

Association Technique based on Classification for Classifying Microcalcification and Mass in Mammogram

Herwanto¹, Aniati Murni Arymurthy¹

¹Faculty of Computer Science, University of Indonesia,
Kampus Baru UI, Depok 16424, Indonesia

Abstract

Currently, mammography is recognized as the most effective imaging modality for breast cancer screening. The challenge of using mammography is how to locate the area, which is indeed a solitary geographic abnormality. In mammography screening it is important to define the risk for women who have radiologically negative findings and for those who might develop malignancy later in life. Microcalcification and mass segmentation are used frequently as the first step in mammography screening.

The main objective of this paper is to apply association technique based on classification algorithm to classify microcalcification and mass in mammogram. The system that we propose consists of: (i) a preprocessing phase to enhance the quality of the image and followed by segmentating region of interest; (ii) a phase for mining a transactional table; and (iii) a phase for organizing the resulted association rules in a classification model. This paper also illustrates how important the data cleaning phase in building the data mining process for image classification. The proposed method was evaluated using the mammogram data from Mammographic Image Analysis Society (MIAS). The MIAS data consist of 207 images of normal breast, 64 benign, and 51 malignant. 85 mammograms of MIAS data have mass, and 25 mammograms have microcalcification. The features of mean and Gray Level Co-occurrence Matrix homogeneity have been proved to be potential for discriminating microcalcification from mass. The accuracy obtained by this method is 83%.

Key words: breast cancer, classification, data mining, association rule, Region of Interest (ROI), Gray Level Co-occurrence Matrix (GLCM).

1. Introduction

Breast cancer is a chronic disease, whereas its total healing remains doubtful. It could also take a long time of medical treatment and money consuming. Breast cancer can be diagnosed with various screening mechanisms, including by mammography, a technique for observing X – Ray photo which is proven to be effectively giving a clue of abnormality in the breast. However understanding the mammogram images in the phase of diagnosing is not an easy job. There are many phases to be carried out, such as: the image preprocessing, segmentation, pattern recognition, classification and conclusion. Comprehensive knowledge in many fields of science is needed, including a vast

specialized experience, particularly to attain relevant specific characters of breast cancer. By diagnosing breast cancer effectively, it can be detected in an earlier step.

Mass and microcalcification are two confusing features present in a mammogram. Masses are identified by their shape and margin characteristics. Microcalcifications are small calcium deposit and appear as a group of bright spots in mammogram. Most of the breast cancer is detected by the presence of microcalcification [8]. Mammogram readings are usually performed by a radiologist. Variety of factors such as the poor quality of the image, eye fatigue factor, the performance of radiologist could greatly affect the diagnose results. To overcome this problem, several computer aided diagnosis systems are developed for automatic detection of breast cancer.

Breast cancer research has been continuing recently, and the challenge has been shifted from gathering data to finding hidden patterns and trends that are most relevant to cancer diagnosis. This paper presents a method for building a classification system, especially to obtain the different characteristics of mass and microcalcification using association technique based on classification. The classification process mainly consists of two phases: training phase and testing phase. In the training phase the properties of typical image features are isolated and the training samples are having their object class information.

In this study, the statistics features and the Gray Level Co-occurrence Matrix (GLCM) based on texture features are extracted. Feature selection is an important step before the process of any classification scheme. In term of the data used, the importance of data cleaning is considered. The data cleaning is applied in the context of image content mining.

This study finds the different characteristics of microcalcification and mass. The information of centroid and radius provided in MIAS database are used to find and extract the microcalcification and mass. The finding is

performed by conducting a data mining process to detect the candidate regions in mammogram into two categories: microcalcification and mass regions. This finding will help radiologists in their tasks to detect subtle abnormalities in a mammogram. Furthermore, it is also helps in making a decision whether a biopsy procedure is necessary.

A decision of the candidate regions were carefully concluded, especially in selecting the threshold value and size of the candidate regions in order to obtain the candidate regions that highly resemble microcalcification or mass. Unfortunately, there are large variations of threshold value from one image to the others, so that a constant threshold will not be good enough.

