circularities value which is greater than (0.70 and less than 0.95) which are best circularities value to present the perfect cells that we have selected by trying many values by training then will be selected as a correct cell, rather than it will be rejected as an incorrect cell image. The tested values have been discovered by experimental results after we tried and test many values in this range. The best one that gives us better result, since if the values came as 100% this means the supposed nuclei cell image contain one texture not image with nuclei cell or if came with values of 50% this mean that images were false negative (cytoplasm, etc.) detected by CHT and a good number of cell images that have been selected.

4) Perfect Ellipse/Cell shape matching using Chamfer Matching and Distance Transform:

The last test step of the automatic cell image selection for the SVM training in the filtration step has been realized on the ellipse and circle shapes. To make sure that we just select the correct nuclei cell images by the perfect ellipse and circle shape, the chamfer matching and distance transform approach have been proposed. They select just the perfect cell image that will be fitted in an ellipse or circle shape. The Chamfer System offers a high-performance solution to shape-based object detection. It covers the detection of arbitrary-shaped objects, whether parameterized (e.g. rectangles and ellipses) or not (e.g. pedestrian outlines). Because the system learns shape distributions of target objects from examples, it is flexible and easily adaptable without reprogramming. It is a pixel-based correlation approach that eliminates the need for error-prone contour segmentation. There are three types of

distance measures, in our case, we used the Euclidean distance between P and Q it is defined as shown in equation (7) [26] [27]. The whole approach of the automatic cell images selection for the SVM training is shown in Fig. 9.

$$D_e(P, Q) = \sqrt{(x - u)^2 + (y - v)^2} \qquad (7)$$

The flowchart of the automatic nuclei cell images selection for training the SVM in the filtration stage is shown in Fig. 11, furthermore, some examples for the nuclei cell images that have been successfully passed and selected by the automatic cell images selection are shown in Fig. 8.

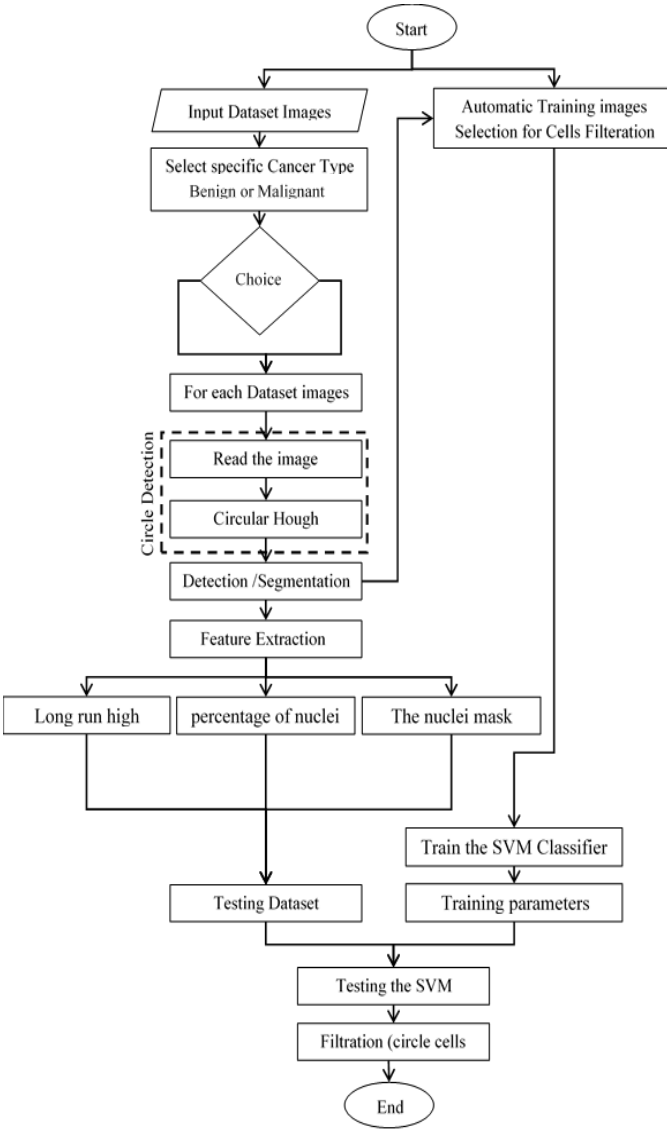


Fig. 9. Proposed approach for the full automatic cell images filteration using chamfer matching and SVM.

