University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-689

Vol.20 No.1 Mar 2025

Using convolutional neural networks and decision tree for Prediction of Breast Cancer

Salah K. moshi
salah@utq.edu.iq

University of Thi-Qar

https://doi.org/10.32792/utq/utj/vol20/1/3

Abstract

Breast cancer is the leading cause of cancer-related mortality among women. Mammography surveillance plans have been proposed to decrease breast cancer mortality. However, the sensitivity of mammography is suboptimal, especially in women with “dense breasts.” In the analysis of breast images for cancer detection, lesion multiplicity and complexity may occur due to dense tissue interactions, making it challenging for radiologists to precisely detect and analyze masses. The present paper introduces a new computer-aided detection (CAD) system for breast cancer prediction. This system consists of three major steps: i) Segmentation strategies**:** The first strategy involves manually assigning the region of interest (ROI), while the second strategy uses a threshold- and region-based method, ii) Feature extraction**:** Features are extracted using a novel convolutional neural network (CNN), iii) Tumor classification**:** A decision tree is employed for classifying tumors. The performance of the proposed method is evaluated using the Digital Database for Screening Mammography (DDSM) dataset. The simulation results indicate that the proposed method outperforms existing methods, with an accuracy improvement of approximately 1.29%.

Keywords: The convolutional neural network, the decision tree, the computer-aided detection.

1- Introduction

Breast cancer is one of the leading causes of death among women. Mammography is currently one of the most widely used techniques for early breast cancer diagnosis. However, Magnetic Resonance Imaging (MRI) has emerged as a highly promising alternative to mammography. Despite its advantages, MRI requires confirmation of tumor presence by radiologists. One of the drawbacks of MRI is the possibility of allergic reactions to contrast agents or skin infections at the injection site, which may cause panic in patients who feel uncomfortable in enclosed spaces. Additionally, masses and microcalcifications (MCs) are significant indicators of potential breast cancer [1].

Despite these advantages, the manual interpretation of MRI images is time-consuming and relies heavily on the radiologist's experience, making it prone to human error. Furthermore, issues such as allergic reactions to contrast agents and the anxiety associated with confined spaces limit the widespread use of MRI.

Several imaging methods and computer-aided systems have been developed to assist radiologists in this area. These advancements have significantly contributed to the detection and treatment of breast cancer. Among various clinical imaging techniques, MRI has become an essential tool for breast cancer detection due to its high sensitivity and resolution, especially in dense breast tissue. Compared to mammography and ultrasound, MRI is the most sensitive imaging technique for detecting breast cancer in high-risk populations. However, accurate interpretation of breast MRI images depends on the quality of the visualized images, the radiologist's experience, and the time available for data analysis. Due to the time-consuming nature of manual MRI analysis and its susceptibility to human error, many automated systems have been developed to assist radiologists in locating and detecting breast lesions. Nevertheless, while several systems are currently used in clinical practice, fully automated breast lesion detection remains an unresolved issue in MRI [2].

The recent growth in computational power and dataset sizes has made convolutional neural networks (CNNs) a feasible solution for image classification challenges. Unlike traditional manual feature extraction techniques, CNNs learn useful features directly from image data, optimizing the classification loss function. These deep learning (DL) models have proven to be highly effective in various domains, including medical image analysis, especially in the imaging of tissue diseases [3].

The goal of this project is to develop a computer-aided diagnosis (CAD) system that can automatically and effectively identify and classify cancer lesions in MRI images using deep learning algorithms. To identify the target areas, this system utilizes two segmentation methods: one manual and one based on a thresholding technique. The primary objective of this study is to reduce reliance on manual analysis while enhancing the accuracy, speed, and reliability of breast cancer diagnosis.

This study aims to automatically extract important information from MRI images and improve cancer lesion identification by leveraging deep learning techniques, particularly convolutional neural networks.

2- Related work

In this section, we review works conducted in the CNN domain and decision tree for breast cancer prediction.

