International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2016): 79.57 | Impact Factor (2015): 6.391

Breast Cancer Diagnosis System Based on Nuclei Cells Localization of Fine Needle Biopsies Images

Ali Fawzi Mohammed Ali[1], Mehdi G. Duaimi[2]

[1, 2]College of Science, Department Computer Science, University of Baghdad, Baghdad, Iraq

Abstract: ***Breast cancer is becoming a leading cause of death among women in the whole world; 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. Accurate diagnosis of this disease can ensure a survival of the patients. This paper presents an analysis of digital histopathology Breast cancer based on cytological images of Fine Needle Biopsy (FNB). The main approach of this study is relying on a localization approach for nuclei cell (cell detection). the nuclei are estimated by circular shape using the Circular Hough Transform (CHT). Then, the cells that have been detected by the (CHT) are then filtered to keep only high-quality and accurate cells that have been estimated for further analysis by using a supervised learning approach. In order to filter the nuclei cells and classify the detected circles as correct (cells) or incorrect Support Vector Machine (SVM) as an approach is proposed to use. A set of 25 features were extracted from the remaining filtered nuclei set. (50 features) produced by calculating the mean and variance for each feature. Support Vector Machine (SVM) and Backpropagation Neural Network (BNN) are the two classification algorithms of the biopsies that used in the final stage. The complete diagnostic procedure is tested on total 130 microscopic images of fine needle biopsies obtained from patients and satisfy (99.88%) classification accuracy by using Resilient Backpropagation Neural Network (RBNN) by selecting only (27 features)from total (50 features). The features selected using mutual information approach which is a distinction between benign or malignant. These results shows that our proposed method consider very promising compared to the previously reported results providing valuable, accurate, and stable diagnostic information.***

Keywords: breast cancer**,** localization, feature extraction, classification

1. Introduction

The International Agency for Research on Cancer (IARC) defines the breast cancer as the most diffuse cancer disease amongst women especially those ages between 40 and 55 years of old. 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 this cancer 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]. An automatic morphometric testing can improve the diagnosis allowing screening and testing on a large scale of medical materials. In some cases, they are uncertain and hard cases which would necessitate 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]. Data mining and machine learning approaches can be utilized in designing a computer based-method using to assist the doctors in the early stage of breast cancer diagnosis. Computer based-system offers a necessary treatment and prevents the influence that may lead to possibility of death [14]. This paper is presents a hybrid intelligent system which combine four methodologies: Circular Hough Transform (CHT) as a preprocessing step to Detect circular shapes in a grayscale image and resolve their center with positions and radii. In order to automatically filtered the nuclei cells from other material Support Vector Machine (SVM) as the classifier using for recognizing the circles supposed to be the nuclei cells as correct or incorrect**.** we extract a set of (25) morphological, texture and topological features from the filtered dataset which are then tested by two different classifiers. Support Vector Machine (SVM) with different three kernels of data separation and Backpropagation Neural Network (BNN) are two algorithms that used in the final classification as benign or malignant**.** The proposed hybrid approach produces a system exhibiting two prime characteristics: first, it attains high classification performance; second, the resulting systems involve a mathematical computation of discriminatory features therefore (human-) interpretable and fast. The rest of the paper is organized as follows. Section 2 gives the background information including breast cancer Diagnosis and classification problem and related work in corresponding area. The proposed localized intelligent system is illustrated in Sect. 3. In Sect. 4, different performance measurements are introduced which are commonly used for testing the effectiveness of automatic diagnosis system. The results obtained are given in Sect. 5. This section also includes the discussion of these results. Consequently in Sect. 6, the conclusion is given with summarization of results by emphasizing the importance of this study.

1.1 Background and motivation

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. 1, 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].

Licensed Under Creative Commons Attribution CC BY

1.2 Breast Cancer Diagnosis

Detection of breast cancer utilizes the screening method. In this method, examination by doctor or nurses to find tumors is used. In addition, screening methods include mammography and other imaging techniques. Screening can detect cancer in its early stages. Breast cancer detection by using the triple-test comprises self-test (palpation), and breast mammography or imaging ultrasonography devices, and fine needle biopsy (FNB). Mammograms screening which is a specialist technique that can be used to check and diagnosis for breast cancer in women that have no signs or symptoms of the disease however 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 which consists of taking material immediately from the suspected tissue mass. The obtained material is then examined using a microscope which conclude the prevalence of abnormality nuclei cancer cells [17].

1.3 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 paper (benign and malignant images) which are shown in Fig. 3

2. Related Works

There were many types of research that recently 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:

D. Belsare et al. [19] This approach proposes a method for breast cancer prediction and classification approach for histopathological images using texture features. This approach proposed different feature extraction scheme. Feature extraction schemes that have been used in this approach were Gray Level Co-occurrence Matrix (GLCM), Graph Run Length Matrix (GRLM) features, and Euler number are extracted. In this approach, the linear discriminant analyzer (LDA) is the classification model that has been used to classify breast histology images. Although, the performance of LDA classifier is compared with k-NN and SVM classifiers. The highest accuracy for this approach was in LDA classifier outperforms over 80% correct classification rate for the non-malignant Vs malignant breast histopathology images respectively.

