

# Breast Mass Classification Based on Hybrid Discrete Cosine Transformation–Haar Wavelet Transformation

Bestan Maaroof Bahaalddin<sup>1,\*</sup>, Hawkar Omar Ahmed<sup>2,3</sup>

<sup>1</sup>Department of Information Technology, College of Commerce, University of Sulaimani, Sulaimani City, Kurdistan Region- F. R. Iraq.

<sup>2</sup>Department of Information Technology, College of Commerce, University of Sulaimani, Sulaimani City, Kurdistan Region- F. R. Iraq.

<sup>3</sup>Department of Information Technology, University College of Goizha, Sulaimani City, Kurdistan Region- F. R. Iraq

\*Corresponding author's email: Bestan2014@gmail.com

Received: 01-06-2020

Accepted: 04-10-2020

Available online: 31-12-2020

## ABSTRACT

Mammography is the most effective procedure for the early detection of breast cancer. In this paper, an efficient computer-aided diagnosis system was proposed to discriminate between a benign and malignant breast mass. The system comprises mainly 3 steps, namely preprocessing of the images, feature extraction, and finally, classification and performance of the analysis. This study sampled mammographic images that originated from the mini Mammographic Image Analysis Society database. In the preprocessing phase, the region of interest was cropped and resized to 128×128. The feature extraction process started with the application of a Haar wavelet transformation to 5 levels, followed by the application of discrete cosine transformations with various coefficients selected to each level. The extracted features were then fed into the feature similarity measure, city block, for the diagnosis of breast cancer. Finally, K-nearest number was used as the classifier to classify the images as benign or malignant.

**Keywords:** Discrete cosine transform, Breast mass, Haar wavelet transform, Feature extraction, Mammogram

## 1. INTRODUCTION

Breast cancer is a type of cancer with the highest incidence rates among women. It is the most common cause of cancer-related deaths among women in many countries. Recent statistics show that breast cancer affects 1 in every 10 women in Europe and 1 in every 8 in the United States. According to the statistics of the Breast Disease Treatment Center (BDTC) in Sulaimani, breast cancer affects 1 in every 10 women.

Breast cancer affects men as well, but at a much lower rate than women. According to the statistics of the BDTC, 651 breast cancer cases were diagnosed between 2007 and 2015, which implies that nearly 81 women are affected by breast cancer every year. Mammography is known to be the most effective procedure for the early detection of breast cancer. Currently, medical knowledge is rapidly increasing to the extent that even experts have difficulties in following all the latest changes and treatments. The aim is to rapidly extend breast cancer detection techniques, which is a popular research field. The use of computer-aided diagnosis (CAD) programs in diagnostics is increasingly being used to provide a second opinion to an expert and to reduce labor intensive tasks, such as examining a large number of images daily, which can cause eye fatigue among radiologists. In addition, CAD programs can be used to decrease the gap between

Access this article online

DOI: 10.25079/ukhjse.v4n2y2020.pp178-187

E-ISSN: 2520-7792

Copyright © 2020 Bahaalddin and Ahmed. Open Access journal with Creative Commons Attribution Non-Commercial No Derivatives License 4.0 (CC BY-NC-ND 4.0).

specialized breast radiologists and general radiologists. The aim of this paper was to present a mass classification system for mammogram images based on texture extraction techniques, because texture is used to describe the variation in intensity between the pixels to supply information, making it one of the central characteristics used to analyze medical images. Therefore, it can help to distinguish and identify the objects.

## 2. RELATED WORK

Currently, medical image processing and CAD systems are hot topics in the medical research field. Many studies have been conducted to assist medical staff, including doctors, with diagnoses and decisions about treatment. Therefore, CAD systems perform the role of a second evaluator in clinical practice. In the following sections, we provide an overview of the previous studies related to this research area. CAD was invented for clinical practice in 1985 when it was introduced by the University of Chicago (Mathews, 2019). Later on, in 1998, after completing all the required trials and tests, the United States Food and Drug Administration approved the use of the CAD system in screening mammographic images. Thereafter, many CAD systems have been proposed for clinical applications (Bagchi and Huang, 2017). Murakami et al. (2013) showed that CAD can improve the rates of early detection of tumors in the size range of 1 to 10 mm by correctly identifying 83% of cases (10 out of 12) using full field digital mammography. For tumors larger than 10 mm in size, the sensitivity was increased to 92% (i.e., 129 out of 140). Various methods have been used for this purpose, including segmentation, feature extraction, classification, and similarity measurement.