In [4], researchers proposed a novel Deep Convolutional Neural Network (DCNN) model with a transfer learning strategy to differentiate cancerous mammograms from normal ones. Initially, the authors extracted hand-crafted features, which were then input into a machine learning method for classification. Subsequently, they applied the transfer learning approach to the deep learning algorithms.

In [5], an integrated dimension reduction convolutional neural network (IDRCNN) model was proposed. The study showed that the accuracy of the model increased when the IDRCNN approach was used, demonstrating its effectiveness.

In [6], the authors demonstrated a DCNN multipurpose transfer learning method, which transfers knowledge learned from non-medical images to medical diagnostic tasks via supervised training. This approach improves the generalization ability of DCNN for real-world tasks. The dataset included digital screen-film mammograms (SFMs), digital mammograms (DMs), and images from Institutional Review Board (IRB)-approved patient files, as well as additional SFMs collected for mammography screening from a digital database.

In [7], the authors developed an algorithm using deep convolutional neural networks to distinguish invasive stem tissue cancer from benign biopsies. The algorithm successfully identified experimental biopsy sets as invasive cancer, based solely on stem tissue characteristics, differentiating them from benign biopsies.

In [8], the significance of transfer learning was explored, and the best adaptation approach for training the CNN model was empirically determined. The study demonstrated how recent and powerful CNNs could be successfully applied in practice, yielding better results compared to other developed techniques that group similar datasets.

In [9], the authors proposed an improved semi-supervised learning (SSL) scheme based on a graph, applying DCNN for breast cancer detection. Unlike traditional CNNs, which require large labeled datasets for training, this method only needed a small amount of labeled data for training. The system consists of four modules: feature selection, CNN, dividing shared training data labels, and data weighting.

In [10], a computer-aided diagnosis (CAD) system using CNN was presented for classifying breast lesions in optical tomographic images. The system utilized a dense optical tomographic setup suitable for repetitive mass screening. The dataset comprised 63 optical tomographic images from women with dense breasts, and a set of 1,260 2D grayscale images was generated from these 3D images.

In [11], a cross-domain multipurpose transfer learning framework was proposed for transferring trained knowledge from non-medical images to medical image recognition tasks, thus improving generalization. The first step involved transferring knowledge from ImageNet-trained DCNNs to DCNNs trained on mammography images.

In [12], two segmentation strategies were applied. The first strategy involved manually assigning regions of interest (ROI), while the second applied thresholding and area-based methods. DCNN was used for feature extraction, and the well-known AlexNet framework was fine-tuned to classify two levels instead of the original 1,000. The final fully connected (FC) layer was linked to a support vector machine (SVM) classifier to achieve good accuracy.

In [13], a machine learning model was developed to provide physicians with a second opinion for predicting early-onset breast cancer using image data. The goal was to reduce errors in breast cancer staging and improve healthcare efficiency. The authors tested multiple classifiers, including MLP, Logistic Regression, Ridge, Random Forest, Decision Tree, Support Vector Machine, and Convolutional Neural Network, to identify the most accurate model.

In [14], an end-to-end system based on the You-Only-Look-Once (YOLO) concept was proposed to detect and categorize suspicious breast tumors in mammograms. This approach preprocesses raw images, detects abnormal regions as breast lesions, and classifies them as masses or calcifications.

3- The proposed method

The proposed method includes some stages as: (1) image pre-processing, (2) image segmentation, (3) feature extraction, (4) classification, (5) evaluation. Fig. 1 illustrates diagram of the proposed method.

