Classification the Mammograms Based on Hybrid Features Extraction Techniques Using Multilayer Perceptron Classifier

Hayder A. Saleh1*, Enaas M. Hussain1, Enam A. Khalel2

1Department of Computer Science, Mustansiriya University, Baghdad, IRAQ
2Specialist Radiology Department-Oncology Teaching Hospital, Medical City Complex, Baghdad, IRAQ

*Contact email: haider8883@gmail.com

ABSTRACT
Cancer of the breast is one of the world’s most prevalent causes of death for women. Early and efficient identification is important for can care choices and reducing mortality. Mammography is the most effective early breast cancer detection process. Radiologists cannot however make a detailed and reliable assessment of mammograms due to fatigue or poor image quality. The main aim of this work is to establish a new approach to help radiologists identify anomalies and improve diagnostic precision. The proposed method has been applied through the implementation of preprocessing then segmentation of the images to get the region of interest that was used to find a texture features that were calculated based on first Order (statistical features), Gray-Level Co-Occurrence Matrix (GLCM), and Local Binary Patterns LBP (LBP). In the features selection phase mutual information (MI) algorithm is applied to choose from the extracted features collection suitable features. Finally, Multilayer Perceptron has been applied in two stages to classify the mammography images first to normal or abnormal, and secondly, classification of abnormal images into benign or malignant images. This method was implemented and gave an accuracy of 92.91 % for the first level and 93.15% for the second level classification.

KEYWORDS: Breast cancer, Digital mammography, Texture feature, Mutual information (MI), Classification, MLP classifier.

INTRODUCTION
Breast cancerous is the most epidemic disease in women’s lives and is the second woman cause of death between them [1]. Based on the most recent preliminary data, breast cancer accounts for 29% of new cancerous states and 15% of cancerous deaths worldwide in the USA. According to the complexity of breast cancer, the probability of developing cancer at the individual level is very difficult to predict [2]. soon and efficient detection is the best ways to minimize mortality improve the care choice, where in the recent years there are many screening methods are available for breast cancer diagnosis, as the biopsy test, Magnetic Resonance Imaging (MRI), Ultrasoundography, mammography, etc.. the biopsy is taken of the tissues of the breast, this test gives higher
accuracy results but this test is very painful and pathetic. Therefore, most patients avoid this test [3]. Currently, mammography-depend screening programs are widely considered one of the safest and most common approaches for breast damage detection. Mammography is a low-dose of an X-ray technique for visualizing the inner breast structure [4]. Mammogram images usually contain many things and noises that make medical images very difficult to diagnose and understand cancer in their soon stages. Therefore, uniform image quality and ROI extraction are primary to restrict the search for anomalies [5]. Visual analysis of the radiologist’s mammograms can diagnose extremely high-level breast cancer, however often contributes to the loss of those features and a lower diagnosis due to radiologist stress and poor picture quality. Studies show that a radiologist who uses visual inspection has a 10 % to 30% false rate to diagnoses of malignant masses. The false rate can be greatly decreased by using the classification methodology.

The classification model for breast cancer screening programs endorses the final decision and is a second opinion on the diagnosis by the radiologist in classifying breast tumors [6]. There are usually five steps in a mammogram (CAD) system to detect breast cancerous [7].

This paper proposed a classification model to support human observations, which can sometimes be subjected to errors and inaccurate diagnosis due to the fatigue of doctor and diagnostic experts, poor image quality, and a great consistency of image texture, particularly in fatty and glandular forms. The proposed structure consists of five main steps: image preprocessing, image segmentation, extraction, and selection features, and finally classification. For image pre-processing, medium filter, and binary image with a global threshold, they were applied to remove noise and small artifacts if any. In the segmentation phase, a Hybrid Bounding Box and Region Growing (HBBRG) algorithm are utilizing to remove the pectoral muscles [8]. Also, a geometric method has been applied to cut the largest possible square that can be obtained from a mammogram which represents the ROI.

The output of the segmentation phase is used at an efficient feature extraction phase based on these techniques to set up strong texture features. In this paper, firstly first order (statistic) features have been calculated for ROI, it is a set of useful features that can be directly extracted from the spatial domain of the image histogram based on pixel values only [9] these features are mean is the average value of the intensity of the image. Standard deviation measures the difference between a set of data and the amount of dispersion between it the variance tells the intensity variation about the mean, Skewness tells the symmetries of the histogram about the mean, the kurtosis is the flatness of the histogram [10].

