International Journal of
Intelligent Systems and Applications in Engineering

ISSN:2147-6799 www.ijisae.org Original Research Paper

Detection of Micro Calcifications in Mammogram Images Using Texture Analysis and Logistic Regression

Kemal Tutuncu[1,*], Ozcan Cataltas[2]

Submitted: 16/09/2019 Accepted : 11/12/2019

Abstract: Micro-calcification in the breast is a symptom of breast cancer. Therefore, detection of micro-calcification in mammogram image plays an important role in the early diagnosis of breast cancer. Because the mammogram images are 2-dimensional, different tissues in the breast are seen on top of each other. Therefore, it is a compelling task for radiologists to identify the masses found in mammogram images. There are different methods for detecting micro-calcification in mammogram images. In this study, different image processing techniques were applied on mammogram images and a region of 80x80 pixel was taken from breast tissue. Texture features of this region were extracted using co-occurrence matrix and classified by logistic regression analysis. Classification success of 88% was achieved with the proposed model.

Keywords: *Breast Cancer, Classification, Image processing, Logistic Regression, Texture Analysis*

1. Introduction

Breast cancer is a serious danger for women's health. One in eleven women develops breast cancer at any time in their lives [1]. In the studies, 570.000 women are diagnosed with cancer every year on earth and 31% of the cancer cases are breast cancer. In women, 17% of cancer-related deaths are due to breast cancer [2]. For this reason, screening studies have started in developed countries, especially in women over the middle age. As a result of these studies, a breast x-ray (mammogram) is developed to be examined to diagnose the related disease.

Mammogram is the most commonly used method in the diagnosis of breast diseases compared to other imaging methods such as magnetic resonance imaging, ultrasonography and etc.[3]. It is widely used since mammogram is inexpensive and can show the first signs of cancer. Mammograms are formed by projecting a three-dimensional object onto a two-dimensional film [3]. Therefore, different tissues can overlap and weaken the perception of radiologists. A certain number of women die each year due to the overlooked signs of cancer on mammograms [3,4]. For this reason, the necessity of developing computerized diagnostic systems that can help radiologists to diagnose the disease by using image processing techniques on mammograms has emerged.

Radiologists look at signs or symptoms of cancer on the mammogram when evaluating the mammogram. These symptoms include calcium deposits (micro-calcifications) that appear as small bright spots, significant masses (tumors), and structural deterioration in tissue integrity of the breast. Especially micro-calcifications are one of the earliest symptoms of breast cancer [5]. Micro-calcifications appear as bright spots on mammograms and are often found in clusters in cancerous tissues.

[1]*Electrical and Electronics Engineering, Selcuk University, Konya – 42031, Turkey, ORCID ID: 0000-0002-3005-374X*
[2] *Electrical and Electronics Engineering, Selcuk University, Konya – 42031, Turkey, ORCID ID: 0000-0002-7136-6574*
** Corresponding Author Email: ktutuncu@selcuk.edu.trr*

2. Literature Review

In [6], micro-calcification in breast tissue was determined by using Deep Convolution Neural Network. They characterized micro-calcifications by descriptors obtained from deep learning and handcrafted descriptors. As a result of the experiment, they stated that the result obtained by deep learning was superior to handcrafted descriptors. In this study, the classification precision was 89.32%.

In [7], a new system for detecting micro-calcification was proposed. The discrete wavelet transforms (DWT), the contourlet transform, and the principal component analysis (PCA) were used for feature extraction. Extracted features were classified using support vector machines. According to the results, the highest classification rate was achieved by using the discrete wavelet transforms (100%). With principal component analysis and contourlet, classification rates of 73.33% and 56.67% were obtained, respectively.

In [8], an automated computer aided diagnosis (CAD) system for micro-calcification detection was proposed. Multilayer perceptron (MLP) neural network was used to detect suspicious micro-calcification regions. Statistical properties and gray level co-occurrence matrix (GLCM) of the images were used as an input to the neural network. They then implemented cascade correlation neural network (CCNN) using gray level run length matrix (GLRLM) features to increase the accuracy. According to the experimental results, the system has a sensitivity (recall) rate of 86%.

