Abstract: Breast cancer is one of common cancers in the developing countries. Detection at an early stage is very crucial for better chance of treatment. The techniques used to detect breast cancer are complex and time consuming. Computerized extraction of tumor areas from mammogram images is challenging due to shape and density of breast tumors which can sometimes surrounded by mucous (mucin). One of the challenges is to detect boundaries which can be blurred under noise factor. In this paper, we are introducing a clustering technique combined with specific structural features operations. A new noise elimination algorithm eases the noise problem and enhance the segmentation process using discrete cosine transform. Followed is the segmentation phase where classifying breast tumor from normal tumor are performed using a combined DCT and fuzzy c means algorithm. The contributions of the research are utilizing new filtering technique for noise removal. We also use Fuzzy C mean clustering algorithm using DCT information to determine the initial number of clusters. The tumor extracted segments are then transferred to the frequency domain using DCT and is used to for classifications. A test algorithm is implemented to classify new mammograms. Experimental results for all the proposed algorithms are extensively performed. The noise removal algorithm are proven robust. The experimental results of the search algorithm depicted different match and mismatch cases. 93% of the cases were a match case and predicted correctly, 5% were light cases and could not be detected from the images.

Keywords—Breast cancer; DCT; Data mining; ID3; Image processing; Classification.

I. INTRODUCTION

Breast cancer is well known to be the most widespread type of cancer in the female population [1]. Cancer cells are usually induced from normal cells due to mutation of DNA or RNA [2]. Breast cancers usually starts in the duct cells. Others begin in the lobules cells (lobular cancers [3], X-ray mammography of the patient’s breast region is a valuable tool for early detection. In this paper we are proposing a novel feature extraction technique to extract features blindly from mammograms that can differentiate between normal cells and cancerous cells in the breast. Classification will be done using data classification technique.

Since most of the tested cases which are diagnosed with breast cancer come from poor neighborhoods, a cheap solution that can be handled by non-trained personnel is required. Therefore, mass mammogram screening could take place in nurse centers and handled by our proposed system. The field of study in our research is clustering, feature extraction and classification. We are presenting a novel method for detecting breast cancer tumors based on extracting features using discrete cosine transform (DCT). The proposed method uses the DCT coefficients as the features to distinguish between Normal and Cancer tissues. The DCT coefficients were proved to be very efficient in classification. Cancer features from the mammography images. The system is trained to deduce classification rules based on the extracted features of labeled images to build a classifier system. Then, these rules will be tested by classifying new unlabeled examples. In our experiments, we will use the digitized mammograms from the DDSM: Digital Database for Screening Mammography [4]. For each image, we will compute statistical features including histograms and DCT for segmented parts of the breast.

The remainder of the paper is organized as follows: Section 2 depicts the background and the related work. Section 3 describes the methodology used and details the proposed algorithms. Section 4 presents the experimental results. Finally, the article is concluded in Section 5.

II. BACKGROUND

Breast cancer is a form of cell mutation in which one or both breasts show tumors. Detection of breast cancer if done at early stages, its therapy can become treatable. On the other hand, late detection would complicate the therapy process. Moreover, the task for detecting breast cancer as a mass screening is a more complicated task due to the fact that it is costly, time consuming, and might be impossible, if done by medical professionals. Thus, researchers from around the world worked on developing different applications for easing the process for detecting breast cancer and measuring cancer’s growth. In [5], the results of their work is integrated to a computer-aided detection mechanism to aid the radiologist to detect the breast cancer in early stages. The study in [6] study classifies the regions of interest as either benign or malignant. In [7], they used edge detection process. Before we apply edge detection algorithm, to locate the background, breast and cancer areas. The authors in [8], based their cancer tumor classifications on geometric algebra and differential evolution. In [8], feature extraction was performed by extracting first order statistical and textural features extraction methods. In [10] tumor segments were classified using SVM and Bayesian from thermogram images. In [11], texture features are imported from Gray Level Co-occurrence Matrix of different segments of a mammogram. In [12], GSP techniques such as Discrete Fourier Transform, obtained by Welch’s averaged periodogram method are utilized.