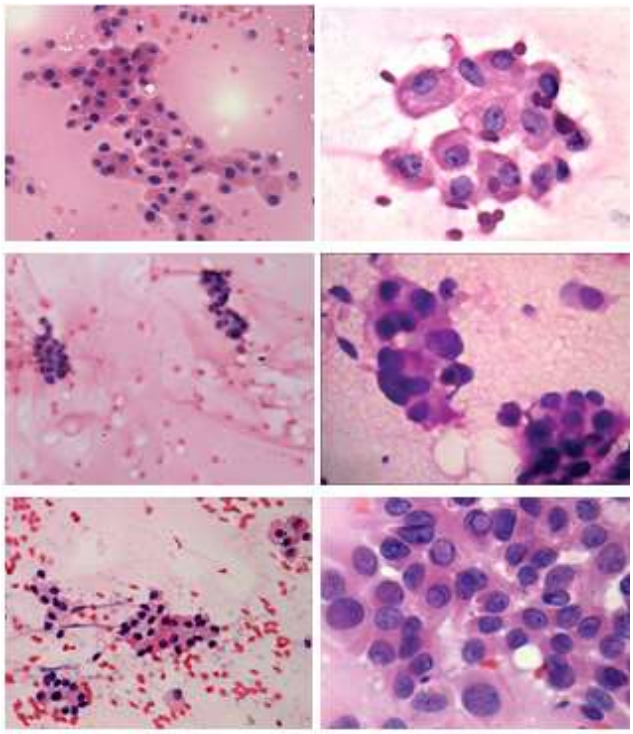


Figure 2: Maligned Breast Cancer image samples (a) is a benign case, (b) a malignant case

Fadzil et al. [20] This approach for Brest cancer prediction and classification approach based on intelligent breast cancer diagnosis using a hybrid genetic algorithm (GA) and artificial neural networks (ANN). This work introduces an automatic breast cancer diagnosis technique using a genetic algorithm (GA). This approach has been designed for simultaneous feature selection approach and parameter optimization of artificial neural networks (ANN). The performances of this work were based on implementation three different variations of the backpropagation technique for the fine tuning of the weight of artificial neural networks (ANN) are compared. The best accuracy of this work was based on an average 99.43% and 98.29% correct classification respectively on the Wisconsin Breast Cancer Dataset.

Niwas et al. [22] 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 the 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. [23] 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

Licensed Under Creative Commons Attribution CC BY

95%.

3. Proposed System

The proposed approach depends on four main steps to predict and classify the breast cancer. The first step (Circularity nucleus Cells Detection): In this step, we proposed and implement the Hough transform to detect just the circular, elliptical or concavity cells in the Brest cancer microscopic tested images. The localization approach based on circle detection using the Hough transform created a new dataset containing (12650) circle cells images automatically cropped from originally slide images, followed by the second step which is the circle cell filtration using a support vector machine. The set of perfect cells images are used to train the (SVM) classifier. This step is designed to retain only those circles that are most likely to represent the nuclei cells. The third step is Feature Extraction: a set of (25) morphometric, textural and topological features were extracted then tested by two different classifiers. final step is the classification of the selected circles. In this step, two different classification approaches have been used. Support vector machine (SVM) and different kernels for data separation are used and tested in the SVM classifier. Although, in this step, we also use a powerful classification framework by proposing a resilient backpropagation Neural Network as a final classification approach. The entire automatic diagnostic procedure was tested on microscopic images of fine needle biopsies images. The code used to process these steps was developed in MATLAB environment. Figure 3 demonstrates the flowchart of our proposed method, and the details of each step are described below.

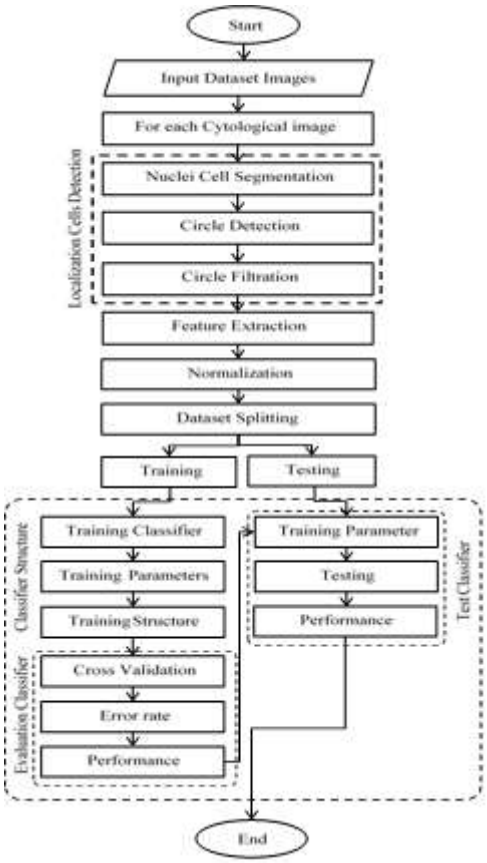


Figure 3: Block digram of the proposed system

3.1 Localization of Nuclei Cells Approach

The Localization approach is proposed to determine and detect nuclei cells. The cells need to be filtered and isolated from the background and from other objects on the tested microscopic slide images. 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 a set of perfect nuclei cells images selected with considering physicians consultant to train the SVM classifier. The whole detected circle cells of the original dataset are automatically classified to correct nuclei cells and incorrect ones. In our case, we set the perfect cells selection to 500 cells images as a training set to train the SVM classifier. By experimental result, we found that number is the perfect one to reach to the higher accuracy of the training and the test of the localization approach. Fig.4. demonstrates the flowchart of the localization method.

