# Breast cancer diagnosis based on support vector machine techniques

#### Ruaa H. Mohammed Ameen<sup>1</sup>, Nasseer Moyasser Basheer<sup>1</sup>, Ahmed Khazal Younis<sup>2</sup>

<sup>1</sup>Department of Medical Instrumentation Technology Engineering, Technical Engineering College of Mosul, Northern Technical University, Mosul, Iraq

<sup>2</sup>Department of the Computer Techniques Engineering, Technical Engineering College of Mosul, Northern Technical University, Mosul, Iraq

# Article Info

#### Article history:

Received Feb 4, 2023 Revised Jun 9, 2023 Accepted Jun 17, 2023

### Keywords:

Breast cancer CAD system Image processing Machine learning Support vector machine

# ABSTRACT

In general, breast cancer is a fatal disease; however, early detection can significantly reduce the risk of death. A physician's experience in detecting and diagnosing breast cancer can be aided by automated feature extraction and classification procedures. Clinical exams and imaging studies are typically used to make a diagnosis of breast cancer. Mammography is by far the most common imaging technique used to detect the early warning signs of breast cancer. The goal of this paper is to design a computer-aided diagnosis/ detection (CAD) system by utilizing image processing techniques. These techniques will represent the first stage in the system, and they will significantly contribute to improving diagnostic accuracy. Next is the "Histogram of oriented gradients (HOG)" technique, which was used to extract features. The final stage involves applying machine learning techniques (MLT), in this case the support vector machine (SVM), a widely used method for detecting breast cancer using mammograms. In testing, the proposed model was found to be 94.74% accurate.

*This is an open access article under the <u>CC BY-SA</u> license.* 



## **Corresponding Author:**

Ruaa H. Mohammed Ameen Department of Medical Instrumentation Technology Engineering, Technical Engineering College of Mosul, Northern Technical University Mosul, Iraq Email: ruaa.hassan1@ntu.edu.iq

## 1. INTRODUCTION

Breast cancer is among the leading global causes of death for women. A higher mortality rate is caused by dense population, patients' ignorance of sickness symptoms, and their delayed seeking for medical advice. Additionally, the lack of medical experts and professionals in rural locations makes it more difficult to provide an early diagnosis accuratly [1]. Utilizing information systems and medical data to develop medical supporting systems that are able to aid the doctor in thinking about and identifying cases of breast cancer is one way that may be used to improve the early diagnosis of breast tumors. As a direct consequence of this, there is a higher probability of making a full recovery and a lower risk of passing away [2].

Imaging analysis is the greatest active method of detecting breast cancer. Many medical imaging methods used for diagnosis, including ultrasound (US), magnetic resonance imaging (MRI), digital mammography (DM), and Histopathological images. These are used to help doctors and radiologists to diagnose the condition [3]. The use of information technology is necessary since image interpretation is operator-dependent and needs expertise. This may speed the process and provide the specialists another perspective, improving the accuracy of the diagnosis [4].

A helpful tool for professionals and doctors to identify problems is the computer-aided diagnosis/detection (CAD) system [5]. There have been several attempts to create CAD systems that make use of advances in artificial intelligence, image processing, and pattern recognition. By use CAD systems, fewer operators are required, diagnostic accuracy increases, and the cost of auxiliary medical modalities decreases [6]. In general, the design of any CAD system relies on image processing techniques to improve the images used, extract features, and use artificial intelligence techniques to classify medical images [7].

In order to assure better care and lower the risk of breast cancer-related death, this work aims to raise likelihood of early breast tumor detection. This research suggests, however, that a completely automated CAD system be created for identifying, localizing, and categorizing breast masses as benign or cancerous using mammogram images. Despite the excellent work that has previously been done to create tools to aid in the diagnosis of malignant illnesses, it is still rather difficult because of the complexity of the images reflecting the patients inner condition [8].