Hanan A. Hosni Mahmoud
Department of Computer Science, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, KSA, Dept. of Computer and Systems Engineering, Faculty of Engineering, University of Alexandria, Egypt
Email: hanhosni@pnu.edu.sa

Revised Manuscript Received on January 5, 2020
A Novel Clustering Algorithm Introducing New Denoising Technique

The proposed method has given satisfied results for differentiation between normal and cancerous cells.

Mammogram digital images usually have low pixel contrast, and sometimes it is hard to diagnose tumor areas due to noise. Therefore, we propose using of segmentation to lesson this problem. Medical image segmentation techniques are recorded in the literature [13]. In [14] the authors utilizes Deep Learning in health informatics. Anatomy-specific classification of medical images using deep convolutional nets was displayed in [15]. The Classification of ECG utilizing 1-D Convolutional Neural Networks was depicted in [16]. Morphological segmentation algorithms of medical images were presented in [17]. In [18], the authors introduced Deep Learning computerized medical aid for Non-Medical personal. Convolutional networks were used extensively in image segmentation, the authors in [19]-[22] utilized neural Convolutional networks. For breast cancer specific research, in [23] the authors presented a survey of classification methods for mammograms. In [24], they performed preprocessing of mammograms by proposing an adaptive median filter technique. Segmentation of thermogram images of Breast tumor are done using SVM and Bayesian classifiers [25].

Clustering algorithms such as K-means and Fuzzy C means are very helpful in the segmentation process of mammogram images. [26]. K-means algorithm computes the means of clusters, computes and calculate the distances between pixels and the clusters centroids, and decide which cluster should contain this pixel. Fuzzy C-means clustering utilizes the fuzzy set theory [26]. It was used extensively in medical images clustering. Fuzzy C-means iteratively estimates the mean, and other statistical coefficients using the maximal likelihood and clustering measures [25]. The problematic matter with the Fuzzy C-means is its sensitivity to noise. In [26], the authors proposed A noise insensitive fuzzy c-means technique for image segmentation and classification are proposed in [25]. Robust fuzzy c-means algorithm in wavelet transform domain is proposed in [24]. Noise-robust techniques for image segmentation Performance of fuzzy C-means were depicted in [25].

Several medical image segmentation techniques decrease the sensitivity of noise to some extent, the stability of segmentation is still a huge challenge. The stability of a clustering against noise is an important criteria to choose the parameters of any clustering model. Reliable clustering technique should be robust against noises. For this purpose we are proposing a new algorithm aiming at improving the stability of the clustering. We propose the DFC which is DCT-based fuzzy c-means algorithm. DFC has two phases, the first phase is an intelligent Denoising: DCT-IDCT algorithm which is using iterative DCT-IDCT technique to remove noise according to the skewness measure. The second phase is a novel DCT-based fuzzy C-means technique is used for segmentation. Finally classification algorithm in the DCT domain is presented.

III. METHODOLOGY

In our study, we aim to complete a data classification system in which it classifies any given image of the breast region to its corresponding cancer detection. Mammogram of a normal breast and cancerous breast are shown in fig. 1 and fig. 2 respectively. The methodology will be collecting images of the patients’ breasts along with their classes attached to them. The attached classes can be cancerous case or normal case. Clustering algorithm will be applied to mammograms of cancerous case to locate the tumor in the mammogram. Feature extraction of normal and cancerous cases will be done blindly. Association rules will be established. Then the collection of images will be divided into two corpus, the training corpus and testing corpus. The schematic diagram of the proposed segmentation system is depicted in Fig. 3.

The implementation of the overall system consists of three steps as follows:
1. Preprocessing step: Denoising of mammograms images are performed.
2. Training step: The features of each image in the training corpus are extracted and saved in the database along with their diagnostic label. Then generating rules from the extracted features are established.
3. Testing step: where new and unlabeled images from the testing corpus are fed into the system to be classified.

![Fig. 1. Mammogram of a normal breast.](image1)

![Fig. 2. Mammogram of a cancerous breast.](image2)
Algorithm 1: ISA (input: MAM- image, Output: X,Y)

**Description:** Image Segmentation Algorithm will take Mammogram image (MAM- image) as an input, and forms a segmented tumor image(SEG-Image) and output X, Y which are the upper left point of the square (Sq) enclosing the tumor.