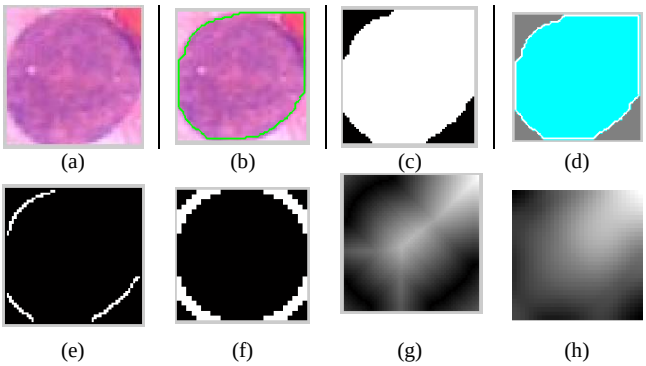


Fig. 10. An example of correct neclei that has been passed and selected, (a) original cell image, (b) traced automatic boundray detection, (c) binary mask, (d) metric area, (e) edge detection, (f) image tamplate, (g) Distacne Map, (h) Matching Error

5) Feature Extraction for SVM Training Set

After the perfect cell images have been selected for training the SVM classifier to filter the whole dataset and select the nuclei cell images, we use a support vector machine as a classification approach for nuclei cells filtration and isolation. In term of training the SVM, some features are extracted. The training dataset for the SVM classifier is prepared by computing three features extracted from inside the nuclei cells areas and outside. in addition, with the features inside the regions of circularity nuclei cells, the features from the surrounding texture of wholly cropped cells images that detected by using (CHT) were calculated the three features as it has been proposed in the previous approach [23]: the mean value of pixels inside the circle in the blue channel as it shown in the in Table 1 [23].

Table 1. Filtration Features step

Feat. No.	Description
1	The mean value of pixels inside the circle in the blue channel
2	Long run high gray-level
3	The percentage of nuclei pixels per the nuclei mask

Where all the three features that have been used in this step which are: first the mean value of pixels inside the circle in the blue channel in the tested image, second A long run high gray-level emphasis determined by using gray -level run length matrix. Then, the percentage of nuclei pixels according to the nuclei mask. Finally, the nuclei mask which is obtained by conducting Otsu's thresholding on the red channel of the image.

Otsu's method: Otsu's method is aiming to find the optimal value for the global threshold. This method relies on a measure of the region homogeneity which is the variance. In another word, the regions with high homogeneity will have low variance. Otsu's method selects the threshold by minimizing the within-class variance of the two groups of pixels separated by the thresholding operator. It does not depend on modeling the probability

density functions, however, it assumes a bimodal distribution of gray-level values (i.e., if the image approximately fits this constraint). The Otsu's method [28] that has been proposed in this step as a global and primal thresholding.

- **Binary Mask:** The result of the binary mask is obtaining like where the dark objects like nuclei are zeros as well as the bright background pixels are ones. The final value of the feature is computed by equation (8) [23].

$$PNM = \frac{n_{mask}}{n_{all}} \quad (8)$$

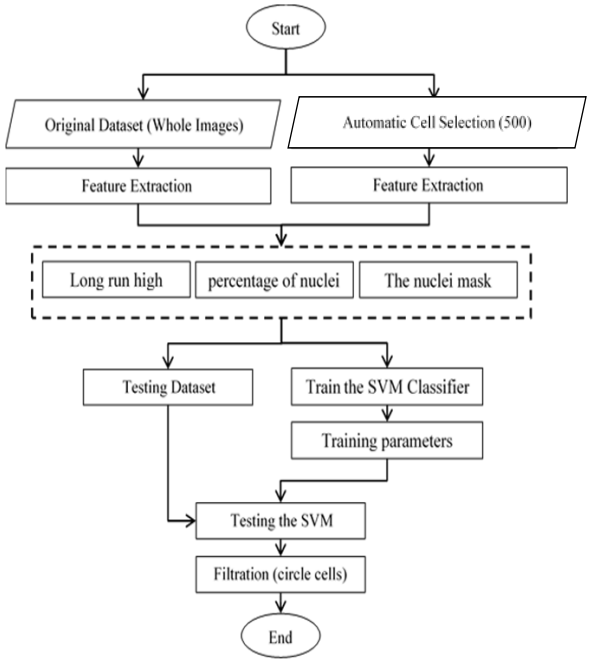


Fig. 11. SVM neclei cells filteration approach

Where n_{mask} is the number of pixels inside the circle, for which the mask value is 0, and n_{all} is the number of all pixels inside the circle.