The most crucial step is feature extraction because the mammograms contain content rich in textures and shapes (Kaur and Doegar, 2019). Therefore, using an appropriate feature extractor is very important, because the classification of the results are dependent on it. In the study by Abubacker et al. (2017), they used discrete wavelet transformation (DWT) and the gray-level co-occurrence matrix (GLCM) as feature extractors. The CAD proposed by Berbar (2017) focused on the feature extraction stage by designing a new hybrid feature extractor in which wavelet and contourlet features for breast mass classification were merged. In the study by Sharma and Khanna (2015), the Zernike of different

order was used as a feature extractor and a support vector machine (SVM) was used for the classification of images from the Digital Database for Screening Mammography (DDSM). The study by Mohamed et al. (2014) proposed a CAD system for microcalcification detection in which they used GLCM for the extraction of features in the ROI in combination with different classifiers for the classification, with the SVM classifier providing the best result. In the CAD designed by Mohanty et al. (2019), they used a combination of 2-dimensional block DWT (2D-BDWT) and GLCM. Chougrad et al. (2018) designed a system for breast mass detection in mammograms based on deep convolutional neural networks. The CAD system proposed by Junior et al. (2013) is based on diversity indices (i.e., Shannon-Wiener, Simpson, J, Ed, Buzas-Gibson, Camargo, Hill, McIntosh, Total Diversity, Brillouin, and Berger-Parker), geostatistical indices (i.e., Global Getis, Local Ripley K, Joint-Counte, and Nearest Neighbor), and geometric indices extracted from concave geometries (i.e., eccentricity, circularity, solidity, orientation, circular disproportion, circular density, square density, ring density, and quadratic density) to detect masses using SVM. The images were selected from the mini MIAS database and the DDSM database. Azar and El-Said (2013) used a probabilistic neural network for breast cancer classification. Gray-level first order statistics, GLCM features, and Laws texture energy measures were extracted from the original image.

A segmentation method was proposed by Gao et al. (2010) to identify suspected mass regions in mammograms. Yu and Huang (2010) proposed a wavelet filter to detect the suspicious areas using the mean pixel value. Eltoukhy et al. (2010) used multiscale curvelet transformation coefficients to propose, conduct, and evaluate a supervised classifier for mammograms. Verma et al. (2010) investigated a novel soft clustered-based direct learning classifier which creates soft clusters within a class and learns using direct calculations of weights. Moon et al. (2011) extracted 3-D features for 147 breast cancer cases (76 benign and 71 malignant breast masses), including the texture, shape, and ellipsoid fitting and, based on the segmented 3-D tumor contour, classified the tumors as benign or malignant on the basis of a logistic regression model. They used the Student t test, Mann-Whitney U test, and receiver operating characteristic (ROC) curve analysis for statistical analysis. From the AZ values of the ROC curves, the shape features (0.9138)

were better than the texture features (0.8603) and the ellipsoid fitting features (0.8496) for classification. Elfara and Abuhaiba (2013) proposed the square centroid lines gray level distribution method for feature extraction. Yufeng (2010) used the k-nearest number (KNN) method along with fuzzy c-mean clustering as a classifier.

### 3. FEATURE EXTRACTION METHODS

In this section, 2 feature extraction methods that are widely used in machine vision, namely pattern recognition and signal processing have been described.

#### 3.1. Haar Wavelet Transformation

Haar wavelet transformation (HWT) is the simplest form of DWT that operates by calculating the sums of and differences between the intensity values (Lim et al., 2001). The cropped image is taken as the input for the process and at each level the algorithm finds 4 coefficients, namely approximation, vertical, horizontal, and diagonal. The process is repeated again for the generation of approximation coefficients. In this study, it was repeated for 5 levels and for the last level coefficients (horizontal, vertical, and diagonal) to form the feature set. Figure 1 shows the first level decomposition of HWT.