In [9], a three-stage approach to micro-calcification detection was proposed. In the first stage, it was aimed to highlight the potential regions in the breast tissue. In the second stage, the characteristic of the region of interest was extracted by using Hough transform. In the last stage, micro-calcification determination was made by using clustering algorithms. According to the study, the true-positive rate (recall) of the system was determined as 91.78%.

3. Materials and Methods

3.1. Image Database

In this study, MIAS mammogram data set was used [10]. The data set contains 330 images, including right and left mammograms of 161 people. The images have a resolution of 1024x1024 pixels.
209 of the 330 images, mammograms of patients who have no signs of disease and can be called clinically healthy are available. The anomalies in other images and their numbers are shown in Table 1.

Table 1. Mias database

***Anomaly* name**	***Number***
Asymmetry	15
Calcification	28
Circumscribed masses	25
Other, ill-defined masses	15
Architectural distortion	19
Spiculated masses	19
Normal	209

3.2. Image Filtering

An image obtained in computer environment may contain various noises. Different image processing filters are available to eliminate these noises.
One of these filters is the median filter. The new value of a pixel in the median filter is obtained by using the density values of the surrounding neighbors. Firstly, a window size is determined. The values of the neighboring pixels of that pixel are sorted from small to large. The median value is assigned as the new value of the pixel. Applying a median filter to an image makes the image blurred [11].

3.3. Local Histogram Equalization

The details in the images are often not clear due to low contrast. In such cases it is important to increase the contrast to obtain useful information from the image. Histogram equalization determines the value of each pixel of the image using the entire image, which helps to adjust the contrast of the image.
Assume that any numerical image is represented by f and that the density value of each pixel of that image ranges from 0 to L-1. Here L represents the total number of possible amplitude values in the image and is 256 for 8-bit grayscale images [12]. The normalized histogram of such an image is given by (1).

$$p_f(f_k) = \frac{n_k}{n} \quad k = 1,2,\ldots,L-1 \qquad (1)$$

In equation (1), n represents the total number of pixels in the image and nk represents the number of pixels with fk amplitude. The graph obtained by drawing pf (fk) according to fk is called image histogram. T, generalized histogram equalization, is given by (2).

$$T(f_k) = \sum_{i=0}^{k} p_f(f_i) \qquad (2)$$

The calculation of the histogram equalization transformation of a digital image consists of three basic steps. The first process is to count the density value at each pixel of the digital image and increase the amplitude index by one when the same amplitude value is found for the other pixels. As a result, the histogram of the digital image is calculated. The second step is to calculate the cumulative histogram, also called the cumulative distribution function. The stacked histogram is obtained by adding the index assigned to each color value to its predetermined index value. The last step is to multiply each value of the stacked histogram by the maximum desired value in the image and divide the result by the total number of pixels. As a result of these three processing steps, the problem of low contrast in the image is solved since the density values will be proportionally distributed in the histogram of the digital image.

3.4. Co-occurrence Matrix

The Gray Level Co-occurrence Matrix (GLCM) aims to obtain the texture information by using the relational connections of the gray values in the image. In general, it provides information about the number of repeating pixel pairs in the image [13].
Texture information of an image is obtained as follows. First, a window such as 3x3, 5x5 and etc. is defined. The gray values of the pixels in this window are scaled according to the selected number of gray values. The parameters such as the number of gray levels, direction and distance must be defined to create the GLCM. The number of gray levels determines the size of the GLCM. For example, in an image where the gray values range from 0 to 255 (2n), the size of the GLCM is scaled from 1 to 8 (n). The goal is to create a matrix that reduces the number of gray values. The generated matrix is always square and can be selected in size such as 8x8, 16x16. The direction parameter is used to define the direction of the pixel pairs. The defined pairs of pixels in directions such as 0 °, 45 °, 90 °, 135 ° and etc. are processed into the matrix [14].
The GLCM of a selected pixel is generated as shown in Fig. 1. In the next step, the probabilities are calculated by dividing the total number within each matrix. Then the statistical information such as mean, variance, entropy, homogeneity, contrast and correlation of the related pixel is calculated with the help of the window defined on the image [13].