6) Cells Filtration using Support Vector Machine

In this step, the whole dataset that we have extract through the localization approach using the (CHT) is used as a testing set after we trained the SVM classifier on the training set that we have selected automatically using a full automatic cell images selection approach. The main flowchart of the circle cell filtration is illustrated above in Fig. 11.

7) SVM Classifier Algorithm:

In what follows the cell images filtration to be used in the classifier design. The Support Vector Machine (SVM) is trained and applied with the intention of enhancing the predictive power of our cell image filtration (classifiers). In this case, the input space is not linearly separable and we need to rely on soft margin SVM which both maximizes the margin w and minimizes the errors Eq. (9) subject to Eq. (10) and Eq. (11).

$$\min \frac{1}{2}|w|^2 + C\sum_i \xi_i \quad (9)$$

$$s.t.\ \ y^{(i)}\left(w^T x^{(i)} + b\right) \geq 1 - \xi_i \quad (10)$$

$$\xi_i \geq 0\ \forall_i \quad (11)$$

The final Lagrangian dual formulation becomes Eq. (12), Eq. (13) and Eq. (14).

$$\max_{\alpha \geq 0} \mathcal{L}(\alpha) = \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y^{(i)} y^{(j)} x^{(i)^T} x^{(j)} \quad (12)$$

$$s.t \sum_i \alpha_i y^{(i)} = 0 \quad (13)$$

$$0 \leq \alpha_j \leq C\ \forall_i \quad (14)$$

Now α_i's upper bound is $C$and the solution is Eq. (15)

$$.w = \textstyle\sum_{i \in Ns} \alpha_i y_j x^{(i)} \quad (15)$$

Then we use Sequential Minimal Optimization (SMO) to solve for each pair of αi and αj by freezing other variables. C is left at its default value of $C = 1$. The three kernel functions that have been used in the experiments to compute the inner product in the Lagrangian dual formulation Eq. (15) are the Linear Eq. (16), Gaussian Radial Basis Function Eq. (17), and Polynomial Kernels Eq. (18).

Linear:

$$G(x_i, x_i) = x_i^T x_j \quad (16)$$

Gaussian RBF:

$$G(x_i, x_i) = e^{-\|x_i - x_j\|} \quad (17)$$

Polynomial:

$$G(x_i, x_i) = \left(1 + x_i^T x_j\right)^2 \quad (18)$$

The Gaussian Radial Basis Function (RBF) tuning as it given in Eq. (19) can be achieved by scaling the input vectors by a scalar value σ before the kernel transformation Eq. (20), resulting in Eq. (21). Alternatively, Matlab can automatically select the optimal scaling via heuristic procedure using subsampling.

$$\mathrm{K}\left(\mathrm{x}^{(\mathrm{i})}, \mathrm{x}^{(\mathrm{j})}\right) = \left(1 + \mathrm{x}^{(\mathrm{i})^{\mathrm{T}}} \mathrm{x}^{(\mathrm{j})}\right)^{\mathrm{p}} \quad (19)$$

$$x' = \frac{x}{\sigma} \quad (20)$$

$$G(x_i, x_i) = e^{\frac{-\|x_i - x_j\|}{\sigma^2}} \quad (21)$$

According to what we have discussed above and in order to obtain perfect nuclei cells, we suggest a circle cell classification using support vector machine (SVM). In this step, the circles are then classified as correct or incorrect using a support vector machine with a Gaussian radial basis function kernel at scaling factor $\sigma = 0.8$. The classifier was trained on a fully automatic nuclei cells prepared database of 500 circles. The database contained 500 properly detected nuclei and 500 incorrect detections, which included red blood cells, joined and overlapped nuclei, as well as false positives. The wholly localized 12350 Circles cells by using CHT were used as a testing set. The filtration approach using support vector machine (SVM)is described and illustrated in the Fig. 11.