A. Circularity Nuclei Cells Detection

CHT is used as a first step on the localization approach. Hough transform is a method that can be applied to detect such a circular shape in a given image [25]. Circular Hough Transform (CHT) was designed to find a circle shape characterized by the center point (x_0, y_0) of the circle in additional to the radius (r). However, Ellipse Hough Transform (EHT) also applied to find the elliptical formations coded by detect the center (x_0, y_0) and the orientation of the ellipse θ of its semi .axes a and b . CHT algorithms is mainly used to detect circles and ellipses which are computationally more expensive than line detection algorithms in any tested image according to the 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 transfer 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) [25]:

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].

$$x = a + r\cos(q) \quad (2)$$

$$y = b + r\sin(q) \quad (3)$$

The first step of the preprocessing the dataset before localization algorithm is enhancing the contrast of images by transforming the values in the intensity image (I) using Contrast-limited Adaptive Histogram Equalization (CLAHE)

Licensed Under Creative Commons Attribution CC BY

[26] Unlike Histogram, it operates on small data regions (tiles), rather than the entire image. Each tile's contrast is enhanced, so that the histogram of the output region approximately matches the specified histogram. In the processing, we use only the red channel where the difference in values between nuclei and red blood cells is the greatest. Also, the cytoplasm, which surrounds the nuclei, is barely visible. After choosing the optimal and an appropriate channel (Red) which the nuclei cells are more visible in. In our implementation of the Circular Hough Transform (CHT) the detection approach searches for circle radius r in the range $1\,\mu m \leq r \leq 8$ where μm is the step size. In pixels, this range corresponds to values $8 \leq r \leq 30$. Some example results of the circle detection approach for cell detection and selection in the localization process is shown in Figure.5. The segmentation of cells images based on discontinuity (Edge-based). Edge based segmentation methods detect edges and produce binary images included edges and their background as the output. we detect edges in the microscopic images by using Canny edge detector which considered as one of the best edge detectors currently in use, Cranny's edge detector guarantees large noise immunity and at the same time ensure detects true edge with least error. A thresholding on the gradient magnitude is performed before the voting process of the Circular Hough transform to remove the 'uniform intensity' (sort-of) image background from the voting process. In other words, pixels that have gradient magnitudes smallest than thresholding value aren't considered in the computation.

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. Circular Cells Filtration on and Isolation using SVM

To select the perfect nuclei cells we proposed performing. The circle cell filtration and isolation. In this case, we need to remove the nonnuclear objects that we have detected during the cells detection step which Formed as a result of the following factors: The nuclear cells sizes are relatively have a high variation. So, seldom a circle which has been detected by using the Circular Hough Transform (CHT) contains more than one nuclei overlapped together. Other objects also exist in the images such as red blood cells. Although they are have much illumination in the red channel of the slide images than the nuclei, they are sometimes detected by the circular Hough transform (CHT). In addition to the false positives generated as a result of some cases, for example, structure composition in the background being inaccurately recognized as the boundary of a cell's nucleus. Those cases are red blood cells and cytoplasm, etc.