**Phase 1: Removing Noise**

**Description:** Removing and minimizing the effect of noise on mammogram breast images without affecting the texture

**Algorithm 1.1. DCT-IDCT (input: MAM -Image, output: Image1)**

**Start**

step = 1

Image1 = MAM -Image

**Do**

a. Obtain the histogram \(H1\) of image1
b. Compute the degree of skewness (Skew1) of the histogram \(H1\)
c. Compute the Discrete cosine transform of image1
d. Set the lower triangular block of (width = step and height =step) of coefficients to zero
e. Compute the Inverse Discrete cosine transform to get Image2
f. Compute the histogram \(H2\) of image2
g. Compute the degree of skewness (Skew2) of the histogram \(H2\)
h. Image1 = image2
i. step = step +1

**While** (skew2 > \(\varepsilon\))

**End**

**A. Clustering**

**Phase 2: Clustering**

**Description:** it has imag1 from phase 1 as an input and output X, Y which are the upper left point of the square (Sq) enclosing the tumor

**Algorithm 1.2 DFC (input: imagel, output: X, Y)**

**Start**

a. Divide image1 into 8x8 blocks of pixels
b. Compute the Discrete cosine transform of each block. Dividing image into blocks is characterized by extracting frequency spectrum and time-space localization.
c. Form SEG-Image from the DCT blocks formed from step 2
d. Use the Ac component and the first 5 DC components of each DCT block in a zig zag manner to determine the number of clusters.
e. Segment SEG-Image utilizing the FCM algorithm.
f. Determine the segment with tumor
g. Define an N\(\times\)N square (Sq) of the tumor
h. Get the coordinate X,Y of the upper left point of Sq

**End**

Fig. 3. Schematic diagram of the segmentation technique

The proposed segmentation technique is divided into three phases.

1. **Preprocessing**
   - Extracting breast surface
   - Denoising the breast mammogram image using the proposed DCT-IDCT technique

2. **Clustering:** DCT-based fuzzy C-means technique is used for segmentation (DFC).

3. **Location of the tumor area.**

Noise in mammogram images affects the identifications of tumor area and may not properly provide true diagnosis. Skewness is taken as a metric to measure noise in images using the histogram of the image. Definition of skewness of a random variable \(X\) is given in equation 1.

\[
\text{skew}(X) = \frac{E(X-\mu^3)}{\sigma^3}
\]  

(1)

Where, \(\mu\) and \(\sigma\) are the mean and standard deviation of \(X\). \(X\) for an image represent a pixel value. Skewness value is ranged approximately between 3 down and 3 up. Skewness at or around zero indicate the absence of noise in the image [26]. In the proposed algorithm we will choose a threshold \(\varepsilon\) that defines no noise if skew is less than it. \(\varepsilon\) Is a very small number near zero.

In this paper, a proposed DCT-IDCT technique is used for removing and minimizing the effect of noise on mammogram breast images without affecting the texture. The proposed DCT-based fuzzy C-means technique uses Fuzzy-C means for locating cluster centroids and then uses the DCT algorithm to perform clustering. The detailed image segmentation algorithm (ISA) takes a mammogram image as an input and outputs a segmented image which separates the tumor segment.

Fig. 4 shows a mammogram image utilizing the FCM algorithm.

Algorithm is done through the DFC algorithm. A mammogram image from the DDSM database is shown in fig. 4, fig. 5 shows the mammogram after Denoising, and fig. 6 shows the clustering of the image.
A Novel Clustering Algorithm Introducing New Denoising Technique

1. For this block
   a. Calculate the histogram $HT$ of the block $BT$.
   b. Calculate the DCT: DCTT of block $BT$.
   c. Save the tuple $TT = <HT, DCTT, Diagnosis>$ in the database: DC

2. Repeat the following steps of each image of a normal breast in Normal-Images.
   a. Divide the breast image into $NNXN$ blocks $Bi,j$
   b. For each block in a sliding window fashion
   c. Calculate the histogram $Hi,j$ of the block $Bi,j$
   d. Calculate the DCT: DCTi,j of block $Bi,j$
   e. Calculate the average histogram $HN$ of the all block $Bi,j$
   f. Calculate the average DCT: DCTN of all block $Bi,j$
   g. Label the image with the doctor diagnosis.
   h. Save the tuple $TN = <HN, DCTN, Diagnosis>$ in the database: DN
   i. Build a decision tree DC from the databases DC and DN.