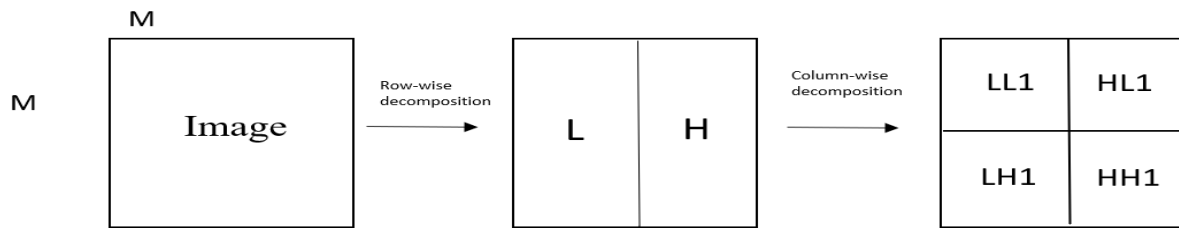


Figure 1. A 1 level 2-D DWT using HWT

#### 3.2. Discrete Cosine Transformation

Discrete cosine transformation (DCT) is one of the most important methods used in the modern computer applications field, implying that the DCT calculation is a part of many systems. It views an image as the sum of

sinusoids of varying magnitudes and frequencies. DCT tries to reduce the number of bits required to represent a signal. The algorithm works by removing redundancy in the signal using equation (Eq.) 1.

$$FF(u, v) = \frac{1}{\sqrt{MN}} \alpha(u)\alpha(v) \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y) \cos \left[ \frac{(2x+1)u\pi}{2M} \right] \cos \left[ \frac{(2y+1)v\pi}{2N} \right] \tag{Eq. 1}$$

Where:  $v = (0,1, \dots N)$ ,  $u = (0,1, \dots M)$

$f(x, y)$  is the intensity of the pixel in row  $x$  and column  $y$

$$\text{and } (\omega) = \begin{cases} \frac{1}{\sqrt{2}} & \omega = 0 \\ 1 & \text{otherwise} \end{cases}$$

### 4. THE FRAMEWORK

In this section, the proposed CAD system is discussed (Figure 2). First, the images are all processed and then stored along with the extracted textures in the database available for use at later retrieval. The following sections describe the proposed system step by step.

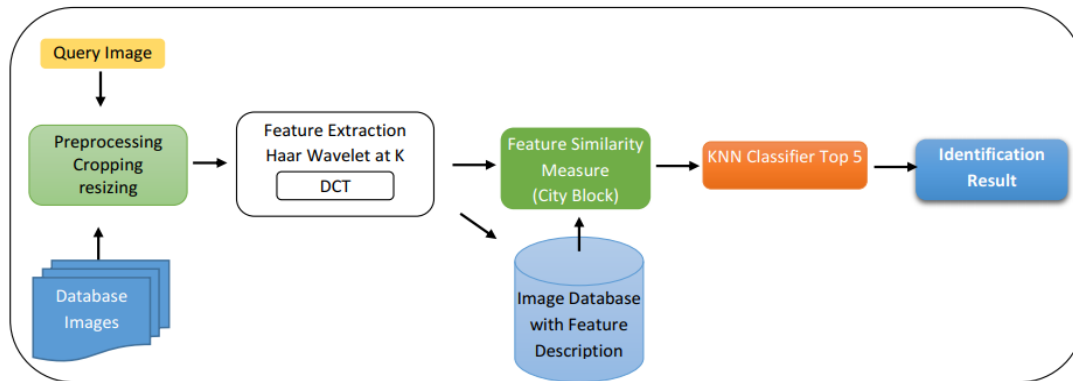


Figure 2. Proposed CAD System

#### 4.1. Image Pre-processing

The images undergo the following 2 stages of image pre-processing (See Figure 3):

- (1) First, the ROI (the region with the mass in it) is cropped from the image;
- (2) Second, the images are resized to  $128 \times 128$  pixels.