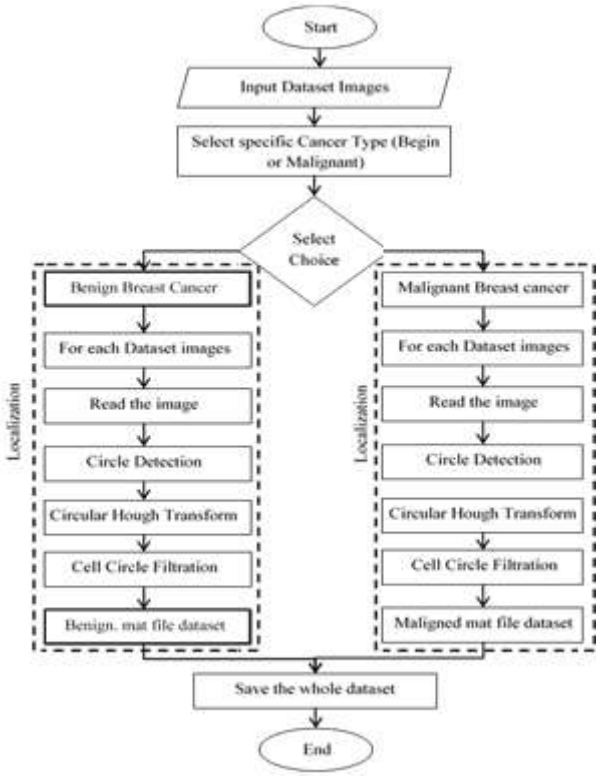


Figure 4: Localization neculei cells detection and filteration approach

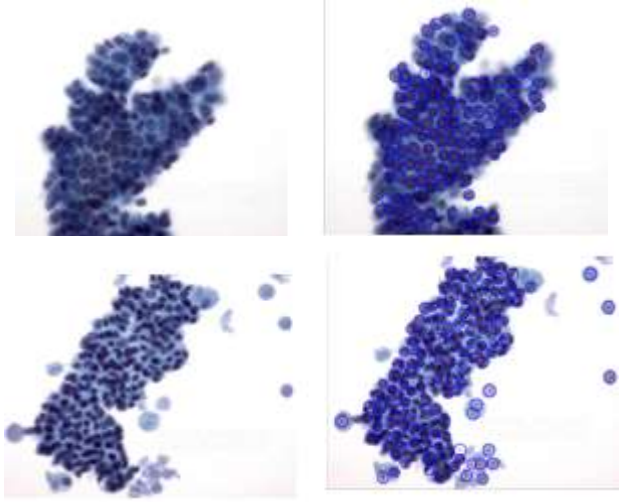


Figure 5: Another example of cell detection using (CHT) Circular Hough transform (a) is the original image, (b) is the CHT image

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 calculating three features as the following:

1) The value of mean pixels inside the localized circle in the blue channel.
2) Long run of high gray-level.
3) The percentage of localized nuclei pixels according to the nuclei mask.

Licensed Under Creative Commons Attribution CC BY

The entire three feature that have been used in these steps which are: first the mean value of pixels inside the circle in the blue channel in the tested image. To make sure that we chose only the infected (cancerous) cells because it obtain the blue dye, more than the normal cell that is not infected because of staining processes of the sections with H&E. Hematoxylin binds to DNA and thereby dyes the nuclei blue/purple, and eosin binds to proteins and dyes other structures (cytoplasm, stroma, etc.) pink. Second; A long run high gray-level emphasis is specified by using gray-level run length matrix. Then, in order to obtain the percentage of infected cells according to all slide image objects. We compute the percentage of detected nuclei's cropped image pixels in the blue channel according to the nuclei mask. Finally, the nuclei mask which is obtained by conducting Otsu's thresholding on the red channel of the image. Which we get binary image contain all objects including nuclear cells, red blood cells cytoplasm in a dark pixel. which means taking the percentage of the pixels in detected circle cells in the blue channel and divide it on the wholly slide image that the detected image cropped from. Containing all dark pixel produced as a result of applying Otsu's thresholding in the red channel. Otsu's method is aimed to finding the optimal value for the global threshold. This method relies on measure of the region homogeneity which is the variance. In other 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 [27] has been proposed in this step as a global and primal thresholding. As illustrated in Figure (6)

Binary Mask: The result of the binary mask is obtained as where the dark objects like nuclei are zeros while the bright background pixels are ones. The final value of the feature is computed by equation (4).

$$PNM = \frac{n_{mask}}{n_{all}} \quad (4)$$

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. If the value of this feature came 0 that's mean is not infected cells and removed else it considers as infected ones.

3.2 Cells Filtration using Support Vector Machine

In this step, the whole (12650) dataset that we have extracted through the localization approach using the (CHT) is used as a testing set after we trained the SVM classifier on the (500 perfect cells) as the training set. The main flowchart of the circle cell filtration is illustrated in Fig. 7. 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. (5) subject to Eq. (6) and Eq. (7).

$$\min \frac{1}{2}|w|^2 + C\sum_i \xi_i \quad (5)$$

$$s.t.\ \ y^{(i)}\left(w^T x^{(i)} + b\right) \geq 1 - \xi_i \quad (6)$$

$$\xi_i \geq 0\ \forall_i \quad (7)$$

The final Lagrangian dual formulation becomes Eq. (8), Eq. (9) and Eq. (10).