The paper is organized as follows. Section 1 is an introduction and background. Section 2 presents a review of existing techniques for mammographical feature analysis. Section 3 provides the detail proposed methodology for microcalcification and mass detection. Section 4 discusses the experiment results. Finally, Section 5 presents the concluding remarks.

2. Related Works

Various computer aided diagnosis algorithms have been proposed for analysis mammograms. These algorithms are based on extracting image features from regions of interest and estimating the probability of malignancy for a given mammogram. A variety of features and classification schemes have been used to automatically discriminate between benign and malignant mammogram. GLCM is generally used for extracting features. Khuzi *et al.* have used GLCM to identify the mass region in mammogram [2].

Many research works have used data mining technique to analysis mammogram. Researches that use data mining approach to classify can be found in [11, 12]. Most of them classify a mammogram into benign or malign, and the candidate regions are captured from the whole original image. Luiza *et al.* [11] proposed a classification method based on association rule mining. The original image was split initially in four parts, for a better localization of the region of interest. And the extracted features were discretized over an interval before organizing the transactional data set. Aswini *et al.* [4] proposed an image mining techniques using mammograms to classify and detect the cancerous tissue. The mammogram image is classified into normal, benign and malignant class and to explore the feasibility of data mining approach

Zhang *et al.* [9] have proposed, an automatic segmentation method using a modified region-growing technique. The outputs are a number of segmented regions considered

being the most suspicious. Karahaliou *et al.* [3] have investigated whether or not texture properties of the tissue surrounding microcalcification clusters can contribute to breast cancer diagnosis. The steps of the proposed method are as follows: (i) preprocess using a wavelet-based spatially adaptive method for mammographic contrast enhancement; (ii) thresholding technique to exclude microcalcification area; and (iii) texture analysis the tissue surrounding microcalcification area. The study have used four models for textural features extraction, i.e : First Order Statistics (FOS), Grey Level Co-occurrence Matrices (GLCM), Grey Level Run Length Matrices (GLRLM) and Law's Texture Energy Measures (LTEM). The tissue surrounding the microcalcification area are classified using k-Nearest Neighbor (kNN) method. The texture analysis of the tissue surrounding the microcalcification area has shown a promising result in computer-aided diagnosis of breast cancer and may contribute to the reduction of benign biopsies.

Khuzi *et al.* [2] proposed a simple Computer Aided Design (CAD) system to automatically detect areas that have a high probability of masses in digital mammogram. The texture features are extracted using GLCM. The matrices are constructed at a distance of $d = 1$ and for direction of θ given as $0^\circ, 45^\circ, 90^\circ$ and 135° . A single direction might not give enough and reliable texture information. For this reason, four directions are used to extract the texture information for each masses and non masses tiles area. The results show that the GLCM at $0^\circ, 45^\circ, 90^\circ$ and 135° with a block size of 8×8 give significant texture information to identify between masses and non-masses tissues. The detection process of the candidate regions is done based on block processing windows or tiles. The entire mammogram is divided into tiles area before feature extraction is done to each tile. Thus, the segmented image is divided into tiles with a size of 32×32 and a tile would be categorized as suspicious if its average intensity is more than 200. The intensity comparison is applied to each region in the segmented image and regions or tiles that do not fall into this category are rejected. The 32×32 windows are divided into smaller windows with size of 8×8 . Furthermore, a tile is considered to be suspicious if its average intensity within 8×8 tiles is more than 210. After that, its texture criteria are evaluated. The tiles are considered as masses if their texture criteria values are within the defined range of masses texture values.

3. Methodology

The overview processes in the proposed methodology is presented in Figure 1.