#### 4.2. Feature Extraction

In this study, a hybridization of 2 widely used texture transformation techniques, namely HWT and DCT, was proposed. The textures were extracted according to the following steps:

- 1) first level decomposition was applied, which decomposed the image into 4 subparts, namely HH, HL, LL, and LH;
- 2) the DCT was applied to the LL subpart of the image with various coefficients;

- 3) The second level decomposition was applied and again the LL subpart was decomposed into 4 subparts, namely HH1, HL1, LH1, and LL1. The third level decomposition was applied and again LL1 was decomposed into its 4 subparts, namely HL2, LL2, LH1 and HH;
- 4) The DCT was applied to the LL subpart of the image with various coefficients.

This process was continued until the fifth level of decomposition was reached. In particular, for each image, the 2-D HWT at 5 levels of decomposition was performed, thereby generating 20 subimages (i.e., LL1, HL1, LH1, HH1, LL2, HL2, LH2, HH2, . . . LL5, HL5, LH5, HH5). Figure 3 shows the image decomposition for only 3 levels. In terms of how the textural features are derived from such a decomposition, for each of the LL-band subimages, DCT is applied with different coefficients.

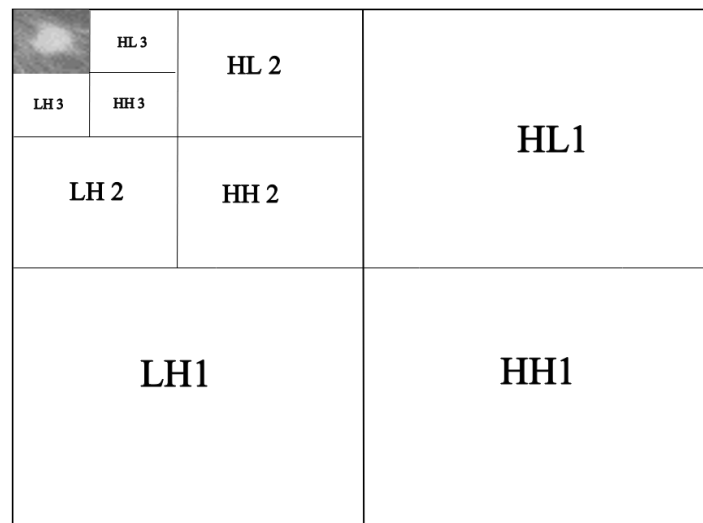


Figure 3. Decomposition of the ROI images with 2-D HWT for 3 levels

### 4.3. Classification

The images used in this study were of 2 different classes of images, namely malignant and benign mass images. At the classification stage, the images were classified as either a benign mass image or a malignant mass image. For this purpose, 2 different similarity measurements were used in 2 phases for the classification steps, which were city block at the first phase and KNN for the second phase.

#### 4.3.1. City block

The city block is a similarity measurement or distance metric that measures the path between the pixels based on 4 connected neighborhood pixels whose edges touch are 1 unit apart and those that are touching diagonally 2 units apart. The city block distance between 2 points, a and b, with k dimensions is calculated as follows:

$$\sum_{i=1}^k |a_i - b_i| \quad (\text{Eq. 2})$$

For the first phase of the classification, the city block distance is used as a texture similarity measure between the textures of the query images, with the textures of the training images saved in the database using Eq. 2.

#### 4.3.2. K-nearest number

The KNN algorithm is perhaps the simplest algorithm in machine learning. The model only consists of the training data that learns the entire training set and for prediction gives the output as the class with the majority in the 'k' nearest neighbors calculated according to some distance metric. For this study, the KNN was used to retrieve the top 5 nearest images, which were calculated according to the city block distance metric, and to choose the class with the highest probability as the result of the query image.