End

Algorithm 3: TEST (Input: MAMunknown, output: Diagnosis, Tumor)
Description: New and unlabeled image MAMunknown from the testing corpus are fed into the system to be classified and outputs the diagnosis: Diagnosis and a list Tumor. Tumor is empty if breast is normal and has a number $D$ of coordinates $X, y$ if breast has $D$ tumours.

Start
1. Get the unlabeled mammogram (MAMunknown) of a patient to be diagnosed
2. Normalize the size of the breast to a determined area $A$ of area $nXn$.
3. Denoise MAMunknown using DCT-IDCT algorithm producing MDNunknown.
4. Divide the MDNunknown into blocks( in sliding window fashion): $BU$ of size $NNXN$
5. For every block $BU$ : $U = 1$ to $n$
   a. $[Calculate the histogram: $HU$].$
   b. Calculate the DCTU.
   c. Classify the $BU$ and produce the diagnosis.
   d. Calculate the MSE between each component in $S$ and $S_{unknown}$
   e. Determine the nearest branches in the DT to follow
   f. Output the corresponding Diagnoses $DU$}
6. For every block $BU$ : $U = 1$ to $n$
   (IF $DU$ is tumor
   Then
   $output = tumor$
   $Xtumor = XU$
   $Ytumor = YU$
   Add $X, Y$ to List-Tumor
   IF List-Tumor is empty
   Then Diagnosis = "normal"
   Else Diagnosis = "Tumor"

End

B. The Training and Classification Algorithm
In this subsection, we are presenting the training and classification algorithm. The details of the algorithms are depicted in Fig. 7. Where, the training step aim is to generate rules from the extracted features. The testing step is the phase where new and unlabeled images from the testing corpus are fed into the system to be classified. The flow chart of the whole system is depicted in fig. 8.

Algorithm 2:
Build-Decision-Tree (input: Cancer-Images, Normal-Images output : DT)
Description: it is the training phase to generate rules from the extracted features. Features are extracted from cancerous breast images and normal images.

Start
Repeat for each image of Cancer-Images
   1. Determine a block $BT$ of size $NNXN$ and coordinate $X, Y$ which includes the tumor from DFC algorithm

End
V. EXPERIMENTAL RESULTS

A. Experimental results of DCT-IDCT Denoising Algorithm

We compared the proposed DCT-IDCT Denoising algorithm with Median, Adaptive and Average filters [27]. 100 different mammograms were used for the comparison and are extracted from the database DDSM. The comparison calculated the average PSNR and average SSIM for the 100 mammograms. Zero mean white Gaussian noise were added to mammograms. We used both PSNR and SSIM. PSNR stands for peak to noise ration which is very popular in comparing before and after images to measure the enhancement of the image after some operations. SSIM stands for structural similarity index and it measures loss of correlation, luminance distortion and contrast distortion. Values of SSIM is between [0,1] where, a value of zero means no correlation between images. The Correlation is higher as value of the metric SSIM approaches one [27]. The results of comparison are depicted in Table 1. The results of the PSNR and SSIM are shown graphically in Fig. 9 and 10. Table 1. Comparison of the proposed DCT-IDCT versus Median, Adaptive and Average filters

| Filter       | PSNR | SSIM |
|--------------|------|------|
| Median       | 35.71| 0.73 |
| Adaptive     | 35.71| 0.73 |
| Average      | 35.71| 0.73 |
| DCT-IDCT     | 35.71| 0.73 |

Fig. 9. Comparison of the PSNR of DCT-IDCT versus Median, Adaptive and Average filters

C. Experimental results of the Test Algorithm

For the training phase, the extracted features were used to build an ID3 classifier DT using Algorithm2: Build-Decision-Tree. The DT is built using : $S_{input}$, which is a set of images of patients with known diagnosis from the database DDSM [4]. : $S_{input}$ , for this experiment used 164 images. Testing of a set of 100 images from DDSM ($S_{tested}$), which are chosen blindly (unknown diagnosis) are used in the testing phase using Algorithm3: Test. Tables 2 depicts the confusion matrices of the recognition process. The effectiveness of a classifier is calculated using the number of correct and incorrect classifications [12]. The results are depicted in Table 3.