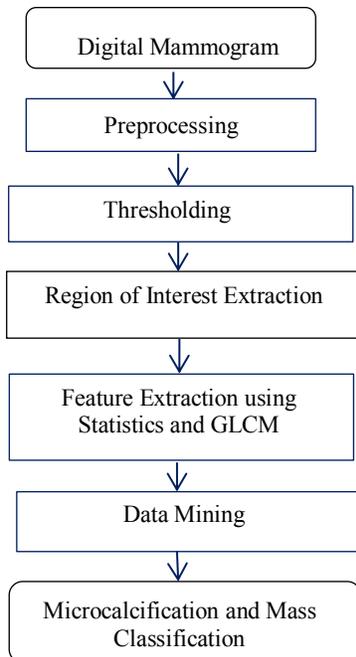


Figure 1. Proposed Methodology

3.1. Digital Mammogram

Mammogram Image Analysis Society (MIAS) database were used for experimental data. The total data consists of 322 images. The database provides some information on abnormal classes types, and the position of individual mass and microcalcification. The mammograms were downloaded from MIAS database.

3.2. Preprocessing

Mammogram is a medical image that difficult to be interpreted, therefore the phase of preprocessing is needed. It is required to increase the quality of the image in order to make the object extraction process easier. The first phase of preprocessing is removing unnecessary parts of mammogram, such as unnecessary background, before improving the quality of image. The image enhancement is used to normalize the extremely bright or dark effect. In this experiment, the method of Contrast Limited Adaptive Histogram Equalization (CLAHE) is used. Figure 2(a-b) shows the original mammogram and the result of cropping process. Figure 2(c-d) shows the morphological opening and the histogram equalization processes.

3.3. Thresholding Value

Thresholding technique is a widely used method for segmentation. It is useful for discriminating objects from background. The simple way is using a technique called global thresholding, where one threshold value is selected for the entire image. All the gray level values below a threshold value will be classified as black (0), and those above threshold value will be white (1). Figure 2(e) shows the thresholding process.

3.4. Region of Interest Extraction

The detection process is important to select suspect area which looks like microcalcification or mass. The election of ROI is based on intensity value and size of ROI, and its texture structures. The ROI election process consists of 2 phases. The first phase is to determine the suspected microcalcification or mass area based on a value of global thresholding and the size of ROI, and mark the connected components. The segmented image is divided into tiles and a tile would be categorized as suspicious if its average intensity is more than 200 [2] and the size of tile is more than 5 millimeter [14]. Figure 2(e) shows the process of marking the connected components.

3.5. Feature Extraction

Texture feature measures the variation in the surface of the image. Two types of measure are the first order and the second order statistics. The ROI feature image using the second order statistic is used as the input for GLCM feature extraction.

GLCM describe the frequency of one gray tone appearing in a specified spatial linear relationship with another gray tone within the area of investigation [1]. The co-occurrence matrix is computed based on two parameters: the relative distance d between the pixel pair d and their relative orientation ϕ . Normally, ϕ is quantized in four directions (0° , 45° , 90° and 135°).

Several statistics are derived from the co-occurrence matrix. These statistics provide information about the texture of an image. In the classical paper [17], Haralick *et al.* have introduced fourteen textural features from the GLCM and the reference stated that only six of the textural features are considered to be the most relevant. Those textural features are Energy, Entropy, Contrast, Variance, Correlation and Homogeneity. These textural features are computed based on the frequency or repetition of the pixel pair in the co-occurrence matrix.

The feature extraction phase is needed in order to create the transactional database to be mined. The extracted features

were organized in a database, which becomes the input for the mining phase. The features that are used in the experiment consist of four statistical features: mean, variance, skewness, kurtosis, and four GLCM's features: contrast, correlation, energy, and homogeneity. All the extracted features have been computed in 8x8 region of interest as shown in figure 3.