## 5. EXPERIMENT AND RESULTS

### 5.1. Data Set

The images used in this study were taken from the mini MIAS database (Suckling et al., 1994). The mini MIAS is a database that provides facilities for researchers that are interested in breast cancer research and is widely used. It contains 322 mammogram images of the breast. There are of 3 classes of images, namely normal, benign, and malignant and on each benign or malignant image, a description is given about the shape, margin, density, and location of the abnormality. A total of 36 images were taken for this study of which 22 contained an image of a benign mass and the other 14 contained images of a

malignant mass. The images are about  $1024 \times 1024$  pixels in portable gray map format.

## 5.2. Results and Discussion

In this section, the performance of the proposed system is examined. The experiments were conducted extensively and can be classified into the following cases based on the feature extraction process:

- (1) Extracting features from the LL subband of the first and second level of the HWT after 24

different coefficients of DCT were applied to the LL subband. In addition, the first level of the Haar wavelet is implemented on the whole mass image, and then DCT is applied on the LL subband with different coefficients. This process is repeated for the second level as well. Table 1 presents the average accuracy for the DCT coefficient separately.

Table1: Accuracy average for each DCT coefficient applied at the first and second level of the Haar wavelet				
DCT block size	LL1		LL2	
	Benign	Malignant	Benign	Malignant
1×1	75	42.857	75	42.857
2×2	85	78.571	85	78.571
3×3	85	78.571	85	78.571
3×4	90	78.5717	90	78.571
4×3	85	64.285	85	64.285
4×4	80	78.571	80	78.571
4×5	90	71.428	90	71.428
5×4	85	78.571	85	78.571
5×5	95	78.571	95	78.571
5×6	95	71.428	95	71.428
6×5	95	78.571	95	78.571
6×6	95	78.571	95	78.571
6×7	95	78.571	95	78.571
7×6	95	78.571	95	78.571
7×7	95	78.571	95	78.571
7×8	90	71.428	90	71.428
8×7	95	78.571	95	78.571
8×8	90	71.428	90	71.428
8×9	90	71.428	90	71.428
9×8	90	64.285	90	64.285
9×9	90	71.428	90	71.428
9×10	90	71.428	90	71.428
10×9	90	78.571	90	78.571
10×10	90	78.571	90	78.571

The results in Table 1 show that the accuracy averages from  $5 \times 5$  to  $7 \times 7$  and  $8 \times 8$  are almost the same and that they are better than the other remaining coefficients.

- (2) To obtain a better understanding of the impact of the hybrid HWT features with DCT, the third level of the HWT is evaluated by extracting

features from the LL subband of the third level of the Haar wavelet separately before DCT is applied. Table 2 presents the average

accuracy for each DCT coefficient on the third LL subband of the Haar wavelet.

Table 2: Accuracy averages for each DCT coefficient applied on the third level of the Haar wavelet		
LL3		
DCT block size	Benign	Malignant
1×1	75	42.857
2×2	85	78.571
3×3	85	78.571
3×4	90	78.571
4×3	85	64.285
4×4	80	78.571
4×5	90	71.428
5×4	85	78.571
5×5	95	78.57
5×6	90	71.428
6×5	90	71.428
6×6	95	78.571
6×7	95	78.571
7×6	95	78.571
7×7	95	78.571
7×8	90	71.428
8×7	90	78.571
8×8	90	71.428
8×9	90	71.428
9×8	90	71.428
9×9	90	71.428
9×10	90	71.428
10×9	90	71.428
10×10	90	71.428

As it can be seen from the results of Table 2, the accuracy of the DCT coefficients for 5×5, 6×6, 6×7, 7×6, and 7×7 on the third level of the Haar wavelet showed better results than the rest.

(3) To have a deeper look at the effect of the hybrid Haar wavelet features with DCT, the fourth

level of the HWT is evaluated by extracting features from the LL subband of the fourth level, as previously, for the Haar wavelet and before DCT is applied. Table 3 shows the average accuracy for each DCT coefficient on the fourth LL subband of the Haar wavelet.