We have used several metrics as defined below.

TP (True Positive) = number of correctly predicted positive cases.
FP (False Positive) = number of incorrectly predicted positive cases.
FN (False Negative) = number of incorrectly predicted negative cases.
TN (True Negative) = number of correctly predicted negative cases.

\[
\text{accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \quad (1)
\]
\[
\text{sensitivity} = \frac{TP}{TP+FN} \quad (2)
\]
\[
\text{specificity} = \frac{TN}{TN+FP} \quad (3)
\]

The calculations of accuracy, sensitivity and specificity for data in table 2, using equations 1, 2 and 3, are depicted as follows:
- Accuracy = 90%
- Sensitivity = 92.4%
- Specificity = 80.95%

Fig. 11 shows the accuracy, sensitivity and specificity of five different run of the classifier, each run with a different set of tested data $S_{\text{test}}$. Also the Fig. shows the average accuracy, sensitivity and specificity of 50 different runs of the classifier with different set of tested data $S_{\text{test}}$. 

VI. CONCLUSION

In this paper, a decision tree (ID3) were built to detect breast cancer automatically without human intervention by generating statistical rules. A new noise elimination algorithm cases the noise problem and enhance the segmentation process using discrete cosine transform was proposed. Followed is the segmentation phase where classifying breast tumor from normal tumor are performed using a combined DCT and fuzzy c means algorithm. The experimental results of the different algorithms were depicted using different match and mismatch cases. The Denoising algorithm outperforms several Denoising filters in the literature. The Test algorithm simulation results show that 95% of the cases were a match case and predicted correctly. 5% were light cases and could not be detected from the images.

ACKNOWLEDGMENT

This research was funded by the Deanship of Scientific Research at Princess Nourah bint Abdulrahman University through the Fast-track Research Funding Program.

REFERENCES

1. L. D. Isaac and C. Suresh Kumar, "Diagnosis prognosis and prevention of breast cancer based on present scenario of human life," 2018 International Conference on Communication information and Computing Technology (ICCICT), Mumbai, 2018, pp. 1-7.
2. Sharma, Ganesh N et al. “Various types and management of breast cancer: an overview.” Journal of advanced pharmaceutical technology & research vol. 1.2 (2010): 109-26.
3. What is breast cancer? American cancer society, 2009. Sep 18, [18 Mar 2010].
4. DDSM: Digital Database for Screening Mammography, http://www.mammoimage.org/databases/
5. P. Král and L. Lenc, “LBP features for breast cancer detection,” 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, 2016, pp. 2643-2647.
6. P. Grgel, A. Sertbas, O.N. Uan, “Computer-aided classification of breast masses in mammogram images based on spherical wavelet transform and support vector machines”, Expert Systems, vol. 32, no. 1, pp. 155-164, 2015.
7. H. Al-Shamlian and A. El-Zaart, "Feature extraction values for breast cancer mammography images." 2010 International Conference on Bioinformatics and Biomedical Technology, Chengdu, 2010, pp. 335-340.
8. Li J, Hong W Sheng wu yi xue Gong Cheng xue za zhi , " Feature extraction for breast cancer data based on geometric algebra theory and feature selection using differential evolution, " Journal of Biomedical Engineering, Dec 2014, 31(6):1218-1222.
9. T. T. Huay and S. S. Maung, “Early Stage Breast Cancer Detection System using GLCM feature extraction and K-Nearest Neighbor (k-NN) on Mammography image,” 2018 18th International Symposium on Communications and Information Technologies (ISCIT), Bangkok, 2018, pp. 171-175.
10. Dinh, "Breast tumor segmentation and classification using SVM and Bayesian from thermogram images", Unique Journal of Engineering and Advanced Sciences, vol. 2, pp. 147-151 J, 2014.
11. R. Biswas, A. nath, S. Roy, "Mammogram classification using GrayLevel Cooccurrence Matrix for diagnosis of breast cancer", International Conference on Micro-Electronics and Telecommunications Engineering, pp. 161-166, 2016.
12. S. M. Naeem, M. S. Mahrouk and M. A. Eldosoky, "Detecting genetic variants of breast cancer using different power spectrum methods," 2017 13th International Computer Engineering Conference (ICENCO), Cairo, 2017, pp. 147-153.