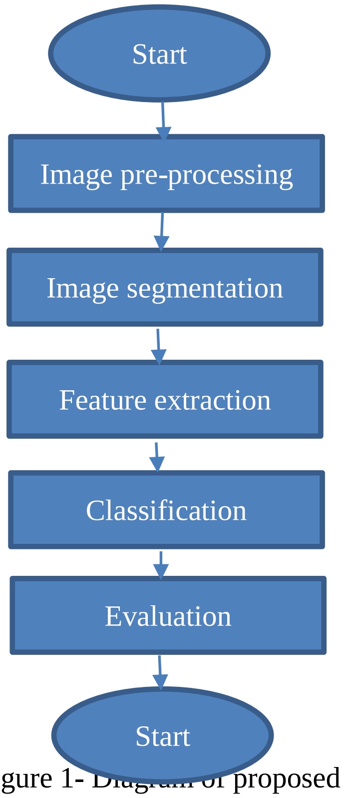


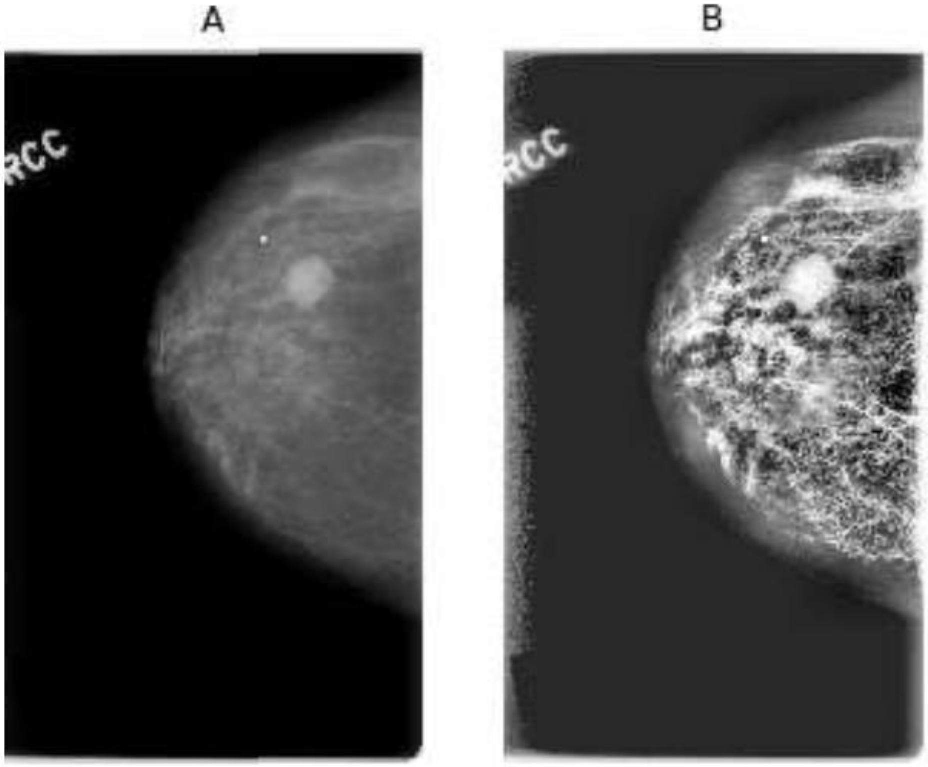
Figure 1- Diagram of proposed method

3-1- Image pre-processing

Intensifier of Image refers to mammographic image processing for raising contrast also avoiding noise for aiding the radiologists in detecting unusual cases. Here, Contrast Limited Adaptive Histogram Equalization (CLAHE) referred to as AHE kind, is applied to enhance image contrast.

The algorithm of CLAHE could be briefed as:

1. share the basic image in background regions with similar size
2. use equalization of histogram in every region,
3. restrict the histogram with the class of cut clip,
4. cut amount share among histogram,
5. achieving the number of pixels through merging histograms.



3-2- Image segmentation

Segmentation of the image is applied for sharing the image in sections having the same attributes and features. Its basic aim is to simplify images by providing a simple and analyzable technique.

Techniques of threshold are the easiest path in sharing the image. Pixels of Images are shared based on the level of intensity. global threshold is the most usual threshold technique kind [1].