Secondly, LBP features are calculated for ROI, it is texture descriptor is considered very interestingly because particularly suitable for real-time quality controlling applications because it is fast and easy to implement [11]. LBP indicates a cognition among the middle pixel and its adjacent in the given block. LBP checked window into cells for example 8x8 for, where compares the pixel to each of its 8 neighbors. It follows the pixels along a circle, i.e. clockwise or counterclockwise. Figure (1) shows an example of a micro pattern and pattern and matching LBP to it.

The histograms are created of the decimal values of all the image blocks. This histogram can be seen as a 256-dimensional feature vector, the length of the feature vector for a single cell reduces from 256 to 59 by optionally normalize the histogram [12].

Thirdly, GLCM features are extracted is also a statistical method to take into account the spatial relationship between pixels. It has been employed widely in numerous applications based on texture analysis in the area of texture analysis [13]. The probability of a pixel with a gray level of I in terms of a certain spatial relationship with a value of j can always be computed with GLCM, and the calculation of the GLCM functional descriptor is performed in angles of a and d. [14]. There are four features of the GLCM are extracted contrast, correlation, energy, and homogeneity. Also, all of
these features passed through the feature selection phase, it is a very important step before the classification phase because irrelevant, outlier, noisy, and redundant features often degrade the performance of classification algorithms through the time consumed and prediction accuracy. Feature selection methods attempt to find minimize features vector, which minimizes the probability of error and maintains the powerful features that provide high-accuracy diagnostic results in the classification stage [13]. In this thesis, the feature selection stage was implemented using the MIFS function, a function that takes into account the interaction between the feature set and the target [15].

Finally, a multilayer perceptron classifier (MLP) was used to classify the image firstly as normal or abnormal image and then categories the abnormal ones into benign or malignant.

RELATED WORK

Many types of research have been done earlier for breast cancer diagnosis, the intelligent techniques are used in most of these researches to attempt to assist radiologists in the detection of abnormal tissue in mammograms image to detect the presence of breast cancer. We will review some of the most important research in this domain as follows:

• In 2015 [16]. They proposed a new two approaches to diagnose abnormalities in mammograms. In the first approach, the automated RG was used to segment the mammogram, while the second approach used a cellular neural network (CNN) to segment the mammogram. In the features extraction phase textural, and shape features are extracted for a segmented tumor. A genetic algorithm is utilized to choose suitable features from the set of extracted features. Finally, ANNs are used for classification the mammograms to benign and malignant. The system has been tested with DDSM and MIAS databases and the results for accuracy, specificity, and sensitivity was 96.47%, 95.94%, and 96.87%, respectively.

• In 2016 [17]. The researchers have proposed system formations of four phases. In the image preprocessing phase laplacian operator, and mean filtering are utilizing to smooth the image and reduction the noise if any to improve the accuracy. Then the output of the preprocessing phase is used for the segmentation phase, which is done using the RG technique. Then for improved detection accuracy, a hybrid feature extraction approach was used by 2D-DWT features, gradient features, and texture features. Finally, a well-organized Feedforward neural network (FFNN) is used for classification. The results were shown to achieve around 91% accuracy obtained.

• In 2018 [18]. In this research, the median filter is utilizing for noise filtering, and global thresholding is using for label removal in the preprocessing stage, and then the bounding box (BB) was used to remove the pectoral muscles. The adaptive fuzzy logic depends on bi-histogram equalization is an efficient algorithm proposed to get better perception by improving the quality of mammograms. ROI dynamically select and separate of mammograms using global thresholds and morphological operations. In the features extraction stage, shape and GLCM features are collected for an ROI, and then ideal features are chosen utilizing the classifier and regression tree (CART). The classification step is finally carried out using backpropagation with Feedforward Artificial Neural Networks.

• In 2019 [19]. The researchers utilized few features than other previous research that used many features sets, many techniques have been used to reduce dimensions. The (KNN) and (ANN) classifiers are used to classify these few features. 50 cases of the 'BAHEYA Foundation to Early Detection and Treatment of Breast Cancer by doctors and radiologists in the hospital have been utilized for the proposed system. The images used are Contrast-Enhanced Spectral Mammograms (CESMs) that have clearer and more contrasting images than typical mammals. The KNN and ANN classifiers were used and the outcomes indicate to achieve accuracy percent with 92 percent with ANN.