The expressions of the GLCM descriptors are as follows.

$$\text{Contrast} = \sum_{i,j} |i - j|^2 p(i, j) \quad (1)$$

$$\text{Correlation} = \sum_{i,j} \frac{(i - \mu_i)(j - \mu_j) p(i, j)}{\sigma_i \sigma_j} \quad (2)$$

$$\text{Energy} = \sum_{i,j} p(i, j)^2 \quad (3)$$

$$\text{Homogeneity} = \sum_{i,j} \frac{p(i, j)}{1 + |i - j|^2} \quad (4)$$

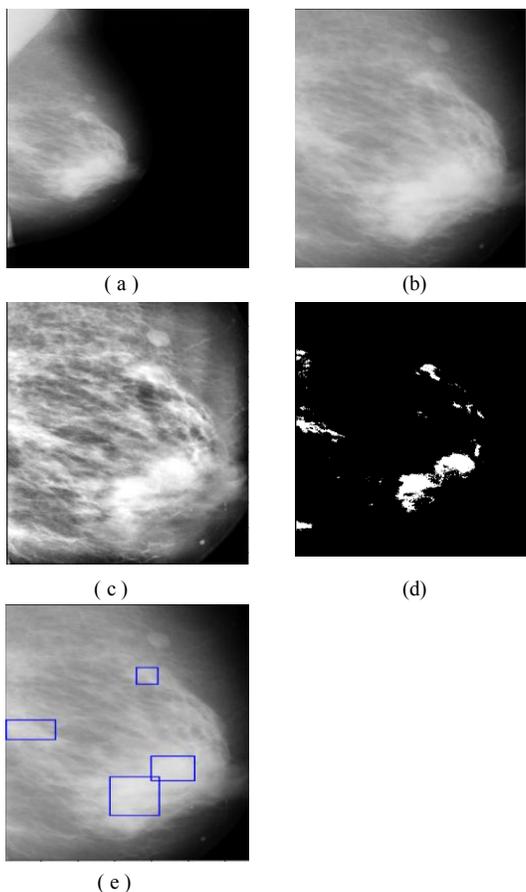


Figure 2. Results of preprocessing phase: (a) original image; (b) crop operation; (c) histogram equalization operation; (d) threshold operation; (e) marking the connected components operation.

3.6. Data Mining

Data mining is the process of discovering meaningful new correlations, patterns, and trends by digging into large amounts of data stored in warehouses. In the experiment the data mining process initially scans the pixels of digital mammograms in the database. This process seeks to discover common and indicative pixel patterns among microcalcification and mass. Using pixel values to search for microcalcification and mass is an important development because the digital mammogram contain subtleties that are not easy to perceive.

The data mining process consists of three steps: (i) the handling of missing data; (ii) the discretization step of data values; (iii) the rule mining and classification.

In this study, the discretization step is done by using k-means clustering algorithm. The k-means clustering algorithm divides a data set into a predetermined number of clusters k . The statistical mean is used as a measure for describing the average location of all members of a particular cluster. Table 1 shows an example of transactional table after the discretization process.

One of the effective algorithm of applying association rules for classification is the Classification based on Predictive Association Rules (CPAR) [6]. The algorithm CPAR is started by reading the data in the form of two-dimensional array which every column represents the attribute where the last attribute represent the object class. The input data is grouped into positive example P and negative example N related to their classes. Weight of positive example $|P|$ and negative example $|N|$ for each attribute is summed up to form PN array. The minimum weight threshold for P is calculated by multiplying the start weight of P by the Total Weight Threshold (TWT) which was set to 0.05 during experimentation. The flowchart of CPAR algorithm is shown in figure 3.

The Laplace accuracy which measures the strength of prediction is calculated using the formula :

$$L.A = (nc+1) / (ntot+f) \quad (5)$$

where f is the number of classes, $ntot$ is the total number of examples satisfying the rule's body, among which nc example belong to c , which is the predicted class of the rule [6].