Segmentation based on region is easier than other techniques. The technique shares images in various regions given the predefined variable. Area-growth algorithm can eliminate the region from an image based on several predefined variables like intensity. Area growth is the strategy of image segmentation where neighbor pixels are checked and joined to the region level where no edges exist. This is grouped as a technique of pixel-based image segmentation due to choosing a basic point of searching [1].

Here, the ROI is extracted from a basic mammographic image with two various techniques. DNN refers to the first ROI determination technique applying helix forms. Tumors in a dataset of DDSM are specialized with a red shape, so, such forms are assigned manually through checking tumor pixel values and applying them for region extraction. The second technique applies techniques based on threshold and area for assigning ROI.

3-3- Feature extraction with convolutional neural networks

A lot of techniques exist for the stage of feature extraction. In recent years, CNN has attracted much attention because of its prominent efficiency. Therefore, novel CNN is applied here. CNN has been successful in problems of classifying images like analysis of images. CNN includes some teachable stages which are stacked on top of each other and pursued by the supervised classification and arrays group known as maps of feature. 3 basic layers exist for creating CNN architectures: (1) convolutional layer, (2) reservoir layer, and (3) fully connected (FC) layer.

The function of activation is the main neural network section. The function of activation is the way "activated neuron" features are kept and mapped via non-linear functions that refer to a key to solving non-linear issues in neural networks. Here, the developed function of activation is applied to developing CNN [15].

The CNN architecture in the proposed method is shown in Table 1.

Table 1- Proposed CNN architecture

Input size	227×227×3		
Layer 1			
Conv1	55×55×96	Filter size	11×11
		Stride	4
SRelu	55×55×96		
Pool1	27×27×96	Pooling size	3×3
		Stride	2
Layer 2			
Conv2	27×27×256	Filter size	5×5
		Stride	1
SRelu	27×27×256		
Pool2	13×13×256	Pooling size	3×3
		Stride	2
Layer 3			
Conv3	13×13×384	Filter size	3×3
		Stride	1
SRelu	13×13×384		
Layer 4			
Conv4	13×13×384	Filter size	3×3
		Stride	1
SRelu	13×13×384		
Layer 5			
Conv5	13×13×256	Filter size	3×3
		Stride	1
SRelu	13×13×256		
Pool5	6×6×256	Pooling size	3×3
		Stride	2
Fully connected layer	4096×2		

3-4- Advanced activation function

The expression of the softplus rectified linear activation unit (SReLU) activation function is as follows:

$$SReLU = \begin{cases} 0. & x < 0 \\ ln(e^{x} + 1). & x \geq 0 \end{cases} \tag{1}$$

Hybrid function completes scattered processing of image and does not vanish neurons of function that are closer to brain neurons smooth nonlinear reflection. To guarantee the dispersion of the network, the increased function of activation returns a value below 0-0. When the argument is above 0, the dependent criterion value should be similar to the softplus function value. activation functions integration has smooth features removes useless info in the info of data and maintains features of data as much as possible. In the reverse process of guidance, the algorithm of backpropagation (BP) is applied. While the rate of error is computed from the final layer output, every layer should multiply this layer’s input neuron value. the layer is exponentially decreased. In the process of multilayer backpropagation, the inverse technique of computation is adapted so the slope is decreased continuously, vanishes, and network learning efficiency is reduced. When SReLU is applied as the function of activation in whole CNN layers, info of data might not be complete because of data features dispersion. Based on the mentioned issues, SReLU activation function integration is presented and a layer of local response normalization (LRN) is applied after every layer of max-convolutional collector, which guarantees speed and data processing scatter and avoids flaws of data because of dispersion. Based on the mentioned issues, the present work offers that multiple nonlinear function SReLU is shown as activation function of the network as [15]:

$$ReLU = \max(o.x)$$

$$SReLU = \begin{cases} 0. & x < 0 \\ ln(e^{x} + 1). & x \geq 0 \end{cases} \tag{2}$$

In CNN made by novel function, the function of ReLU is applied behind a convolutional layer, FC layer matches SReLU as an activation function network structure.