• In 2019 [20]. the authors have proposed a system for detect potential cancer tumors in mammograms, the detection is made through automatically dividing breast images by combining a hybrid density slicing technique with the adaptive k-means algorithm, also by dividing breast images and extracting areas of cancer, then
calculating their properties using the method of the region growing. Gray-level co-occurrence matrices (GLCM) have been used with proposed features that are gray level density matrices (GLDM) to detect abnormal tissue (malignant) or normal tissue (benign) using MLP classifier. Experimental results showed achieved 91.17% accuracy.

Dataset Used
To conduct this study, investigating it all its phases, a special database was established that relies mainly on the Mammogram image analysis Society (MIAS). The database was formed by The United Kingdom national breast screening program and getting from a film-screen mammographic. The database consists of 322 images 208 normal and 114 abnormal, which subdivide to 63 Benign and 51 malignant. The image of the size is 1024 x 1024 pixels and format "PGM" [21]. The database is available on the website http://peipa.essex.ac.uk/info/mias.html [22]. Also, in coordination with the Teaching Oncology Hospital / Medical City / Baghdad, a set of images was obtained and added to the database. These pictures were diagnosed with the help of Dr. “Inaam Khalil Aziz”, a breast cancer specialist and director of the mammography department at the hospital. The pre-processing of these images has also been performed to be in the same MIAS database image specification.

MATERIALS AND METHODOLOGIES
The main aim of this work is to design a computer-assisted classification model that can classify the mammogram image and it helps physicians and diagnostic experts by providing a second diagnostic view for a more reliable diagnostic decision. Figure (2) displays the block diagram of the stages of implementation of the proposed system and the algorithms used to achieve this.

In one example, this process continues on the mask until a value of zero is reached. A square with the co-ordinates from the smallest pixel to the biggest value is drawn. Finally, all columns and rows with the total number equal to zero are omitted to eliminate the black background Figure (5) indicates the steps in which these steps are carried out.

![Figure 2. Block diagram of the proposed method.](image-url)
B. Segmentation

Many segmentation algorithms have been used for many applications in medical images. In this work, this stage was applied in two phases to remove pectoral muscles, and cutting the largest possible square to get the ROI mammogram image.

In the first phase For the removal of pectoral muscles hybrid bounding box and region growing (HBBRG) algorithm was proposed to achieve better results and minimize defects in each algorithm. Based on the fact that the muscles of the pectoral area that are approximately Trenchard and appear in the upper left or upper right corner of the breast contour depending on whether the breast is left or right, the region's rising algorithm is applied to extract the piece of the pectoral muscle. Furthermore, the function must locate the maximum distance between the seed point and the adjacent pixel pixels. Figure (4) displays the results of the implementation of the HBBRG algorithm.

In step two the breast image was segmented to get the maximum possible square region from the image of the mammography. As this area is square, it provides ease and speed when applied in the feature extraction stage. This process has been implemented by a geometrical segmentation method by converting the mammographic image into the binary form, then finding the mask which contains only the breast image with its one's value, and then performing the reverse search process that starts with the penultimate pixel of the mask at the lower right corner. Compared to the right, bottom and diagonal of its three neighbors, the least value is found among them and then increases its value. In one example, this process continues on the mask until a value of zero is reached. A square with the co-ordinates from the smallest pixel to the biggest value is drawn. Finally, all columns and rows with the total number equal to zero are omitted to eliminate the black background Figure (5) indicates the steps in which these steps are carried out.

C. Features Extraction and Selection

In this step, a set of features that describe the tissue properties is assigned to the ROI. These features can be a set of real numbers that allow the normal tissue to varying from abnormal and benign tissue from malignant tissue. In this paper, to design the classifier model, mammograms of
high quality can be diagnosed and powerful texture features created. A features vector consists of 73 features determined for the ROI using three techniques. Firstly ten features of the first-order (statistical) features are (mean (M) and standard deviation (SD) for the mean, M and SD of standard deviation, M and SD of variance, M and SD of skewness, M and SD of kurtosis) Secondly, the fifty-nine feature of the LBP technique is calculated for the region of interest and counted from feature eleven to feature sixty-nine. Thirdly, four features of the GLCM are contrast, correlation, energy, and homogeneity. All of these features were calculated with the aim of creating a powerful texture features. Table (4.3) display values a sample of these features are (M (mean), SD (Standard Deviation), LBP1, LBP59, Contrast, Homogeneity). Finally, the extraction features passed through the phase of the feature, it attempts to find a minimize features vector, which minimizes the probability of error and maintains the powerful features that provide high-accuracy diagnostic results in the classification stage. In this paper, the feature selection stage was implemented using the MIFS function, a function that takes into account the interaction between the feature set and the target. After testing more than one value the number of features was reduced to only 50 features Depending on the weight produced by the algorithm, as this number of features provided the best results in the classification stage. Figure (6) shows analyzing the weight value that was extracted for each feature after applying the MI function.