The recognition rate was the percentage of successfully recognized circles (as correct or incorrect) among all 500 circles. The best combination of a classifier and feature subset gave 98.64% recognition rate. The filtration accuracy is the percentage ratio of successfully classified circles, as correct or incorrect, to the total number of circles. Table (2) shows the detailed results. Three features were used: the mean value of pixels in the blue channel, Long Run High Gray-Level emphasis determined using Gray-Level Run-Length Matrix, and the percentage of nuclei pixels a cording to the nuclei mask

Table 2. Results of filtration accuracy as correct or incorrect using Support Vector Machine (SVM) on database of 500 fully automatic Cells Selection

Training Set				
Experimental Results				***Training Accuracy***
Confusion Matrix		***Performance Results***		
TP	98.74	***Sensitivity***	98.74%	98.64%
FN	1.260	***Specificity***	98.54%	
FP	1.460	***Precision***	98.54%	
TN	98.54	***Accuracy***	99.64%	
Testing Dataset				
Experimental Results				***Testing Accuracy***
Confusion Matrix		***Performance Results***		
TP	97.01	***Sensitivity***	97.01%	96.91%
FN	2.990	***Specificity***	96.81%	
FP	3.190	***Precision***	96.82%	
TN	96.81	***Accuracy***	96.91%	

4. Final Prediction and Classification using Convolutional Neural Network (CNN)

After the isolation of nuclei from the images, as determined by the circles classified as correct in the previous step, we proposed a convolutional neural network (CNN) as a final classifier for the breast cancer diagnosis system. A Convolutional Neural Network (CNN) is a function g that mapping data x, (e.g. an image) to an output vector y. The function g is the combination of a sequence of simpler functions f_l, which we call computational blocks, or layers; $g = f_l \dots f_L$. Assume the network input is $x_0 = x$, and the network outputs are, $x_1, x_2, \dots, x_L$. Each output $x_L = f_l(x_{1-1}; w_l)$ is computed from the previous output x_{1-1} by applying the function f_l with parameters w_1 [29]. The data flowing through the network represents a feature field; $x_1 \in \mathbb{R}^{H_l \times W_l \times D_l}$. Since the data x has a spatial structure, H_l and W_l are spatial coordinates, and D_l is a depth of channels. The functions f_l act as local and translation invariant operators, therefore, the network is called convolutional. CNNs are applied to distinguish between different classes by producing a vector of probabilities $\hat{y} = f(x)$ for all image labels. If y is the true label of image x, CNN performance of true label y of image x is measured by a loss function $\ell_y(\hat{y}) \in \mathbb{R}$ which assigns a penalty to classification errors [30]. However, CNN consists of three main types of layers: convolution layers, Max-pooling layers, and a fully connected layer [32].

- **Convolutional layer:** Convolutional layer convolves the result of the previous layer with a set of learnable filters [32] as shown in Fig. 13, where the weights specify the convolution filter. Each filter is slide to across the width and height of the input volume, producing a 2-dimensional activation map of that filter. The filters have the same depth as in the input [32] [29]. The size of the output can be controlled by three hyper parameters which are the depth, stride and zero-padding
- **Pooling layer:** Pooling layer reduces the size of their input and allows multi-scale analysis. Max- pooling and average-pooling are the most popular pooling operators which are used to compute the maximum or the average value within a small spatial block [32]. Pooling with filters size of 2 × 2 with a stride of 2 are considered ideal [33].
- **Fully-connected layer:** Fully-connected layer connects to all the neurons of the previous layer [33]. Fully connected layers are typically used as the last layer of the network and perform the classification. A sample of CNN is depicted in Fig. 13, which shows all the three previously demonstrated layers.

Z-score normalization is used to normalize the attribute of the features vector in the fully connected layer in the CNN structure. It normalizes the attribute value such that the mean and standard deviation after normalization become zero and one respectively. For this property of normalization, z-score is also called as zero mean normalization. Its mathematical equation is given in equation (22) [32].

$$x_i' = \frac{x_i - \mu(x)}{\sigma(x)} \quad (22)$$

where σ: is the standard deviation of the attribute (x).In this paper, we proposed a formulated nuclei cell classification

as a binary classification problem and it learned the mapping from the original cell image that has been detected and filtered using SVM classification to a binary classification mask. The overall CNN structure that is used in the final classification approach is shown in Table (3).