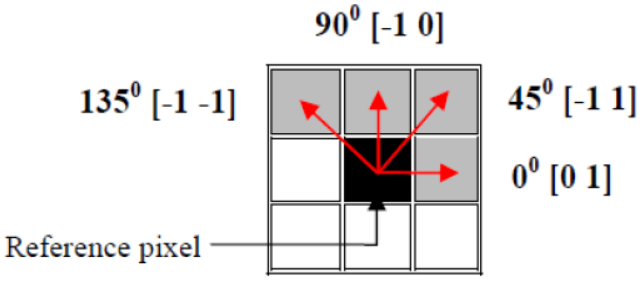

Fig. 1. Co-occurrence matrix direction for extracting texture features

Some texture properties and calculation formulas used in this study are shown in Table 2.

3.5. Logistic Regression

Binary logistic regression is a machine learning algorithm most useful when we want to model the event probability for a categorical response variable with two outcomes (yes/no, true/false, etc.). Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables. It is easy to apply mathematically and is used in many studies today because it gives meaningful results [15].

Table 2. Texture features and formulas

Feature	*Formula*
Energy	$\sum_i \sum_j [p(i,j)]^2$
Contrast	$\sum_i \sum_j (i-j)\, p(i,j)$
Correlation	$\frac{\sum_i \sum_j (i-\mu)(j-\mu)\, p(i,j)}{\sigma^2}$
Homogeneity	$\frac{p(i,j)}{1+(i-j)^2}$
Sum Average	$\sum_{i=2}^{2N} i p_{x+y}(i)$
Entropy	$-\sum_i \sum_j p(i,j) \log_2 p(i,j)$

The logistic function based on the logistic model is given in (3).

$$f(z) = \frac{1}{(1+e^{-z})} \tag{3}$$

The logistic regression formula is given in (4).

$$E(Y=1 \mid X_i) = \frac{1}{(1+e^{-(a+\beta X_i)})} \tag{4}$$

When (5) is substituted for (4), (6) is obtained.

$$z_i = a + \beta X_i \tag{5}$$

$$P_i = \frac{e^{z_i}}{1+e^{z_i}} \tag{6}$$

Equation in (6) is called the logistic distribution function, which indicates the probability of an event will result in predicted way. When the logarithms of the above equations are taken, a linear equation is obtained from the parameters and variables of the equation. The representation of the linear prediction function is as follows:

$$\ln\left(\frac{P_i}{1-P_i}\right) = \ln(e^{z_i}) = z_i = \beta_i X_i \tag{7}$$

Equation in (7) is called the logit model. Here X_i represents the variables and β_i represents coefficients.

4. Proposed Method

In this study, it was aimed to determine whether a piece of tissue obtained from mammography images is diseased or healthy using image processing techniques. For this purpose, 25 images containing diseased mass (containing benign or malignant calcification) and 25 healthy tissue images were taken from the database.

Firstly, in order to reduce the noise in the images, a median filter was applied to the entire image. As a result of the experiments, 5x5 window was selected as the filter window.

In order to reduce the effect of the background on the filtered image, thresholding was performed. The density value of the pixels whose density value is less than 10 has been changed to 0.

After this procedure, local histogram equalization was performed in order to enlarge the image contrast and make the diseased mass more apparent.

Texture samples of 80x80 pixels were taken from diseased images using the central coordinate of the diseased region in the database. Similarly, 80x80 pixel random texture samples were taken from mammogram images of healthy individuals. In this way, an image data set consisting of 50 images was formed.

The gray level co-occurrence matrix (GLCM) values of each of these images were calculated and the texture properties of each image were extracted using this matrix. Thus, a 1x6 feature set was created for each image. Logistic regression model was used to classify the obtained data set. Details of the results obtained are presented in the next section.

Fig. 2 shows the block diagram of the proposed system.