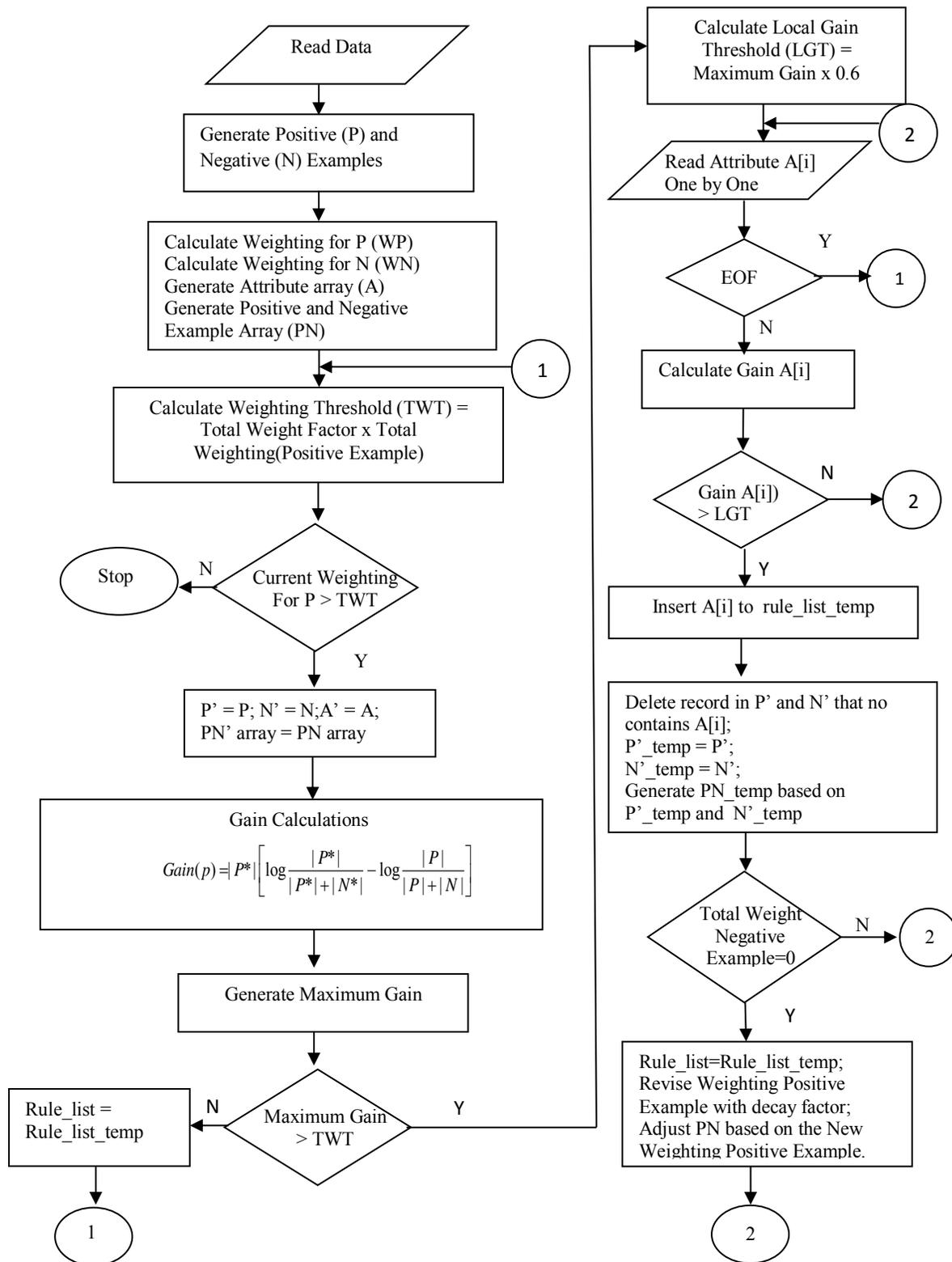


Figure 3. Flowchart of CPAR algorithm

Table 1. Feature's transactional table after discretization

Mean	Median	Variance	Kurtosis	Skewness	Contrast	Correlation	Energy	Homogeneity	Class
01	06	11	18	23	28	31	39	44	46
05	10	11	16	25	28	33	36	44	46
03	08	11	20	25	30	31	36	41	46
04	09	11	16	21	27	34	36	43	46
01	06	11	20	21	28	35	39	44	46
05	10	11	16	23	29	34	36	44	46
04	09	11	20	24	30	35	36	41	46
05	10	11	18	25	27	34	36	43	46
01	06	11	16	25	29	35	39	43	47
03	08	13	20	21	30	33	40	44	47
01	10	11	18	23	29	35	36	44	47
05	10	11	16	24	29	33	39	43	47
05	10	11	20	21	29	34	36	44	47
04	09	11	18	23	29	34	39	43	47
01	06	14	17	23	28	35	36	41	47
03	08	11	20	25	30	31	36	41	47

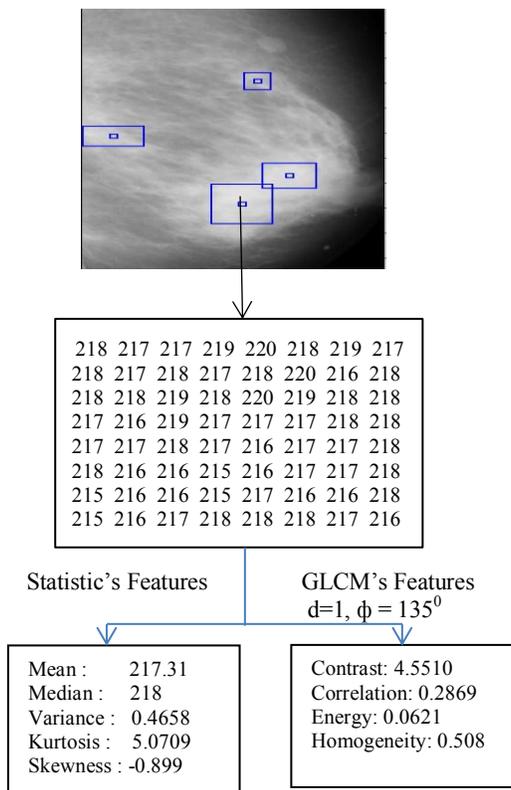


Figure 4. Feature Extraction Process

4. Experimental Results

There are 322 images of MIAS data used in the experiment. The data in the collection consists of the location information of microcalcification, mass, and their radius. It could be summarized that the data processing consists of: preprocessing, a transactional database organization, and mining the transactional database. CPAR algorithm has been used for classifying microcalcification and mass. Table 1 shows a typical example of transactional database after the discretization process. There are ten columns which every column is given a specific attribute. The columns consist of five statistical feature attributes: mean, median variance, kurtosis, skewness, and four GLCM's feature attributes: contrast, correlation, energy, homogeneity. The