#### 2. LITERATURE REVIEW

There are some relevant studies, especially those that relied on machine learning algorithms (ML) and in particular the support vector machine (SVM) classifier, a tool for diagnosing mammograms and displaying the results of previous studies on the basis of accuracy: Ribli *et al.* [9] demonstrated an R-CNN-based CAD system. Without assistance of the human, the technology can identify and categorize tumors on a mammogram as benign or malignant. On the open INbreast database, the suggested technique achieves cutting-edge classification performance with an accuracy of 95%. The proposed method placed second in "the digital mammography DREAM challenge", with an accuracy of 85%.

Shravya *et al.* [10] compare several classification algorithms for breast cancer detection that is logistic regression (LR), k-nearest neighbor (kNN), and SVM on mammogram datasets. The performance of each method is evaluated and compared with the value of accuracy. The scientific python development environment, Spyder, is used to implement these methods once they have been programmed in Python. With an accuracy of 92.7%, their experiment demonstrates that the SVM is the best for predictive analysis.

Farhan and Kamil [11] proposed a technique to identify breast cancer using mammograms. This study makes use of preprocessing, segmentation, feature extraction, and classification. A differentiation is established between various classification techniques after the results have been obtained. Breast cancer is categorized using (SVM), LR, and kNN classifiers, in that order, the accuracy which is 90%, 92.5%, and 89.3% respectively. That shows that the SVM is the best classifier.

Darabi *et al.* [12] proposed a CAD system, utilized for feature selection the random subset feature selection (RSFS) algorithm and genetic algorithm (GA). Additionally, as classifier techniques, the SVM and kNN algorithms were used. The proposed CAD system was verified using matrix laboratory (MATLAB) 2017 and a dataset that was composed of breast images from 78 patients. The implementation results demonstrated that using the RSFS algorithm for feature selection and kNN and SVM algorithms as classifiers had an accuracy of 85.36% and 75%, respectively.

Al-Fahaidy *et al.* [13] proposed a CAD system by using mammographic image analysis society (MIAS) dataset through several stages of image pre-processing, image segmentation using the region growing (RG) algorithm, extraction of features using the sequential forward selection (SFS) technique, and finally the SVM algorithm was used as a binary classifier to classify the images as either benign or malignant. The CAD system was been applied using MATLAB software. The proposed system attained classification accuracy equal to 87.1%.

Alfian *et al.* [14] proposed use a highly randomized trees classifier (extra-trees) is paired with a SVM to offer an early diagnosis of breast cancer based on danger variables. SVM was used to detect the breast cancer status while the extra-trees classifiers were employed to eliminate unnecessary specifics. Using a dataset of 116 breast cancer patients, machine learning algorithms were utilized to predict breast cancer. The maximum accuracy was attained by the suggested merged SVM and extra-trees model, which was 80.23%.

In this study, a CAD system based on SVM classifier is developed to categorize breast cancer pictures into benign and malignant forms. The dataset from the MIAS was used for this investigation. The remainder of the essay is structured as follows: The related work is presented in section 2, the dataset and the proposed method are presented in section 3, the preprocessing method is explained in section 4, the classification results are presented in section 5, and the conclusions and future work are presented in the final section. Table 1 shows a summary of the results of related studies and their comparison with the results of the proposed method.

Table 1. A summury of the previous researches					
Ref	Dataset	Classification technique	Evaluation results		
			Acc	Sn	Sp
Ribli et al. [9]	INbreast database	R-CNN	95%	90%	
Shravya et al. [10]	Private mammogram dataset	LR	92.10%	91%	93%
		kNN	92.2%	90.5%	95.12%
		SVM	92.7%	91.07%	95.14%
Farhan and Kamil [11]	Mini-MIAS	LR	92%	88%	90%
		SVM	90%	72%	100%
		kNN	84%	32%	99%
Darabi et al. [12]	Private mammogram dataset	SVM	75%	79%	-
		kNN	85.36%	94.11%	-
Al-Fahaidy et al. [13]	MIAS dataset	SVM	87.1%	90%	85.7%
Alfian <i>et al.</i> [14]	Private mammogram dataset	SVM+extra trees	80.23%	78.57%	78.57%
The proposed work	MIAS dataset	SVM	94.7%	100%	87.5%

Table 1. A summury of the previous researche

#### 3. PROPOSED METHOD

The CAD system's stages for diagnosing breast cancer should all be included, preprocessing stage for the mammogramphy dataset images, feature extraction stage the entered data for this stage consisting of all features and the output is a data with only the selected features. The extracted data is then fed to classifier to be classified in the classification stage. The proposed system's block diagram as shown in the Figure 1.