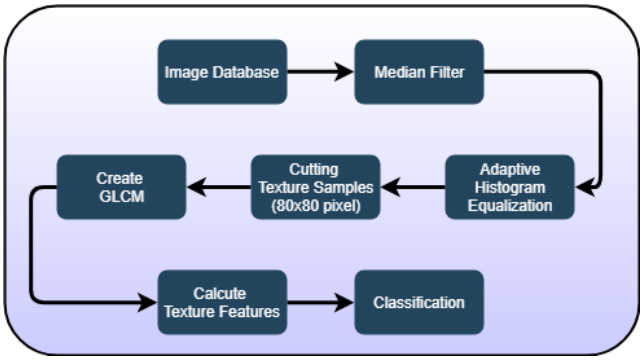

Fig. 2. Proposed method for detection micro-calcification

5. Results

Firstly, 25 mammograms containing micro-calcifications were selected. An example of the selected images is shown in Fig. 3 (a). The median filter was first applied to the selected images. While applying the median filter, 5x5 neighborhoods of the related pixels were examined and the median value was selected as the new value. The median filtered image of Fig. 3 (a) is shown in Fig. 3 (b).

Then, the local histogram equalization algorithm was applied to the related image and the obtained image is shown in Fig. 3 (c).

By using the micro-calcification coordinate in the database, 80x80 pixel regions were cut from the related images. The database containing the diseased image regions is shown in Fig. 4.

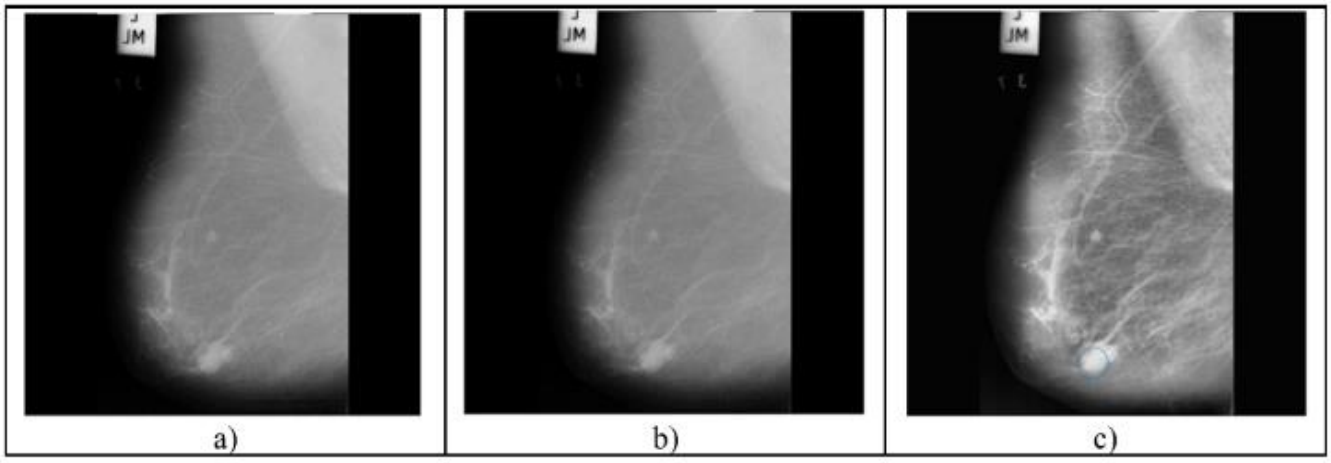

Fig. 3. a) Original mammogram image taken from mias database, b) Filtered and thresholded image in (a), c) Local histogram equalization of the image in (b)