13. R. Merjulah and J. Chandra, "Segmentation technique for medical image processing: A survey." 2017 International Conference on Inventive Computing and Informatics (ICICI), Coimbatore, 2017, pp. 1055-1061.

14. Daniele Ravi, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo, Guang-Zhong Yang, "Deep Learning for Health Informatics", IEEE Journal of Biomedical and Health Informatics, vol. 21, no. 1, pp. 4-21, 2017.

15. Holger R. Roth, Christopher T. Lee, Hoo-Chang Shin, Ari Seff, Lauren Kim, Jianhua Yao, Le Lu, Ronald M. Summers, "Anatomy-specific classification of medical images using deep convolutional nets", IEEE 12th International Symposium on Biomedical Imaging (ISBI), pp. 101-104, 2015.

16. Serkan Kiranyaz, Turkler Ince, Moncef Gabboj, "Real-Time Patient-Specific ECG Classification by 1-D Convolutional Neural Networks", IEEE Transactions on Biomedical Engineering, vol. 63, no. 3, pp. 664-675, 2016.

17. Ashwani Kumar Yadav, Ratnadeep Roy, Rajkumar Vaishali, Devendra Somwanshi, "Thresholding and morphological based segmentation techniques for medical images", International Conference on Recent Advances and Innovations in Engineering (ICRAIE), pp. 1-5, 2016.

18. Yaniv Bar, Idit Diamant, Lior Wolf, Sivan Lieberman, Eli Konen, Hagit Greenspan, "Chest Pathology Detection using Deep Learning with Non-Medical training", IEEE International Symposium on Biomedical Imaging(ISBI), pp. 294-297, 2015.

19. O. Ronneberger, P. Fischer, T. Brox, "U-net: Convolutional networks for biomedical image segmentation", International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234-241, 2015.

20. Sishchok Chaichulee, Mauricio Villarroel, Joao Jorge, Carlos Arteta, Gabrielle Green, Kenny McCormick, Andrew Zisserman, Lionel Tarassenko, "Multi-Task Convolutional Neural Network for Patient Detection and SkinSegmentation in Continuous Non-Contact Vital Sign Monitoring", IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pp. 266-272, 2017.

21. Raghav Mehta, Jayanthi Sivaswamy, "M-net: A Convolutional Neural Network for deep brain structure segmentation", IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), pp. 437-440, 2017, 2017.

22. Nima Tajbakhsh, Jae Y. Shin, Suryakanth R. Gurudu, R. Todd Hurst, Christopher B. Kendall, Michael B. Gotway, Jianming Liang, "Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?", IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1299-1312, 2016.

23. S. Kowsalya, D. S. Priya, "A survey on dignosis methods of breast cancer using Mammography", International Journal of Engineering Technology Science and Research, vol. 2, no. 4, 2016, pp. 2394-3386.

24. K. M. Sundaram, D. Sasikala, P. A. Rani, "A study on preprocessing a mammogram image using Adaptive Median Filter", International Journal of Innovative Research in Science Engineering and Technology, vol. 3, no. 3, pp. 1033-1037, March 2014.

25. Dinska, "Breast tumor Segmentation and classification using SVM and Bayesian from thermogram images", Unique Journal of Engineering and Advanced Sciences, vol. 2, pp. 147-151, 2014.

26. A.A. Rachman, Z. Rustam, "Cancer classification using Fuzzy C-Means with feature selection" in ICMSA Mathematics Statistics and Their Applications, 2016.

27. Y. Wei, J. Qu, P. Shi, H.K. Lam, A new design of H-infinity piecewise filtering for discrete-time nonlinear time-varying delay systems via T-S fuzzy affine models. IEEE Trans. Syst. Man Cybern. Syst. 47(8), 2034-2047 (2017).