Figure 1. Diagrammatic representation of the planned computer system

#### 3.1. Image dataset

A suggested system's accuracy was evaluated using the MIAS dataset. This database, which was established in 1994 and is often used to assess automated mammography analysis methods, is openly accessible for scientific study. All images (.pgm) format and size is (1024×1024) pixels. The database contains of (113 abnormal cases) and (209 normal cases) spread among 161 pairs of medio lateral oblique (MLO) view mammography (right and left views)."The National Breast Screening Program in the United Kingdom" used a film-screen mammographic imaging procedure to create the images in the database [15]. Figure 2 shows one of MIAS dataset images.



Figure 2. One of the MIAS dataset images [15]

#### 3.2. Preprocessing stage

The mammogram image contains several details, including important details in the classification process, and some that are undesirable and fall within the noise that have a role in affecting the classification accuracy [16]. Figure 3 shows the components detected in the original mammogram image and numbered as: 1. Pictorial muscle, 2. The background label, 3. Black background, 4. The tumor, and 5. Breast muscle. The proposed steps of the preprocessing stage tend to remove all the unwanted details from the mammogram images and improve the images quality. Labels in the background, black background, and pectoral muscle are among the noise and unwanted details present in the mammography image.



Figure 3. Components of the mammogram image

This stage's goal is to get the images ready in order to get the very best results for classification. In this stage, the region of interest (ROI) is defined, extraneous details are removed from the images as well as The enhancement of digital mammography is accomplished utilizing "wiener filter" and "contrast limited adaptive histogram equalization" (CLAHE) methods respectively. These noise and unwanted details should all be removed. The breast area is divided once the noises have been eliminated by using a unique threshold approach, which blackens the backdrop solely to distinguish it from the mammography image content. Morphological erosion image analysis modifies the spatial shape or structures of objects within an image and is used to eliminate discrete areas [17].

There have been a great number of investigations into the problem of noise abatement over the last several decades. Therefore, the "wiener filter" is an ideal stationary linear filter that uses mean square error (MSE) to assess the difference between the original, undistorted images of an item and the image that has been filtered. Therefore, the linear noise in the mammography images was removed, and the resulting images displayed a greater was present in the mammography pictures were removed, and the resulting images displayed a greater degree of clarity [18].

CLAHE is an enhancing technique used to increase the image's contrast. Instead of working on the full image, it divides it into equal-sized tiles and then performs operations on each tile. Histogram equalization is applied to each tile to increase its contrast, and each histogram is then cropped using a boundary that relies on the desired contrast development and the size of the surrounding area [19]. The challenge of noise amplification is then removed by joining the neighboring areas using bilinear interpolation to reduce arbitrarily created borders and restrict the contrast in the homogenous regions [20]. Two factors should be taken into account while using

240

CLAHE. One of these factors is the threshold rate that is set (by default) to 40 in the clip limit. The other factor is tile grid size, which is set (by default) to 8\*8 and determines how many tiles will be in each row and column.

The output image has significant improvement and provides much detail and also reduces the noise in the output. The image that has a preprocessed applied to it shows more details than the original image. Figure 4 shows putting an image from the MIAS dataset through a series of preprocessing procedures that include: Figure 4(a) the original image, Figure 4(b) remove the background label, Figure 4(c) remove the black background, Figure 4(d) Remove the pectoral muscle, Figure 4(e) apply wiener filter, and Figure 4(f) apply CLAHE.



Figure 4. Applying preprocessing steps to an image of a MIAS dataset: (a) the original image, (b) remove the background label, (c) remove the black background, (d) remove the pectoral muscle, (e) apply wiener filter, and (f) apply CLAHE

## 3.3. Feature extraction stage

The attributes of each image are determined through feature extraction during the classification step. Histogram of oriented gradients (HOG) was used in this study's feature extraction. The characteristics that used to filter the CAD output are based on Dalal and Triggs's HOG [21]. These attributes have been effectively used by the computer vision community for the purposes of object localization and detection. The HOG method is based on the idea that the histogram of the local intensity gradients or edge directions in a dense grid may be utilized to determine the mass's natural look and structure. This is the approach's fundamental assumption. The distribution of the bulk structures that have been categorized is captured by the HOG histogram, which combines structural and statistical methods [11].