Table 3. Architecture of the CNN training/Testing model for the Final Classification Approach

Layer Number	Layer Type	Parameter	
		Kernel No.	*Kernel Size*
Layer 1	Batch Normalization	-	-
Layer 2	Convolution	32	5×5×3
Layer 3	Pooling	ReLU	3×3
Layer 4	Convolution	32	5×5×32
Layer 5	Pooling	ReLU	3×3
Layer 6	Convolution	64	5×5×64
Layer 7	Pooling	ReLU	3×3
Layer 8	Convolution	64	4×4×64
Layer 9	Fully Connected	-	1×1×64
Layer 10	Softmat with log-loss	-	Activation F.

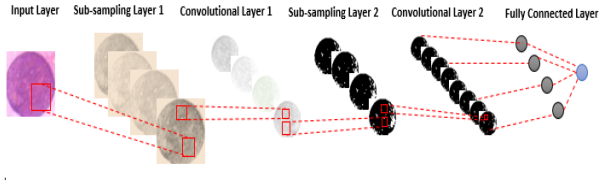


Fig. 12. A sample of CNN architecture .

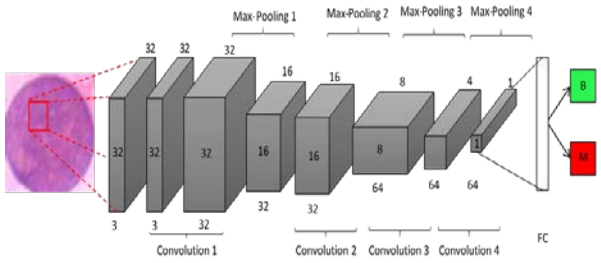


Fig. 13. Nuclei cells classification using convolutional neural networks.

5. Experimental Results

To evaluate the effectiveness and accuracy of the fully automatic breast cancer diagnosis-based localization approach for nuclei cells selection and filtration with the convolutional neural network (CNN). A confusion matrix framework is defined as $m\ x\ m$ matrix, where m denotes the number of classes. In our methodology, a binary classification problem is an appropriate approach to classify the nuclei cells image to a benign or malignant case. In this case, the confusion matrix contains information about actual and predicted classification which is done by the Convolutional Neural Network (CNN). The performance of such systems is commonly evaluated using the data in the whole matrix. Each column of the matrix represents the instances in a predicted nuclei cell image class, while each row represents the instances in an actual nuclei cell image class.

A. Performance Evaluation Measures

Performance of nucleus detection was assessed in terms of accuracy Eq. (23), precision Eq. (24), sensitivity Eq. (25), and specificity Eq. (25), where True Positive (TP) refers to correct classifications of positive cases, True Negative (TN) refers to correct classifications of negative cases, False Positive (FP) refers to incorrect classifications of positive cases into negative class, and False Negative (FN) refers to incorrect classifications of negative cases into class positive.

$$Accuracy = \frac{TP + FN}{TP + FP + FN + TN} * 100 \quad (23)$$

$$Precision = \frac{TP}{TP + FP} \quad (24)$$

$$Sensivity = \frac{TP}{TP + FN} \quad (24)$$

$$Specificity = \frac{TN}{TN + FP} \quad (25)$$

B. Training and Testing approach

The training and testing framework approach for the dataset that used for the final classification using CNN is done by splitting the whole dataset which is after the filtration the total number of cells is (865 cell images), we split the data set to 80% of training which is (692 cell images) and 20% of testing which is (173 cell images).

C. CNN Prediction and Classification Results

Table (4) shows the classification results of the final classification results relies on using the CNN approach. The highest accuracy was (99.89%) in the training set and (99.42%) in the testing set.

Table 4. CNN classifier results using feature selection approach

Training Set				
Experimental Results				***Training Accuracy***
Confusion Matrix		***Performance Results***		
TP	99.981	***Sensitivity***	99.98%	99.9%
FN	0.0188	***Specificity***	99.78%	
FP	0.2188	***Precision***	99.78%	
TN	99.781	***Accuracy***	99.88%	
Testing Dataset				
Experimental Results				***Testing Accuracy***
Confusion Matrix		***Performance Results***		
TP	99.899	***Sensitivity***	99.90%	99.8%
FN	0.1001	***Specificity***	99.70%	
FP	0.3001	***Precision***	99.70%	
TN	99.699	***Accuracy***	99.80%	

D. Comparing with other approaches

Table (5) shows the performance result of our methodology against the related works that have been proposed in the same area. It has been clear that the closet accuracy is (98.51%) that was proposed from Pawel [23], by using the CHT for cell detection, by extracting 50 features that have been used with the SVM. Our approach differs on the previous one Pawel [23], by replacing the hand craft feature extraction part and the SVM classifier by (CNN). CNN approach gives us higher performance for feature extraction and classification. After the first part of our approach has filtered the nuclei cell images and isolate them to train and test the CNN classifier. By that, our approach has reached to (99.42%) with (0.91) more than the previous approach that was achieved (98.51%).