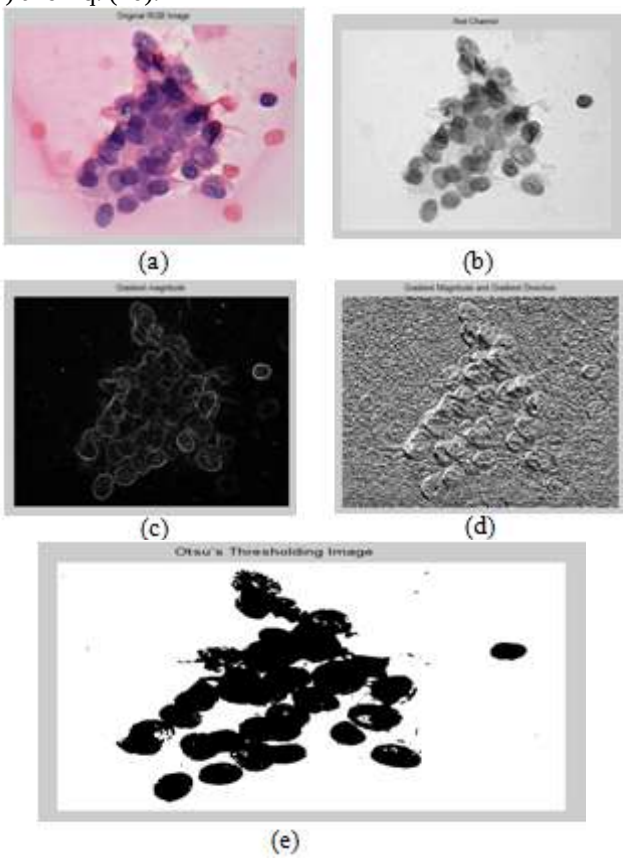


Figure 6: An examples of edge detection required for the circular Hough transform (a) is the original image, (b) is Red channel selection, (c) is the image gradient, (d) is Gradient Magnitude and Gradient Direction, (e) otsu thresholding

$$\underset{\alpha \geq 0}{\text{MAX}}\, \mathcal{L}(\alpha) = \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y^{(i)} y^{(j)} x^{(i)^T} x^{(j)} \quad (8)$$

$$s.t \sum_i \alpha_i y^{(i)} = 0 \quad (9)$$

$$0 \leq \alpha_j \leq C\ \forall_i \quad (10)$$

Now α_i's upper bound is C and the solution is Eq. (11).

$$w = \sum_{i \in Ns} \alpha_i y_j x^{(i)} \quad (11)$$

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. (12) are the Linear Eq. (13), Gaussian Radial Basis Function Eq. (14), and Polynomial Kernels Eq. (15).

Linear:

$$G(x_i, x_i) = x_i^T x_i \quad (12)$$

Gaussian RBF:

$$G(x_i, x_i) = e^{-\|x_i - x_j\|} \quad (13)$$

Polynomial:

$$G(x_i, x_i) = \left(1 + x_i^T x_j\right)^2 \quad (14)$$

The Gaussian Radial Basis Function (RBF) tuning as it given in Eq. (15) Can be achieved by scaling the input vectors by a

Licensed Under Creative Commons Attribution CC BY

scalar value σ before the kernel transformation Eq. (16), resulting in Eq. (17). Alternatively, Matlab can automatically select the optimal scaling via heuristic procedure using subsampling

$$K(x^{(i)}, x^{(j)}) = \left(1 + x^{(i)^T} x^{(j)}\right)^t \quad (15)$$

$$x' = \frac{x}{\sigma} \quad (16)$$

$$G(x_i, x_i) = e^{\frac{-\|x_i - x_j\|}{\sigma^2}} \quad (17)$$

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 manually prepared database of 1000 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. Circle cell filtration approach using support vector machine (SVM)is described and illustrate in the next Fig. 7.

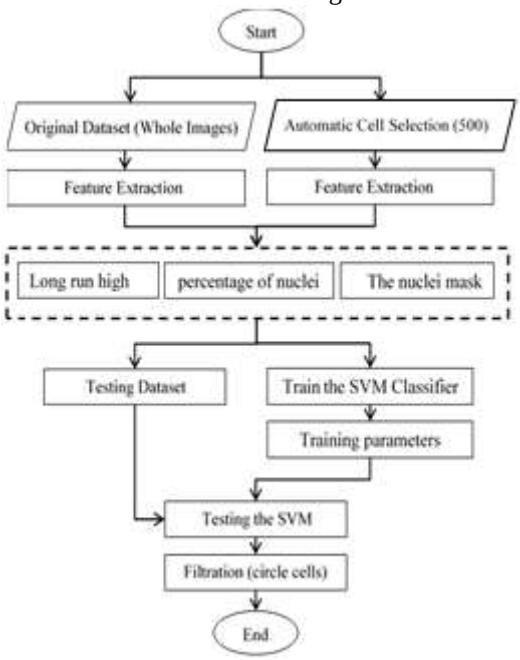


Figure 7: SVM neclei cells filteration approach

3.3 Feature Extraction Approach

After the isolation of nuclei from the images, as determined by the localized circular shape classified as correct in the previous step, 50 global features are extracted and used in the classification procedure. First, for each nucleus we calculate all 25 features, described below.

Note that each nucleus is represented features calculations by the pixels in the interior and on the boundary of the circle that determined it. For each of the 25 features, the mean and variance are calculated giving a total of 50 global features.