This method uses the gradient of the local grey levels to capture the structural form inside the ROI. In the HOG descriptor, the ROI is partitioned into tiny, related sections called cells, and produced the histogram of gradients for every cell rather than the complete ROI. The gray scale value is calculated for blocks, which are bigger overlapping sections [22]. These blocks are created by cells that have gathered in ROI. The cells inside the blocks are normalized using the contrast normalization method and the computed gray scale value for each block. The HOG features were then produced using the normalized ROI [23].

#### **3.4.** The classification stage

One of the many machine learning techniques (MLT) that may identify, predict, and diagnose breast cancer is the SVM. The SVM is becoming more popular as an effective method for dealing with regression and classification issues. The fundamental idea behind SVM is to generate a hyperplane by maximizing the difference between two classes. The input data is mapped into a space with a higher dimension by the Kernel function, which may be a linear, quadratic, polynomial, or radial base function [24], [25]. A schematic representation of the SVM classifier is shown in Figure 5. After the classifier has been trained with the use of examples of training data, it is then given test data to classify to evaluate how well the classifier is doing its job.



Figure 5. SVM classifier [25]

#### 4. METHOD

The proposed model of the CAD system was implemented in MATLAB 2020 environment. After entering MIAS dataset into the proposed model and applying it to preprocessing steps and feature extraction using HOG technology, the data was entered into the SVM classifier. In this study, the linear kernel function was used and the dataset has been split into two parts, training data (90%), and testing data (10%) The model achieved an accuracy of 94.7%.

The samples are divided into two distinct groups: those that are positive and those that are negative. Positive samples are those that represent malignant (cancerous) samples, whereas negative samples are those that represent benign (non-cancerous) samples. The primary category of importance is the positive samples, which represent malignant (cancerous) samples.

## 5. RESULTS AND CALCULATIONS

The model's performance is evaluated using the commonly used evaluation metrics for CAD systems are: Accuracy (Acc) reflects the degree to which the actual class matches the one that was predicted. It implies

- a) Accuracy (Acc) reflects the degree to which the actual class matches the one that was predicted. It implies that it shows the percentage of properly categorized samples (non-cancer and cancer) concerning the total number of samples. In other words, it means that it indicates the proportion of samples that were correctly categorized.
- b) Sensitivity (Sn) is the so-called "true positive rate," which indicates the proportion of (cancer) samples that have been correctly identified.
- c) Specificity (Sp) is the rate at which samples that do not contain cancer are correctly detected, and it serves as a measure of the genuine negative rate. Equations of (Acc, Sn, and Sp) and the calculation for each of them are given as shown in (1)-(3) [13]:

$$Acc = (TP + TN)/(TP + TN + FP + FN) \times 100$$
  
= (22 + 14) /(22 + 14 + 2 + 0) × 100 = 94.7 % (1)

$$Sn = TP/(TP + FN) \times 100$$

$$= 22/(22 + 0) \times 100 = 100\%$$
(2)

$$Sp = TN/(TN + FP)$$
(3)  
= 14/(14 + 2) × 100 = 87.5%

where TP: true positive, TN: true negative, FP: false positive, FN: false negative. These evaluation metrics can be calculated by using the confusion matrix of this model that is shown in Figure 6.





Breast cancer diagnosis based on support vector machine techniques (Ruaa H. Mohammed Ameen)

#### 6. CONCLUSION AND FUTURE WORK

In this study, the breast cancer diagnosis system is based on artificial intelligence techniques was designed. The system was implemented in MATLAB 2020 environment. The mammogram images were processed by using image processing techniques, which led to an increase in the classification accuracy of about 20% and this research work reveals also that feature selection and feature extraction by using HOG method can help improve the diagnosis of benign and malignant tumors using machine learning techniques. In many previous studies, The SVM classifier was used to achieve high classification accuracy. The proposed system achieved an accuracy of 94.7% and it is worth noting that misclassification may occur due to the complex nature of breast tissue. The proposed model will be improved in the future to recognize the shapes and sizes of tumors making them more suitable. Also, the study can be expanded in the future by integrating the SVM classifier with neural networks and working on improving the accuracy.