3-5- Classification

At this stage, ROI is grouped into benign/malignant given the features. Several classification techniques exist; like Binary Decision Tree (DT) and SVM, Linear Discriminant Analysis (LDA), and Artificial Neural Networks (ANN). Here, a decision tree is applied.

4- Discussion and evaluation of results

We used a Python environment for implementing CNN. In Python, there are several libraries in the field of deep learning. One of the libraries that we used is the TensorFlow library.

4-1- Description of data set

The images used in this study were collected from a digital database for screening mammography (DDSM). The DDSM database contains 322 cancer images of benign and malignant breast tissue. The basic image is sized 1024 x 1024 pixels and is in the format of DICOM. This must be considered that basic cancerous masses boundaries in database images were first performed by radiologists and a report on malignant/ benign masses is accessible as well in this database for every image. Fig. 3 illustrates the DDSM database image sample.

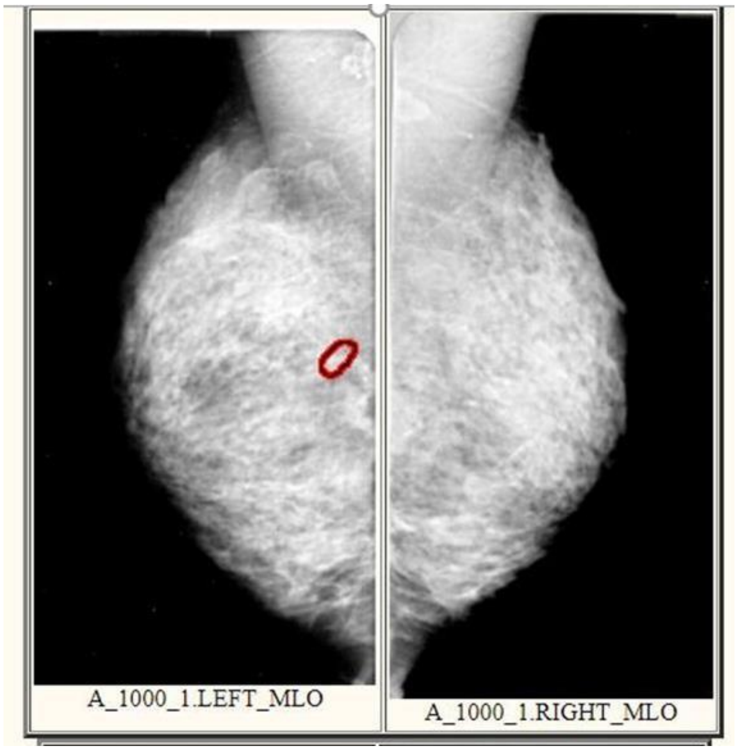


Figure 3- A sample of DDSM database images [16]

The table 2 highlights the essential aspects of the data utilized in the study and includes information required to analyze and identify abnormalities in medical imaging. The first column contains the database reference number, a unique identifier for each sample or image in the dataset. The second column describes the type of background tissue in which the aberration appeared, which is divided into three categories: fat (F), sebaceous glands (G), and dense glands. The third column specifies the type of abnormality seen, which covers categories such as calcification (CALC), well-defined masses (CIRC), spiny masses (SPIC), ill-defined masses (MISC), structural distortion (ARCH), asymmetry (ASYM), and normal state. The fourth column specifies the degree of the anomaly, which is classified as either benign (B) or malignant (M). The fifth and sixth columns provide the coordinates of the anomaly's center in the image, which are represented by x and y coordinates and are needed to precisely locate the anomaly. The seventh column displays the anomaly's approximate radius (in pixels), which shows the size of the circle that surrounds it. These properties lay the groundwork for advanced medical picture analysis, modeling, and automated diagnosis by providing detailed information on the type, location, and degree of anomalies. Details of the DDSM are described as follows.