**Table 3: Accuracy averages for each DCT coefficient applied on the fourth level of the Haar wavelet**

LL4		
DCT block size	Benign	Malignant
1×1	75	42.857
2×2	85	78.571
3×3	85	78.571
3×4	85	78.571
4×3	85	64.285
4×4	80	78.571
4×5	90	71.428
5×4	85	78.571
5×5	95	78.571
5×6	95	78.571
6×5	95	78.571
6×6	95	78.571
6×7	95	71.428
7×6	95	78.571
7×7	95	71.428
7×8	95	71.428
8×7	95	78.571
8×8	90	71.428

The results of Table 3 demonstrate that the accuracy of the DCT in the fourth level of the Haar wavelet showed better results than the previous 3 levels of the Haar wavelet.

as mentioned in the previous sections with different numbers of coefficients. Table 4 illustrates the accuracy average for each DCT coefficient applied on the fifth level of the Haar wavelet.

(4) Finally, the fifth level of the Haar wavelet is evaluated and the DCT coefficient is applied

**Table 4: Accuracy average for each DCT coefficient applied on the fifth level of the Haar**

LL5		
DCT block size	Benign	Malignant
1×1	75	42.857
2×2	85	78.571
3×3	80	78.571
3×4	85	78.571



4×3	85	78.571
4×4	80	78.571

The results shown in Table 4 indicate that the features are not accurate. Because the number of features decreased to such an extent that it was insufficient for comparison with other image features, the accuracy of the results were reduced. From the results of this study, it can be seen that the system achieved the optimum average accuracy in the application of DCT on the fourth level of the LL subband of the Haar wavelet.

### 5.3. Comparison to Existing Results

Thus far, there is no standardized data set available for mammography that take standardized environmental condition into consideration. This drawback is reflected in the difficulty to compare the proposed approach with other previous studies. However, a comparison with some of the previous studies is presented. In Table 5, the test results of the proposed scheme are compared with the results of Li et al. (2015), Dheebea and Singh (2015), and Ericeira et al. (2013). The results demonstrate that the accuracy in identification achieved with the proposed scheme outcompetes that of the previous studies.

Table 5: Comparison between the results of the proposed scheme and those of previous studies

Reference	Database size	Features method	Classifier	Accuracy (%)
Li et al., 2015	114 samples of images from the DDSM database	Texton analysis with multiple subsampling strategies	KNN SVM	85.96
Dheebea and Singh, 2015	MIAS database	GLCM	Wavelet Neural Network (WNN)	89.87
Ericeira et al., 2013	180 samples of images	Variogram function as texture feature extractor	SVM	90.26
Proposed Scheme	MIAS database	HWT+DCT	KNN	95

## 6. CONCLUSION

In this paper, a new approach for breast mass identification was presented based on DCT feature extractors in a DWT domain to classify breast mass images into one of 2 groups, namely benign and malignant. A KNN classifier was used to classify the selected feature vector. The proposed approach

outperformed the approaches of competitors in previous studies (Li et al., 2015; Dheebea and Singh, 2015; Ericeira et al., 2013). The test results illustrate that the performance of the proposed approach and the average reached 95% accuracy.