#### REFERENCES

- A. Makandar and B. Halalli, "Mammography image analysis using wavelet and statistical features with SVM classifier," in *Lecture* Notes in Networks and Systems, Springer Singapore, vol. 14, pp. 371–382, 2018.
- [2] F. Bray, J. S. Ren, E. Masuyer, and J. Ferlay, "Global estimates of cancer prevalence for 27 sites in the adult population in 2008," *International Journal of Cancer*, vol. 132, no. 5, pp. 1133–1145, Jul. 2013, doi: 10.1002/ijc.27711.
- [3] A. Jalalian, S. B. T. Mashohor, H. R. Mahmud, M. I. B. Saripan, A. R. B. Ramli, and B. Karasfi, "Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review," *Clinical Imaging*, vol. 37, no. 3, pp. 420–426, May 2013, doi: 10.1016/j.clinimag.2012.09.024.
- [4] D. H. Lee and S. N. Yoon, "Application of artificial intelligence-based technologies in the healthcare industry: opportunities and challenges," *International Journal of Environmental Research and Public Health*, vol. 18, no. 1, pp. 1–18, Jan. 2021, doi: 10.3390/ijerph18010271.
- [5] B. Halalli and A. Makandar, "Computer aided diagnosis medical image analysis techniques," in Breast Imaging, InTech, 2018.
- [6] S. Radhakrishna *et al.*, "Role of magnetic resonance imaging in breast cancer management," *South Asian Journal of Cancer*, vol. 07, no. 02, pp. 069–071, Apr. 2018, doi: 10.4103/sajc.sajc\_104\_18.
- [7] T. Kooi et al., "Large scale deep learning for computer aided detection of mammographic lesions," Medical Image Analysis, vol. 35, pp. 303–312, Jan. 2017, doi: 10.1016/j.media.2016.07.007.
- [8] R. Touahri, N. Azizi, N. E. Hammami, M. Aldwairi, and F. Benaida, "Automated breast tumor diagnosis using local binary patterns (LBP) based on deep learning classification," in 2019 International Conference on Computer and Information Sciences (ICCIS), Apr. 2019, doi: 10.1109/ICCISci.2019.8716428.
- [9] D. Ribli, A. Horváth, Z. Unger, P. Pollner, and I. Csabai, "Detecting and classifying lesions in mammograms with deep learning," *Scientific Reports*, vol. 8, no. 1, Mar. 2018, doi: 10.1038/s41598-018-22437-z.
- [10] C. Shravya, K. Pravalika, and S. Subhani, "Prediction of breast cancer using supervised machine learning techniques," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 8, no. 6, pp. 1106–1110, 2019.
- [11] A. H. Farhan and M. Y. Kamil, "Texture analysis of breast cancer via LBP, HOG, and GLCM techniques," *IOP Conference Series: Materials Science and Engineering*, vol. 928, no. 7, Nov. 2020, p. 72098, doi: 10.1088/1757-899X/928/7/072098.
- [12] N. Darabi, A. Rezai, and S. S. F. Hamidpour, "Breast cancer detection using RSFS-based feature selection algorithms in thermal images," *Biomedical Engineering - Applications, Basis and Communications*, vol. 33, no. 3, p. 2150020, Mar. 2021, doi: 10.4015/S1016237221500204.
- [13] F. A. K. Al-Fahaidy, B. Al-Fuhaidi, I. Al-Darouby, F. Al-Abady, M. Al-Qadry, and A. Al-Gamal, "A diagnostic model of breast cancer based on digital mammogram images using machine learning techniques," *Applied Computational Intelligence and Soft Computing*, vol. 2022, pp. 1–17, Sep. 2022, doi: 10.1155/2022/3895976.
- [14] G. Alfian *et al.*, "Predicting breast cancer from risk factors using SVM and extra-trees-based feature selection method," *Computers*, vol. 11, no. 9, p. 136, Sep. 2022, doi: 10.3390/computers11090136.
- [15] A. A. Nahid and Y. Kong, "Involvement of machine learning for breast cancer image classification: a survey," Computational and Mathematical Methods in Medicine, vol. 2017, pp. 1–29, 2017, doi: 10.1155/2017/3781951.