In our approach, the features chosen reflect the observations of cytologists. Below a detailed description of all the features that have been used in our proposed system:

1) Area feature value this feature presents the total number of pixels that the nuclei cell has.
2) Perimeter feature value: This feature presents the actual measurement of the nuclei shape by testing the nuclei cell border shape
3) Distance to centroid of all nuclei feature value: This feature presents the gematrical distance between the nuclei cell center and the all pixels (area) around the boundary
4) The Distance to c-nearest nuclei (distance to c-NN) feature this feature presents the total distance of the geometrical actual distance and the nearest nuclei cell
5) Mean R value feature: This feature presents the mean (sum of the color pixels divided by the total number of the pixels) for the RED channel color.
6) Mean G feature value: This feature presents the mean (sum of the color pixels divided by the total number of the pixels) for the GREEN channel color.
7) Mean B feature value: This feature presents the mean (sum of the color pixels divided by the total number of the pixels) for the BLUE channel color
8) Variance of R value: The feature presents the actual variance of pixel values of the nucleus in channel RED.
9) Variance of G value: The feature presents the actual variance of pixel values of the nucleus in channel GREEN.
10) Variance of B value: The feature presents the actual variance of pixel values of the nucleus in channel BLUE.
11) The next feature values shows the four textural features values that are co-occurrence features which is proposed by [28] and has been calculated by using the GLCMs relies on different values such as 0, 45, 90, and 135, and eight gray-levels.
12) Contrast feature: This feature presents the intensity level deference's (contrast) between each pixel and its neighbor in the tested area
13) Correlation feature: this feature presents each pixel that it is correlated with its neighbors in the tested area.
14) Energy feature: this feature presents the sum (total) square of the GLCM features elements.
15) Homogeneity feature: this feature presents the distribution value of the feature elements in the GLCM and the CLCM diagonal elements.
16) Next features values is the last eleven textural features are run-length features [29] [30] calculated using GLRLMs for 0, 45, 90, and 135, and eight gray-levels:
17) The SRE Feature: This feature presents the Short Run Emphasis feature value which is presented by (SRE).
18) The LRE Feature: This feature presents the Long Run Emphasis feature which is presented by the (LRE).
19) The GLN feature: This feature presents the Gray Level Nonuniformity which is presented by the (GLN).
20) The RLN feature: This feature presents the Run Length Nonuniformity which is presented by the (RLN).
21) The RP feature: This feature presents the Run Percentage which is presented by the (RP).
22) The LGRE feature: This feature presents the Low Gray Level Run Emphasis which is presented by the (LGRE).
23) The HGRE feature: This feature presents the High Gray Level Run Emphasis which is presented by the (GLRE).

Licensed Under Creative Commons Attribution CC BY

24) The SRLGE feature: This feature presents the Short Run Low Gray Level Emphasis which is presented by the (SRLGE).
25) The SRHGE feature: This feature presents the Short Run High Gray Level Emphasis which is presented by the (SRHGE).
26) The LRLGE feature: This feature presents the Long Run Low Gray Level Emphasis which is presented by the (LRLGE).
27) The LRHGE feature: This feature presents the Long Run High Gray Level Emphasis which is presented by the (LRHGE). Because the classification is related to whole images and not individual nuclei, each image mean and variance of each feature are computed .All of the image features are then standardized to be 50 features.

3.4 Feature Selection Approach

The initial set of candidate features presented in the previous section is relatively large store after we used the mutual information on the whole dataset. In order to improve the accuracy, we find the optimal subset of features for each classifier (SVM and BNN) by applying the Mutual Information (MI) [31] approach. In this method, Estimates the mutual information between features and associated class labels using a quantized feature space. The procedure is repeated until the addition of any feature does not improve the effectiveness. The total feature score that have been tested by using mutual information are 50 features. And the total features that have been selected to the final classification approach are 27 features.

3.5. Final Prediction and Classification

The classification is performed with two different classification approaches: SVM and three kernels were used (polynomial, Gaussian linear) and Resilient Backpropagation NN with Multi-Hidden Layer [32] [33] [34]. Our design of neural network for breast cancer prediction and classification approach has an input layer, thee hidden layer, and one output layer. Z-score normalization is used to normalize the attribute of the features vector. Classification algorithm performance was measured with the n-fold cross-validation technique [36].

4. Experimental Results

To assess the validation and accuracy of the fully automatic breast cancer diagnosis-based localization approach for nuclei cells selection and filtration with the SVM. A confusion matrix framework is defined as an $m \times m$ 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 benign or malignant case. In this case, the confusion matrix contains information about actual and predicted classifications which is done by the SVM and resilient backpropagation neural network. 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.

4.1 Performance Evaluation Measures

Performance of nucleus detection was assessed in terms of accuracy− Eq. (18), precision− Eq. (19), sensitivity− Eq. (20), and specificity− Eq. (21). , 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.

$$Accuracy = \frac{TP + FN}{TP + FP + FN + TN} * 100 \quad (18)$$

$$Precision = \frac{TP}{TP + FP} \quad (19)$$

$$Sensivity = \frac{TP}{TP + FN} \quad (20)$$

$$Specificity = \frac{TN}{TN + FP} \quad (21)$$

4.2 Localization Approach Experiential Results

During the localization approach, we used the whole dataset which is (130 images). Each image gives different number of cell detection after applying the circular Hough Transform (CHT). The total numbers of the cell images that have been detected during the localization approach are shown in the next Table (1).

Table 1: Localization Approach Experimental Results

Total No Dataset		Localization Approach	
130		Cell Detection	Cell Filtration
Benign	Malignant	12604 cell images	580 cell images
65 cell images	65 cell images		

The localization approach has three stages. The first one is the cell detection which is the total number of the images (130 benign and malignant images). The dataset has 65 images for benign and 65 images for malignant.