Table 2-. The characteristics of breast cancer disease

Features	Number
Column 1: The reference number of the database	1
Column 2: Background tissue feature: F. fat G. sebaceous gland D. Dense glands	2
Column 3: Available anomaly class: CALC - Calcification CIRC - Fully determinate/restricted masses SPIC - Spoked masses MISC - Other Undefined Masses ARCH – distortion of architecture ASYM - Asymmetry NORM – it is normal	3
Column 4: intensity of anomaly; B - Benign M - malignant	4
Columns 5, 6: Image coordinates y, x coordinates of the anomaly center.	5
Column 7: Approximate radius (in pixels) of a circle surrounded by an anomaly	6

4-2-The evaluation criteria

Matrix of confusion has info on certain and predicted classifications. Model efficiency could be assessed given the matrix data. If the data are two classes of normal and abnormal, the confusion matrix of this data is in the form of Table 3.

Table 3- Confusion matrix

The detected category			
Abnormal sample	Normal sample		
FN	TP	Normal sample	Actual category
TN	FP	Abnormal sample	

- True negative (TN): unusual instances’ number which were accurately diagnosed.
- False negative (FN): usual instances’ number which were diagnosed as unusual.
- True positive (TP): usual instances’ number which were diagnosed accurately.
- False positive (FP): unusual instances’ number which were diagnosed as usual.

In the following, we will use more accurate measurement parameters to compare these two methods more accurately. One of the solutions that is used to display the precision of data classification is to find the accuracy value of output rules. Which is calculated as follows:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \qquad (3)$$

The parameter of sensitivity refers to the other variables' kind for illustrating laws' efficiency. For computing precision percent to diagnose sickness laws. When the value is higher, laws of diagnosis have a higher ability to detect.

$$Recall = \frac{TP}{TP+FN} \qquad (4)$$

Test capability to find healthy items is known as a parameter of the feature. For computing test specificity, true negatives rate to total true negatives and false positives should be achieved. Mathematically, the rate could be mentioned as:

$$Specificity = \frac{TN}{TN+FP} \qquad (5)$$

The precision parameter is another parameter that indicates the probability of being sick and the correctness of that prediction. If this value is higher, it has higher rules for disease diagnosis.

$$Precision = \frac{TP}{TP+FP} \qquad (6)$$

F1 Score is defined as the harmonic mean of sensitivity and precision.

$$F1\ score = \frac{Precision \times Recall}{Precision+Recall} \qquad (6)$$

4-3- The initialization of parameters

To examine the presented algorithm quality, we set parameters based on paper parameters [12]. In such a way we considered the total number of images equal to 322 and the number of classes was considered as two, i.e. benign and malignant. Table 4 shows the settings related to the parameters.

Table 4- The initial values for parameters

Parameters	Values
Total number of images	322
Number of classes	2
Images size	227
Filter	10
Core size	[11.11]
Epoch number	20

4-4- The evaluation of results

Firstly, instances were shared and augmented by applying two techniques mentioned in the "presented technique". Then, features were extracted by applying CNN. Instances applied the technique of SVM for classification.

The presented technique architecture input layer needs an image size of 227 × 227 × 3. So, a pre-processing stage exists for changing whole images of input to the needed size with no size consideration. In the training and testing stages, the dataset is divided into two parts so that the model can be trained with the training data and then evaluated on the new data (test data). The data set split ratio is as follows: 80% of the total data is used to train the model.

20% of the data is used for model testing.