## REFERENCES

- Ericeira DR, Silva AC, et al. (2013). Detection of masses based on asymmetric regions of digital bilateral mammograms using spatial description with variogram and cross-variogram functions. *Comput Biol Med*, 43(8), 987-99.
- .Z, Y. (2010). Breast Cancer Detection with Gabor Features from Digital Mammograms. *Algorithms*, 3, 44-62.
- Apanveer kaur and Amit Doegar. (2019). Classification of Mammograms using Various Feature Extraction Methods and Machine Learning. *International Journal of Recent Technology and Engineering (IJRTE)*, 8(2), 5401-5405.
- Azar, A.T., El-Said, S.A. (2013). Probabilistic neural network for breast cancer classification. *Neural Compp. Appl.*, 23(6), 1737-1757.
- Belal K. Elfara, Ibrahim S.L. Abuhaiba. (2013). New Feature Extraction Method for Mammogram Computer Aided Diagnosis. *International Journal of Signal Proc., Image Proc., and Pattern Rec.*, 6(1).
- Berbar, M. A. (2017). Hybrid methods for feature extraction for breast masses classification. *Egyptian Informatics Journal*, 2018(19), 63-73.
- Eltoukhy, M.M., Faye, I., Samir, B.B. (2010). Breast cancer diagnosis in digital mammogram using multiscale curvlete transform. *Comput. Med. Imaging Graph.*, 34, 269-276.
- Figlu Mohanty, Suvendu Rup, Bodhisattva Dash, Banshidhar Majhi, M. N. S. Swamy. (2019). Digital mammogram classification using 2D-BDWT and GLCM features with FOA-based feature selection approach. *Neural Computing and Applications*, 1-15.
- G.B. Junior, S.V. da Rocha, M. Gattass, A.C. Silva, A.C. de Paiva. (2013). A mass classification using spatial diversity approaches in mammography images for false positive reduction. *Expert Syst. Appl.*, 40(18), 7534-7543.
- Gao, X., Wang, Y., Li, X., Tao, D. (2010). On combining morphological component analysis and concentric morphology model for mammodraphic mass detection. *IEEE Trans. Inf. Technol Biomed*, 14, 266-273.
- Hayat Mohamed, Mai S. Mabrouk, Amr Sharawy. (2014). Computer Aided Detection System for Micro-calcifications in Digital Mammograms. *Computer methods and programs in biomedicine*, 116(3), 226-235.
- Hiba Chougrad, Hamid Zouaki, Omar Alheyane. (2018). Deep Convolutional Neural Networks for Breast Cancer Screening. *Computer methods and programs in biomedicine*, 157, 19-30.
- J. Dheeba, N. Albert Singh. (2015). Computer Aided Intelligent Breast Cancer Detection: Second Opinion for Radiologists. *Computational Intelligence Applications*, 397-430.
- Li Y, Chen H, Rohde GK, et al. (2015). Texton analysis for mass classification in mammograms. *Pattern Recogn Lett*, 52(15), 87-93.
- Mathews, R.P. and Mathews. (2019). CAD in Medical Imaging: A Review of Current Trends and Future Directions. *IJERT*, 8(6), 199-2013.
- Moon WK1, Shen YW, Huang CS, Chiang LR, Chang RF. (2011). Computer-aided diagnosis for the classification of breast masses in automated whole breast ultrasound images. *Ultrasound Med Biol*, 4, 539-48.
- Nirase Fathima Abubacker, Azreen Azman, Shyamala Doraisamy, Masrah Azrifah Azmi Murad. (2017). An integrated method of associative classification and neuro-fuzzy approach for effective mammographic classification. *Neural Computing and Applications*, 28, 3967-3980.
- Ryusuke Murakami, Shinichiro Kumita, Hitomi Tani, Tamiko . (2013). Detection of Breast Cancer with a Computer-Sided Detection Applied to Full-Filled Digital Mmmography. *J Digital Imaging*, 26, 768-773.
- S. Lim, K. Lee, O. Byeon, and T. Kim. (2001). Efficient iris recognition through improvement of feature vector and classifier. *ETRI journal*, 23(2), 61-70.
- Shubhi Sharma and Pritee Khanna. (2015). Computer-Aided Diagnosis of Malignant Mammograms using Zernike Moments and SVM. *Journal of digital imaging*, 28, 77-90.
- Suckling J et al. (1994, december 11). *The Mammographic Image Analysis Society Digital Mammogram Database Exerpta Medica*. Retrieved september 1, 2019, from Mammographic Image Analysis Society: <http://peipa.essex.ac.uk/pix/mias/>
- Susama Bagchi and Audrey Huong. (2017). Signal Processing Techniques and Computer-Aided Detection Systems for Diagnosis of Breast Cancer. *Indian Journal of Science and Technology*, 10(3), 1-6.
- Verma, B., McLeod, P., Klevansky, A. (2010). Classification of benign and malignant patterns in digital mammograms for diagnosis of breast cancer. *Expert Syst. Appl.*, 37, 3344-3351.
- Yu, S.N., Huang, Y.K. (2010). Detection of microcalcification in digital mammograms using combined model-based and statistical textural feature. *Expert Syst. Appl.*, 37(7), 5461-5469.