4.3 Training and Testing approach

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

4.3 SVM cell filtration result

Table (2) Results of filtration accuracy as correct or incorrect using Support Vector Machine (SVM) on database of 500 manually selected circles

Filtration Accuracy		Sensitivity	Specificity
Training	Testing	0.91	0.95
0.9698	0.9193		

4.4 SVM Prediction and Classification Results

Figure (8) illustrate the performance result of our proposed

Licensed Under Creative Commons Attribution CC BY

system in breast cancer prediction and classification approach. Two different approaches have been used and discussed. The first one is using the whole feature space without suing feature selection. The second one is using mutual information to select the highest feature score which is in our case (27) features.



Figure 8: Overall SVM classifier performance results using the whole feature space once and with feature selection approach

Table (3) shows the classification results of the final classification results relies on using the SVM approach. The highest accuracy was (94.1%) in the training set and (97.0%) in the testing set

Table 3: overall SVM classifier performance results

Approach	Kernel	Training	Testing
Using the Whole (50) Feature Space	Linear	74.1%	64.0%
	Gaussian	84.0%	94.0%
	Polynomial	91.5%	95.3%
With Feature Selection Approach	Linear	90.0%	97.0%
	Gaussian	91.0%	94.0%
	Polynomial	94.1%	97.0%

Table (4) shows the classification results of backpropagation Neural Network (BNN) prediction and classification results using the whole feature space (50 features). The highest accuracy was (99.88%) in the training set and (99.15%) in the testing set.

Table 4: BNN classifier results using feature selection approach

Training /Cross-Validation			
Experimental Results			Training Accuracy
Confusion Matrix	Performance Results		
99.980 0.01945 0.2195 99.7805	Sensitivity	0.9978	0.9988
	Specificity	0.9998	
	Precision	0.9998	
	Accuracy	0.9988	
Testing Dataset			
Experimental Results			Testing Accuracy
Confusion Matrix	Performance Results		
99.248 0.7517 0.9517 99.048	Sensitivity	0.9905	0.9915
	Specificity	0.9925	
	Precision	0.9925	
	Accuracy	0.9915	

Figure (9) shows the performance of the Backpropagation Neural Network (BNN) prediction and classification results on the training set using 5-folds cross validation approach. The feature selection approach by selecting (27 features only) has been applied to train the network.

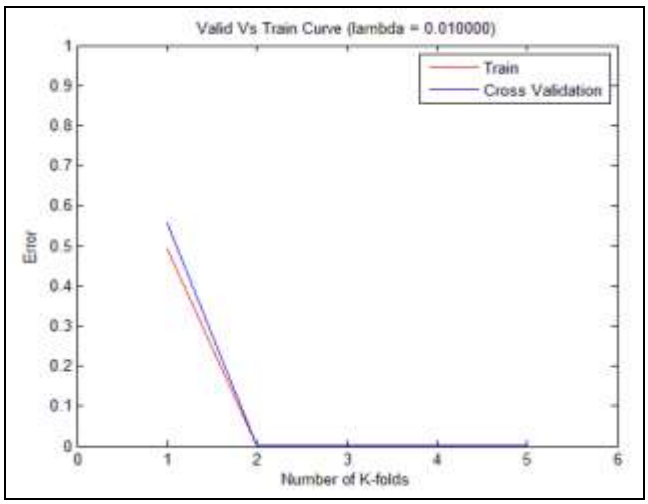


Figure 9: BNN performance results using 5-folds cross validation approach

4.5 Comparison between BNN and SVM classification results

In term of comparing the classification results for our proposed system which is breast cancer detection and classification approach using two different algorithms Backpropagation Neural Network (BNN) and Support Vector Machine (SVM).But by using the feature selection approach with (27 features) that have been selected according to the mutual information scoring. Figure (10) shows the performance results of SVM and BNN depending on using (26 features). In this comparison results, we can clearly see that the highest accuracy between those two algorithms in the training set was on the BNN classifier by (99.88%), (94.19%). Although, in the testing set the highest accuracy was (99.15%) using BNN classifier, but the highest accuracy in SVM classifier using polynomial kernel was (96.88%).

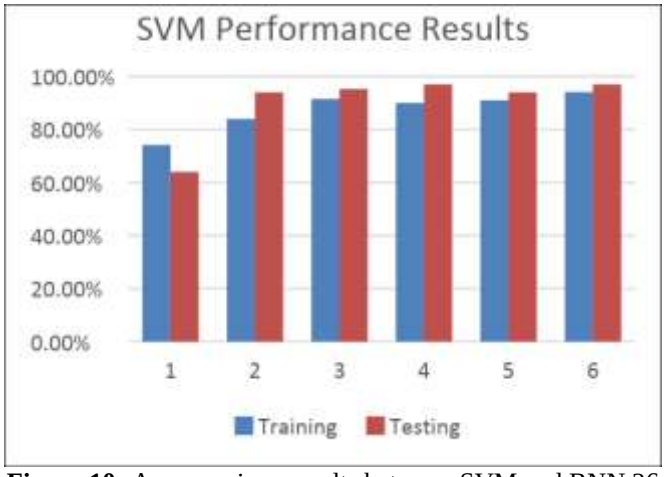


Figure 10: A comparison results between SVM and BNN 26 features by feature selection approach