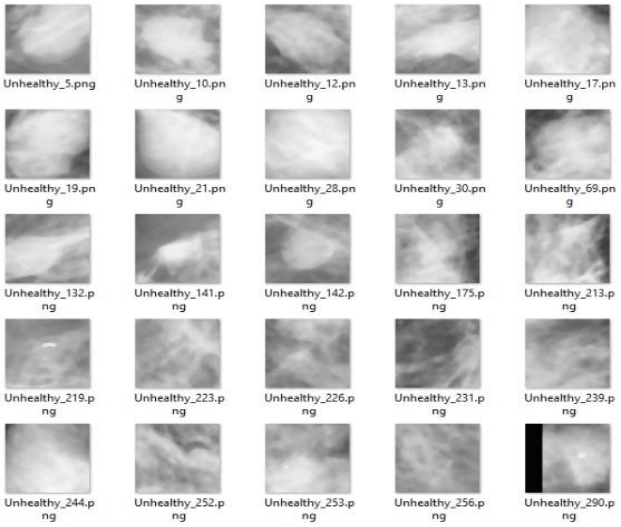

Fig. 4. 80x80 pixel images containing diseased mass

Similar procedures were applied to 25 images randomly selected from the mammogram database without any disease. 80x80 pixel regions containing breast tissue were randomly selected from these images. The database containing the healthy tissue regions is shown in Fig. 5.

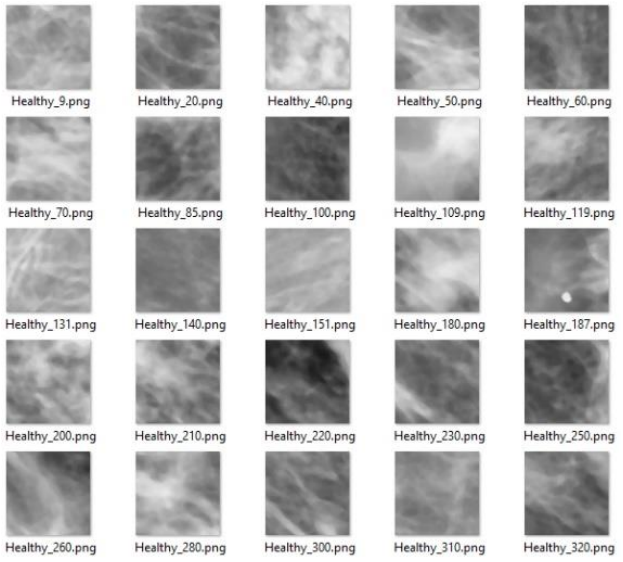

Fig. 5. 80x80 pixel images of healthy people

Co-occurrence matrices of 50 images obtained by application of image processing techniques were calculated. These matrices have been used to calculate features such as energy, contrast, correlation, sum average, homogeneity and entropy mentioned in the previous section.

Using the obtained feature values, a feature set of 1x6 for each image and a total of 50x6 was obtained. This feature set was used to construct logistic regression model. 10-fold cross-validation was used. The confusion matrix of the obtained model is shown in Table 3 and accuracy, precision, recall and f-measure values of obtained model are shown in Table 4.

Table 3. Confusion matrix

		Actual Values	
		Healthy	***Unhealthy***
Predicted Values	***Healthy***	20	4
	Unhealthy	2	24

When Table 3 is examined, it is seen that 44 of 50 breast tissue pieces are classified correctly and 6 of them are classified incorrectly. The classification success of the obtained regression model was found as 88%.

Table 4. Accuracy, precision, recall and f-measure value of obtained model

Measure	***Value***
Accuracy	88%
Precision	83.3%
Recall	90.9%
F-Measure	86.9%

6. Conclusion

In this study, image processing methods and logistic regression, which is a machine learning technique, are used for the detection of micro-calcification in mammogram images. The mammogram images with and without the diseased mass from the Mias database were first subjected to filtering and thresholding. The contrast of the image was improved using local histogram thresholding. Image sections of 80x80 pixels were taken and the co-occurrence matrices of these images were calculated and the feature set was formed. The obtained feature set was modeled by logistic regression. The

classification success of the model was 88%. As can be seen from Table 5, success of the proposed method is promising and compatible with the results obtained in the literature. It is better than [7] with respect to accuracy, better than [8] with respect to precision. In future studies, different filters can be applied to the mammogram images to increase the success of the classification.