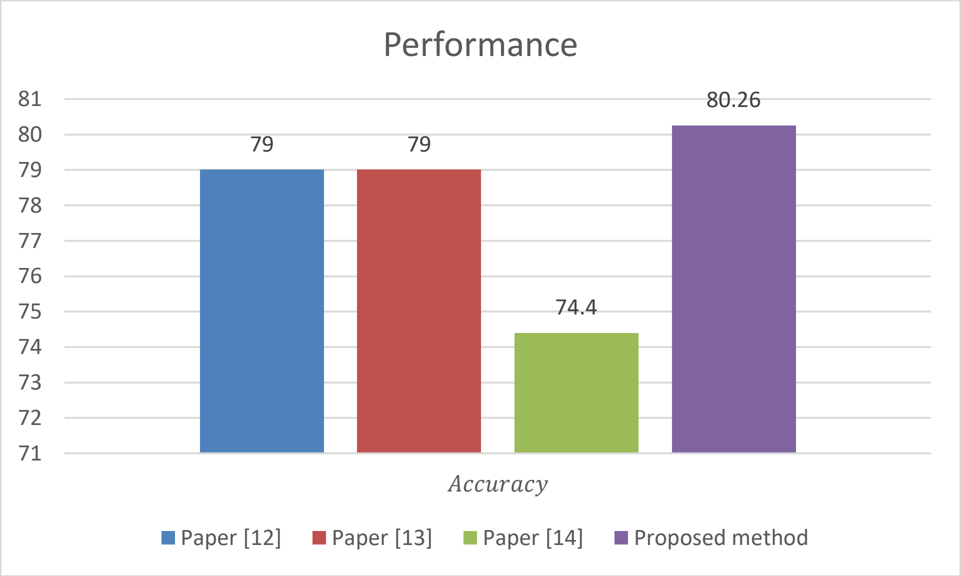
While applying the first technique of segmentation, recently trained architecture accuracy was just 71.01%. the outcome was obtained while extracting and

classifying injuries with CNN. So, while linking CNN to the decision tree to obtain a good outcome, accuracy with linear core performance was 79% with an AUC of 80.26 (80%), accuracy with CNN in [13] was 79%, and accuracy with YOLO in [14] was 74.4%. Fig. 4 illustrates the presented technique efficiency comparison with reference [12-14] among benign and malignant tumors instances. As is obvious from the outcomes, the presented technique has good performance.

The observed advantages in the combined CNN and decision tree method highlight the usefulness of hybrid approaches. While CNN's performance was average on its own, when combined with a decision tree, the extracted features were better utilized, resulting in superior outcomes. These findings indicate that the proposed system has tremendous potential as an automatic and effective method for detecting breast lesions in MRI images.

Table 5- The efficiency criteria of each method in the test phase in the first test

Criteria	Paper [12]	Paper [13]	Paper [14]	Proposed method
Accuracy	79.00	79.00	74.4	80.26
Sensitivity	76.30	-	-	84.54
specificity	82.20	-	-	96.34
Precision	85.00	-	-	81.12
F1	80.00	-	-	81.62



Fi

5- Conclusion

This paper proposes a new CAD system for predicting breast cancer. We will use two segmentation methods: the first approach includes manually determining the ROI, while the second approach uses the threshold method and is area-based. We also use a new CNN for extracting features and classifying the final tumor based on the decision tree. To achieve better accuracy, the last fully connected layer on CNN was replaced with DT. When comparing the two segmentation methods for the DDSM dataset, it was found that DT provides promising results for the second segmentation method. In terms of accuracy, sensitivity, specificity, precision, and F1 score, it achieved 80.26%, 84.54%, 96.34%, 12.81%, and 81.62%, respectively.

References

[1] Patterson, Kellie, Sarah Byun, and John Hines. "Placenta Accreta Spectrum on MRI." Contemporary Diagnostic Radiology 47, no. 1 (2024): 1-5.

[2] Wang, Hui, Bas HM van der Velden, Erik Verburg, Marije F. Bakker, Ruud M. Pijnappel, Wouter B. Veldhuis, Carla H. van Gils, and Kenneth GA Gilhuijs. "Automated rating of background parenchymal enhancement in MRI of extremely dense breasts without compromising the association with breast cancer in the DENSE trial." European Journal of Radiology (2024): 111442.