A feature has been used as input in the MLP neural network and there are connections with the input, hidden, and output layer. paper, the network structure consists of one input layer that contains inputs that correspond to 50 features extracted from the feature selection stage, two hidden layers, the first contains 24 nodes, the second contains 12 nodes, and finally, one output layer contains outputs corresponding to two categories, either natural or unnatural, and benign or malignant according to the level of classification as shown in figure (7).

E. Classification Evaluation
The efficiency of machine learning techniques is assessed by such success assessment metrics. A current and expected class confusion matrix consisting of true positive (TP), false positive (FP), true negative (TN), and false-negative (FN) is created. Positive/negative in this section refers to the decision by the algorithm and true/false refers to the way the decision correlates with the real clinical condition. In wherein the case of the two classes there are only four possible outputs represented elements of the confusion matrix of (2 * 2) for a binary classifier. There are sex statistical
metrics used to evaluate the performance of the proposed system based on the confusion matrix are accuracy (ACC), error rate (ERR), sensitivity (SN), false-positive rate (FPR), specificity (SP), and precision (P).

RESULTS AND DISCUSSION
In this research, the proposed system was applied to all images in the database, where 70% of the image was used for the training phase and 30% testing phase of random instants of image features from the dataset with 100 loop and 20000 iterations for the training phase. The results show that MLP for the first level has achieved the average, best, and worse accuracy they are 82.318%, 92.915%, and 72.725 %, respectively, see figure (8). Also, the results show that the MLP classifier for the second level has achieved the average, best and worse accuracy they are 84.561%, 93.15%, and 73.493 %, respectively, see figure (9).

Table 1. Confusion matrix result of MLP for first-level

| #    | The average results | The best results | The worst results |
|------|---------------------|------------------|------------------|
| ACC  | 82                  | 92.915           | 72.725           |
| ER   | 17.074              | 7.085            | 27.274           |
| SE   | 86.491              | 91.7             | 78.446           |
| FPR  | 18.641              | 5.136            | 29.845           |
| SP   | 81.358              | 94.863           | 70.154           |
| P    | 87.399              | 97.7             | 79.839           |

Table 2. Confusion matrix result of MLP for second-level

| #    | The average results | The best results | The worst results |
|------|---------------------|------------------|------------------|
| ACC  | 84.561              | 93.15            | 73.493           |
| ER   | 15.438              | 6.849            | 12.506           |
| SE   | 85.546              | 98.28            | 67.349           |
| FPR  | 15.96               | 2.761            | 34.95            |
| SP   | 84.039              | 97.238           | 65.584           |
| P    | 87.568              | 96.5             | 63.842           |

Figure 9. displays a performance analysis of the MLP for first level from table 2.

CONCLUSIONS
Mammography is the most powerful tool used to diagnose breast cancer early. Removal of the pectoral muscles represents the biggest obstacle in the treatment of mammograms because they closely resemble tumors and the rest of the breast tissue, especially in the types of fatty and glandular images, as well as their presence greatly affects the results of the following stages such as extracting features and segmentation. The HBBRG algorithm achieved very good results with more than 98% to completely remove the pectoral muscles. A geometric segmentation method was proposed to cut the largest possible square representing a sample of breast tissue, it achieved 100% success with normal images and more than 98% with abnormal images, where the tumor is within the crop area. The proposed texture features are based on three technologies are first-order, LBP, and GLCM, these features give the algorithm more robustness due to its
resistance to many image situation variations that lead to the best discrimination potential for classification the type of images. Finally, there is a set of future works for the development of our paper can be achieved such as model development from the diagnostic model to the diagnostic and prediction model, and testing new segmentation methods that provide better results in discovering and identifying the damage and isolation it from the rest of the breast tissue, especially in fatty and glandular images, where this stage is considered one of the most difficult stages of the CAD system.

REFERENCES

[1] Bray, Freddie, et al. "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries." CA: a cancer journal for clinicians 68.6 (2018): 394-424.

[2] Qiu, Yuchen, et al. "An initial investigation on developing a new method to predict short-term breast cancer risk based on deep learning technology." Medical Imaging 2016: Computer-Aided Diagnosis. Vol. 9785. International Society for Optics and Photonics, 2016.