Table 5. Literature comparison

Method	*Accuracy*	*Precision*	*Recall*
In [6]	-	89.32%	-
In [7]	PCA, 73.33% Counterlet, 56.67%	-	-
In [8]	-	-	86%
In[9]	-	-	91.78%
Proposed method	88%	83.3%	90.9%

References

[1] J. B. Li, "Mammographic image based breast tissue classification with kernel self-optimized fisher discriminant for breast cancer diagnosis" *Journal of Medical Systems*, vol. 36, no. 4, pp. 2235-2244, 2012.

[2] J. Ferlay, M. Colombet, I Soerjomataram, C. Mathers, D.M. Parkin, M. Pineros, A. Znaor, F. Bray, "Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods" *International Journal of Cancer*, vol. 144, no. 8, pp. 1941-1953, 2019.

[3] J. G. Melekoodappattu and P.S. Subbian, "A Hybridized ELM for Automatic Micro Calcification Detection in Mammogram Images Based on Multi-Scale Features" *Journal of Medical Systems*, vol. 43, no. 7, pp. 183, 2019.

[4] V. Ramachandran and V. Kishorebabu, "A Tri- State Filter for the Removal of Salt and Pepper Noise in Mammogram Images" *Journal of Medical Systems*, vol. 43, no. 2, pp. 40, 2019.

[5] A. Taliafico, G. Mariscotti, M. Durando, C. Stevanin, G. Tagliafico, L. Martino, B. Bignotti, M. Calabrese, N. Houssami, "Characterisation of microcalcification clusters on 2D digital mammography (FFDM) and digital breast tomosynthesis (DBT): does DBT underestimate microcalcification clusters? Results of a multicentre study" *Eur Radiol*, vol. 25, no. 1, pp. 9-14, 2015.

[6] H. Cai, Q. Huang, W. Rong, Y. Song, J. Li, J. Wang, J. Chen, L. Li, "Breast Microcalcification Diagnosis Using Deep Convolutional Neural Network from Digital Mammograms" *Journal of Computational and Mathematical Methods in Medicine*, vol. 2019, no. 2717454, pp. 10, 2019.

[7] M. Sharkas, M. Al-Sharkawy, D.A. Ragab, "Detection of Microcalcifications in Mammograms Using Support Vector Machine" *UKSim 5th European Symposium on Computer Modeling and Simulation*,16-18 Nov. 2011, doi: 10.1109/EMS.2011.23.

[8] B. Kurt, V. Nabiyev, K. Turhan, "An Automated Computer-Aided Detection (CADe) And Diagnosis (CADx) System For Breast Microcalcifications In Mammograms" *Selcuk Univ J Eng Sci Tech*, vol. 6, no. 3, pp. 355-376.

[9] T. M. A. Basile *et al*., "Microcalcification detection in full-field digital mammograms: A fully automated computer-aided system" *Physica Medica*, vol. 64, no. 4, pp. 1-9, 2019.

[10] J. Suckling, "The Mammographic Image Analysis Society Digital Mammogram Database Exerpta Medica" *Paper presented at the International Congress Series*, 1994.

[11] D. Vernon, "Machine Vision" Prentice-Hall, 1991.

[12] K. Joung-Youn, K. Lee-Sup, H. Seung-Ho, "An advanced contrast enhancement using partially overlapped sub-block histogram equalization" *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 4, pp. 475-484, 2001.

[13] M. Cui, S. Prasad, M. Mahrooghy, J. V. Aanstoos, M. A. Lee, L. M. Bruce, "Decision Fusion of Textural Features Derived From Polarimetric Data for Levee Assessment" *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 3, pp. 970-976, 2012.

[14] N. Severoğlu, "Mammogram images classification using Gray Level Co-occurence Matrices" *24th Signal Processing and Communication Application Conference*, 16-19 May 2016, doi: 10.1109/SIU.2016.7496106.

[15] H. Midi, S. K. Sarkar, S. Rana, "Collinearity diagnostics of binary logistic regression model" *Journal of Interdisciplinary Mathematics*, vol. 13, no. 3, pp. 253-267, 2010.