TABLE 5. ARCHITECTURE OF THE CNN TRAINING/TESTING MODEL FOR THE FINAL CLASSIFICATION APPROACH

Approach	Accuracy	Methods
Telen [19]	82.6 %	Level Set Segmentaion
Niwer [20]	93.33%	Wavelet Transform
Malek [21]	95.00%	Active Countour
Xiong [22]	96.57%	Prtial Least Square Error
Pawel [23]	98.51%	CHT&SVM
Our Approach	**99.80%**	**Convolutional Neural Network**

6. Conclusion

This paper proposes a new approach that is based on the localization methodology for a fully automatic breast cancer image diagnosis system using convolutional neural network (CNN) classification approach. In this work, we have used an approach for nuclei cells detection and isolation that is a modified one to what was proposed by Pawel [23], with major modifications. The first contribution and modification that we have done in this approach is to make the nuclei cells detection and filtration by using the SVM fully automatic. Instead of relying on manually selecting 500 perfect cell images to train the SVM classifier, we proposed a fully automatic cell images selection to train the SVM classifier. Variation of the Color Intensity in blue channel, Self-organization map (SOM) and Chamfer matching with distance transform were utilized to select a set of perfect cell images to train the SVM classifier. We notice that our approach achieved (99.42%) diagnosis accuracy instead of (98.51%) that has been achieved by the precious approach. The second contribution is that we have replaced the feature extraction part with the SVM classifier that was proposed as a more robust and powerful approach for feature extraction and classification within the CNN framework. This paper shows that our approach has satisfied about (0.91) outperform the previous approach [23], the main challenge for this study lies in the complexity of non-homogeneous high-resolution slide images dataset that we used in our approach this stipulates that the proposed system is more robustness for breast cancer diagnosis tasks and it is fully automatic approach. Future work includes testing our methodology on other deep learning structuring algorithms and optimizing the performance of each individual deep learning method. Future work includes testing our methodology on other deep learning structuring algorithms and optimizing the performance of each individual deep learning method.

References

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

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

[3]. P. Britton, S. Duffy, R. Sinnatamby, M.Wallis, S. Barter, M. Gaskarth, A. O'Neill, C. Caldas, J. Brenton, P. Forouhi, and G. Wishart, "Onestop diagnostic breast clinics:" Br. J. Cancer, pp. 1873–1878, Jun. 2009.

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

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

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

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

[8]. O. Lezoray, A. Elmoataz, and H. Cardot, "A color object recognition scheme: Application to cellular sorting," , vol. 14, no 3, 2003.

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

[10]. Y. Lan, H. Ren, and J. Wan. A hybrid classifier for mammography CAD. In Fourth IEEE, International Conference on Computational and Information Sciences (ICCIS), pages 309–312,2012.

[11]. K. Ganesan, U. R. Acharya, C. K. Chua, L. C. Min, K. T. Abraham, and K. Ng. Computer aided breast cancer detection using mammograms: a review. IEEE Reviews in Biomedical Engineering, 6:77–98, 2013.

[12]. B. Hela, M. Hela, H. Kamel, B. Sana, and M. Najla. Breast cancer detection: a review on mammograms analysis techniques. In 10th IEEE International Multi-Conference on Systems, pages 1–6, 2013.

[13]. LPCC. Programa de rastreio de cancro da mama da Liga Portuguesa Contra o Cancro. Liga Portuguesa Contra o Cancro (LPCC), 2009.