[3] Wang, Lulu. "Early diagnosis of breast cancer." Sensors 17.7 (2017): 1572.

[4] Maitra, Indra Kanta, Sanjay Nag, and Samir Kumar Bandyopadhyay. "Technique for preprocessing of a digital mammogram." Computer methods and programs in biomedicine 107.2 (2012): 175-188.

[5] Makandar, Aziz, and Bhagirathi Halalli. "Preprocessing of mammography image for early detection of breast cancer." International Journal of Computer Applications 144.3 (2016): 0975-8887.

[6] Berbar, Mohamed A. "Hybrid methods for feature extraction for breast masses classification." Egyptian informatics journal 19.1 (2018): 63-73.

[7] Ramadan, Saleem Z. "Methods Used in Computer-Aided Diagnosis for Breast Cancer Detection Using Mammograms: A Review." Journal of Healthcare Engineering 2020 (2020).

[8] Cheng, Heng-Da, et al. "Automated breast cancer detection and Saeed, Enas Mohammed Hussein, and Hayder Adnan Saleh. "Pectoral Muscles Removal in Mammogram Image by Hybrid Bounding Box and Region Growing Algorithm." 2020 International Conference on Computer Science and Software Engineering (CSASE). IEEE, 2020.

[9] Hussain, Alyaa, Alaa Noori Mazher, and Asraa Razak. "Classification of Breast Tissue for mammograms images using intensity histogram and statistical methods." Iraqi Journal of Science 53.5 (2012): 1092-1096.

[10] Hlaing, K. Nyem Nyem. "First order statistics and GLCM based feature extraction for recognition of Myanmar paper currency." Proceedings of the 31st IIER International Conference, Bangkok, Thailand. 2015.

[11] Harefa, Jeklin, A. Alexander, and M. Pratiwi. "Comparison Porebski, Alice, Nicolas Vandenbroucke, and Ludovic Macaire. "Haralick feature extraction from LBP images for color texture classification." 2008 First Workshops on Image Processing Theory, Tools, and Applications. IEEE, 2008.

[12] Nosaka, Ryuksuke, Yasuhiro Ohkawa, and Kazuhiro Fukui. "Feature extraction based on co-occurrence of adjacent local binary patterns." Pacific-Rim Symposium on Image and Video Technology. Springer, Berlin, Heidelberg, 2011.

[13] Vasanth, M., V. Subbiah Bharathi, and R. Dhamodharan. "Medical image feature, extraction, selection, and classification." International Journal of Engineering Science and Technology 2.6 (2010): 2071-2076.

[14] Al Mutaz, M. Abdalla, Safaa Dress, and Nazar Zaki. "Detection of masses in digital mammogram using second-order statistics and artificial neural network." International Journal of Computer Science & Information Technology (IJCIST) 3.3 (2011): 176-186.

[15] Al-Ani, Ahmed, and Mohamed Deriche. "Feature selection using mutual information based measure." Object recognition supported by user interaction for service robots. Vol. 4. IEEE, 2002.

[16] Rouhi, Rahimeh, et al. "Benign and malignant breast tumors classification based on region growing and CNN segmentation." Expert Systems with Applications 42.3 (2015): 990-1002.

[17] Alamin, Ibrahim Mohamed Jaber, W. Jebserson, and H. K. Bajaj. "Improved framework for breast cancer detection using hybrid feature extraction technique and fmn." International Journal of Advanced Research in Artificial Intelligence 5.8 (2016).

[18] Sheba, K. U., and S. Gladston Raj. "An approach for automatic lesion detection in mammograms." Cogent Engineering 5.1 (2018): 1444320.

[19] Mostafa, Shaimaa, et al. "Breast Cancer Detection Using Polynomial Fitting Applied on Contrast Enhanced Spectral Mammography." 2019 International Conference on Innovative Trends in Computer Engineering (ITCE). IEEE, 2019.

[20] Salman, Nassir H., and Semaa Ibrahim M. Ali. "Breast Cancer Classification as Malignant or Benign based on Texture Features using Multilayer Perceptron." International Journal of Simulation–Systems, Science & Technology 20.1 (2019).

[21] LI, Shenglan, et al. "Performance evaluation of a CAD system for detecting masses on mammograms by using the MIAS database." Medical Imaging and Information Sciences 18.3 (2001): 144-153.

[22] J Suckling et al. "The Mammographic Image Analysis Society Digital Mammogram Database Exerpta Medica." International Congress Series 1069 (1994) pp375-378. Available online: http://peipa.essex.ac.uk/info/mias.html