- [16] M. Y. Kamil and E. A. Radhi, "Breast tumor segmentation in mammography image via Chan-Vese technique," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 22, no. 2, p. 809, 2021, doi: 10.11591/ijeecs.v22.i2.pp809-817.
- [17] A. Makandar and B. Halalli, "Pre-processing of mammography image for early detection of breast cancer," *International Journal of Computer Applications*, vol. 144, no. 3, pp. 11–15, 2016, doi: 10.5120/ijca2016910153.
- [18] A. Rakhlin, A. Shvets, V. Iglovikov, and A. A. Kalinin, "Deep convolutional neural networks for breast cancer histology image analysis," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*), vol. 10882, 2018, pp. 737–744.
- [19] M. H. Alhabib and O. H. Alhabib, "Detection of partially overlapped masses in mammograms," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 18, no. 1, pp. 235–241, Apr. 2019, doi: 10.11591/ijeccs.v18.i1.pp235-241.
- [20] B. Gupta and M. Tiwari, "A tool supported approach for brightness preserving contrast enhancement and mass segmentation of mammogram images using histogram modified grey relational analysis," *Multidimensional Systems and Signal Processing*, vol. 28, no. 4, pp. 1549–1567, Jun. 2017, doi: 10.1007/s11045-016-0432-1.
- [21] B. Li and G. Huo, "Face recognition using locality sensitive histograms of oriented gradients," *Optik*, vol. 127, no. 6, pp. 3489–3494, Mar. 2016, doi: 10.1016/j.ijleo.2015.12.032.
- [22] C. H. Salh and A. M. Ali, "Breast cancer recognition based on performance evaluation of machine learning algorithms," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 27, no. 2, pp. 980–989, Aug. 2022, doi: 10.11591/ijeecs.v27.i2.pp980-989.
- [23] A. H. Farhan and M. Y. Kamil, "Texture analysis of mammogram using histogram of oriented gradients method," *IOP Conference Series: Materials Science and Engineering*, vol. 881, no. 1, Jul. 2020, p. 12149, doi: 10.1088/1757-899X/881/1/012149.
- [24] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: aplications, challenges and trends," *Neurocomputing*, vol. 408, pp. 189–215, Sep. 2020, doi: 10.1016/j.neucom.2019.10.118.
- [25] N. M. Ali, N. A. A. Aziz, and R. Besar, "Comparison of microarray breast cancer classification using support vector machine and logistic regression with LASSO and boruta feature selection," *Indonesian Journal of Electrical Engineering and Computer Science* (*IJEECS*), vol. 20, no. 2, pp. 712–719, Nov. 2020, doi: 10.11591/ijeecs.v20.i2.pp712-719.

## **BIOGRAPHIES OF AUTHORS**





**Nasseer Moyasser Basheer b s s e** received his B.Sc. in electrical engineering from University of Mosul, Iraq, at 1980. Also received M.Sc. and Ph.D. from the same college at 1984 and 2008. Currently an assistant professor in the Technical Engineering College-Mosul/Northern Technical University. The research interests include image processing and compression, FPGA applications, and microcontroller applications. He can be contacted at email: nmbasheer@ntu.edu.iq.



Ahmed Khazal Younis **(D)** S S S works at Northern Technical University/Engineering Technical College-Mosul-Iraq/Department of Computer Technology Engineering. He Completed his undergraduate education (1998-2002), master degree (2005-2008) at Computer Technical Engineering Department in Technical College/North Technical University and Ph.D. degree in computer engineering/embedded system design from Yasar University in Izmir/Turkey in 2019. His research interests include: neural networks, neurofuzzy system, microcontrollers, and FPGA devices. He can be contacted at email: ahmedkhazal@ntu.edu.iq.