[14]. Worldwide Breast Cancer. Breast cancer statistics worldwide. Worldwide Breast Cancer,2009.URL www.worldwidebreastcancer.com/learn/breast-cancer-statistics-worldwide.
[15]. Erickson, Carissa" Automated detection of breast cancer using saxes data and wavelet features", (Unpublished doctoral dissertation) university of Saskatchewan,Saskatoon, 2005.
[16]. Breastcancer.org,http://www.breastcancer.org/symptoms/ understandbc/ what_is_bc.
[17]. Simon S Cross, Robert F Harrison," Fine Needle Aspirate of Breast Lesions Dataset", Senior Lecturer, Department of Pathology, University of Sheffield Medical School, Beech Hill Road, Sheffield UK.
[18]. http://www.prevencaoediagnose.com.br/web.inf.ufpr.br/vri/ breast-cancer-database
[19]. L. Jeleń, T. Fevens, and A. Krzyżak, "Classification of breast cancer malignancy using cytological images of fine needle aspiration biopsies," Int. J. Appl. Math. Comput. Sci., vol. 18, no. 1, pp. 75–83, 2010.
[20]. I. S. Niwas, P. Palanisamy, and K. Sujathan, "Wavelet based feature extraction method for breast cancer cytology images," in Proc. 2010 IEEE Symp. Indust. Electron. Appl., 2010, pp. 686–690.
[21]. J. Malek, A. Sebri, S. Mabrouk, K. Torki, and R. Tourki, "Automated breast cancer diagnosis based on GVF-Snake segmentation, wavelet features extraction and fuzzy classification," J. Signal Process. Syst., vol. 55, pp. 49–66, 2009.
[22]. X. Xiong, Y. Kim, Y. Baek, D. W. Rhee, and S.-H. Kim, "Analysis of breast cancer using data mining & statistical techniques," in Proc. 6th Int. Conf. Software Eng., Artif. Intell., Netw. Parallel/Distribut. Compute 1st ACIS Int. Worksh. Self-Assemb. Wireless Netw., 2005, pp.82–87.
[23]. Paweł Filipczuk, Thomas Fevens, Adam Krzyżak,"Computer-Aided Breast Cancer Diagnosis Based on the Analysis of Cytological Images of Fine Needle Biopsies", IEEE TRANSACTIONS ON MEDICAL IMAGING, VOL. 32, NO. 12, DECEMBER 2013.
[24]. D. Kerbyson and T. Atherton, "Circle detection using Hough transform filters," in Proc. 5th Int. Conf. Image Process. Appl., U.K., 1995, pp. 370–374.
[25]. Kohonen, Teuvo (1982). "Self-Organized Formation of Topologically Correct Feature Maps".
[26]. http://www.mathworks.com/help/images/examples/identifying-round objects.html.
[27]. Fukunaga and Hostetler, "The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition", IEEE Transactions on Information Theory vol 21 , pp 32-40 ,1975.
[28]. T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," Communication Technology, IEEE Transactions on, vol. 15, pp. 52-60, 1967.
[29]. D. Comaniciu, et al., "Kernel-based object tracking," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 25, pp. 564-577, 2003.
[30]. Hai Su, Fujun Liu, Yuanpu Xie, Fuyong Xing, Sreenivasan Meyyappan, and Lin Yang. Region segmentation in histopathological breast cancer images using deep convolutional neural network. IEEE, 2015.
[31]. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems, pages 1097–1105, 2012.
[32]. Andrea Vedaldi and Karel Lenc. Matconvnet: Convolutional neural networks for matlab. In ACM International Conference on Multimedia, pages 689–692. ACM, 2015.
[33]. M Iftene, Q Liu, and Y Wang. Very high resolution images classification by fine tuning deep convolutional neural networks. In Eighth International Conference on Digital Image Processing, 2016.

Ali Fawzi was born in Baghdad, Iraq 1985. He received the B.S. degrees in computer sciences from AL Mustansiriya University, Baghdad, Iraq at 2006,he is now M.S.C postgraduate. He worked as programmer analysis in the information technology department, general investigator office, Iraqi, ministry of health and he has some contribution in Iraqi health information system (HIS), and he has contributed to health government programs system including remote medical consultant, medicine control system, and medical devices control system since 2010 until now.

Mehdi G. Duaimi was born in Babylon, Iraq, 1968. He received his B.Sc., M.Sc., and Ph.D. degrees, all in computer sciences from Nahrain University, Baghdad, Iraq at 1992, 1995, and 2007 respectively. In 2009 he joined the University of Baghdad, where he is now an instructor in the Department of Computer Sciences. During the 1999 – 2009 years, he was at the Iraqi commission for computers and informatics - Baghdad where he worked as a database designer and as an instructor. He has some publications related to data mining and information retrieval. His current research interests include areas like data mining, databases and artificial intelligence.