[3] Rayed, Md Eshmam, SM Sajibul Islam, Sadia Islam Niha, Jamin Rahman Jim, Md Mohsin Kabir, and M. F. Mridha. "Deep learning for medical image segmentation: State-of-the-art advancements and challenges." Informatics in Medicine Unlocked (2024): 101504.

[4] Hussain, Lal, Sara Ansari, Mamoona Shabir, Shahzad Ahmad Qureshi, Amjad Aldweesh, Abdulfattah Omar, Zahoor Iqbal, and Syed Ahmed Chan Bukhari. "Deep convolutional neural networks accurately predict breast cancer using mammograms." Waves in Random and Complex Media (2023): 1-24.

[5] Xu, HuanQing, Xian Shao, Shiji Hui, and Li Jin. "Supervised breast cancer prediction using integrated dimensionality reduction convolutional neural network." Plos one 18, no. 5 (2023): e0282350.

[6] Samala, Ravi K., Heang-Ping Chan, Lubomir M. Hadjiiski, Mark A. Helvie, Kenny H. Cha, and Caleb D. Richter. "Multi-task transfer learning deep convolutional neural network: application to computer-aided diagnosis of breast cancer on mammograms." Physics in Medicine & Biology 62, no. 23 (2017): 8894.

[7] Bejnordi, Babak Ehteshami, Maeve Mullooly, Ruth M. Pfeiffer, Shaoqi Fan, Pamela M. Vacek, Donald L. Weaver, Sally Herschorn, et al. "Using deep convolutional neural networks to identify and classify tumor-associated stroma in diagnostic breast biopsies." Modern Pathology 31, no. 10 (2018): 1502.

[8] Chougrad, Hiba, Hamid Zouaki, and Omar Alheyane. "Deep convolutional neural networks for breast cancer screening." Computer methods and programs in biomedicine 157 (2018): 19-30.

[9] Sun, Wenqing, Tzu-Liang Bill Tseng, Jianying Zhang, and Wei Qian. "Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data." Computerized Medical Imaging and Graphics 57 (2017): 4-9.

[10] Xu, Qiwen, Xin Wang, and Huabei Jiang. "Convolutional neural network for breast cancer diagnosis using diffuse optical tomography." Visual Computing for Industry, Biomedicine, and Art 2, no. 1 (2019): 1.

[11] Samala, Ravi K., Heang-Ping Chan, Lubomir Hadjiiski, Mark A. Helvie, Caleb Richter, and Kenny Cha. "Cross-domain and multi-task transfer learning of deep convolutional neural network for breast cancer diagnosis in digital breast tomosynthesis." In Medical Imaging 2018: Computer-Aided

Diagnosis, vol. 10575, p. 105750Q. International Society for Optics and Photonics, 2018

[12] Ragab, Dina A., Maha Sharkas, Stephen Marshall, and Jinchang Ren. "Breast cancer detection using deep convolutional neural networks and support vector machines." PeerJ 7 (2019): e6201.

[13] Leung K. "Classification of Early Breast Cancer using Convolutional Neural Networks. Journal of Student Research." 2024 Feb 29;13(1).

[14] Baccouche A, Garcia-Zapirain B, Olea CC, Elmaghraby AS. "Breast Lesions Detection and Classification via YOLO-Based Fusion Models." Computers, Materials & Continua. 2021 Oct 1;69(1).

[15] Guo, Yanhua, Lei Sun, Zhihong Zhang, and Hong He. "Algorithm Research on Improving Activation Function of Convolutional Neural Networks." In 2019 Chinese Control And Decision Conference (CCDC), pp. 3582-3586. IEEE, 2019.

[16] http://www.eng.usf.edu/cvprg/mammography/database.html