5. Conclusion

In this paper, we design and implementation a new approach for breast cancer diagnosis-based computer-aided system. The goal of this system is to fully automatically distinguish

Licensed Under Creative Commons Attribution CC BY

between the breast cancer types which are benign and malignant cases. The whole approach is based on the analysis of cytological images of FNB images. Depending on the fact that most of the new method of segmentation do not work properly on the new very high resolution and very complicated texture images, in this paper we depend on using a fast and accurate segmentation technique. This technique relies on an prediction of cell nucleus by circularity texture detection and size estimation. By using that we depend on the localization approach for estimation and detection the nuclei cell images for each breast cancer test image by using an accurate circle detection approach. For this goal, the Circular Hough transform (CHT) is proposed and used to detect circle cell. Filtration and isolation is proposed to remove of incorrect or less reliable detections using a supervised technique by SVM classifier. The high-quality isolated nuclei are the main solution and proposed techniques that we rely on to remove the false positive detection like (blood cells and cytoplasm) from consideration inside the tested images. Although, it allows extraction of features were extracted from high-quality isolated nuclei inside the tested images. By adding those two approaches together which are the proposed features and classifiers gave us very good results in distinguish between the two types of the breast cancer (Benign and Malignant). The best that we have obtained effectiveness was reached to (99.15%) using Backpropagation Neural Network (BNN) including subset of features that have been selected depending on the Mutual score feature selection approach.
Finally, indicating our results which present that our proposed system for breast cancer prediction and classification approach is effective and capable of providing valuable diagnostic information between those two types of the breast cancer cases (Benign and Malignant).

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] A. D. Belsare, M. M. Mushrif , M. A. Pangarkar and N. Meshram “Classification of breast cancer histopathology images using texture feature analysis”, TENCON 2015 - 2015 IEEE Region 10 Conference, 1-4 Nov 2015.
[20] Fadzil Ahmad, Shah Alam, Malaysia , Nor Ashidi Mat Isa , Mohd Halim Mohd Noor and Zakaria Hussain,” Intelligent Breast Cancer Diagnosis Using Hybrid GA-ANN”, Fac. of Electr. Eng., Univ. Teknol. MARA, Computational Intelligence, Communication Systems and Networks (CICSyN), 2013 Fifth International Conference.
[21] 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.
[22] 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.
[23] J. Malek, A. Sebri, S. Mabrouk, K. Torki, and R. Tourki, “Automated breast cancer diagnosis based on GVF-Snake segmentation, wavelet features extraction

Licensed Under Creative Commons Attribution CC BY

and fuzzy classification," J. Signal Process. Syst., vol. 55, pp. 49–66, 2009.
[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] 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.
[26] Karel Zuiderveld, "Contrast Limited Adaptive Histogram Equalization",Graphics Gems IV, p. 474-485, code: p. 479-484
[27] T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," Communication Technology, IEEE Transactions on, vol. 15, pp. 52-60, 1967.
[28] HARALICK R., SHANMUGAM K., DINSTEIN I., Textural features for image classification, IEEE Trans. on Systems, Man and Cybernetics, 1973, Vol. 3, No. 6, pp. 610–621.
[29] M. Galloway, "Texture analysis using grey level run lengths," Comput. Graph. Image Process., vol. 4, pp. 172–179, 1975.
[30] B. Dasarathy and E. Holder, "Image characterizations based on joint gray level-run length distributions," Pattern Recognit. Lett., vol. 12, no. 8, pp. 497–502, 1991.
[31] J. Pohjalainen, O. Rasanen & S. Kadioglu: "Feature Selection Methods and Their Combinations in High-Dimensional Classification of Speaker Likability, Intelligibility and Personality Traits", Computer Speech and Language, 2015, for more details.
[32] L. Belanche. Heterogeneous Neural Networks. PhD thesis, Technical University of Catalonia, 2000.
[33] J. S´ıma and P. Orponen. General-purpose computation with neural networks: A survey of complexity theoretic results. Neural Computation, 15:2727–2778, 2013.
[34] Koudelka, V., Svobodova, J., Raida, Z.: 'Impedance Network Simplification: A Combinatorial Optimization Approach', In Proceedings of ICEAA conference on Electromagnetics in Advanced Applications, Torino-Italy, September 2014.
[35] Koudelka, V., Raida, Z., Tobola, P.:'Simple Electromagnetic Modelling of Small Airplanes: Neural Network Approach', Radio Eng., vol. 18, pp. 38-41, 2009.
[36] Wen Zhu, Nancy Zeng, Ning Wang," Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS® Implementations", 1K&L consulting services, Inc, Fort Washington, PA 2Octagon Research Solutions, Wayne, PA.

Author Profile



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 in health government programs system including remote medical consultant , medicine control system, medical devises 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.

Licensed Under Creative Commons Attribution CC BY