last attribute is the object class. The positive class of microcalcification and the positive class of mass is determined based on the information found in the MIAS. In this example, the positive class of microcalcification is 47 and the positive class of mass is 46. For the case of using 145 mammogram, the features of homogeneity and mean can be used to classify microcalcification or mass with a maximum accuracy of 83%. The testing result using the association of a selected feature set is shown in Table 2. Table 2 presents five generated rules for classification. There are no contradicting rules in either results. Microcalcification has mean value higher than mass, it means microcalcification are brighter than mass. Mass has a higher homogeneity value than

microcalcification, it means mass has more homogeneous texture than microcalcification.

Table 2. The obtained rules for classification

No	Rule	L.A
1	IF Mean between 185.74 and 203.74 Then Positive Microcalcification	0.83
2	IF Homogeneity between 0.61 and 0.65 Then Positive Microcalcification	0.80
3	IF Mean between 167.53 and 184.53 Then Positive Microcalcification	0.69
4	IF Mean between 140.82 and 165.82 Then Positive Mass	0.68
5	IF Homogeneity between 0.65 and 0.71 Then Positive Mass	0.67

5. Proposed Improvement Technique

In the previous section, it was shown how homogeneity in GLCM can be used to distinguish between microcalcification and mass. This section will describe the improvement process to increase accuracy by modifying the homogeneity computation.

The co-occurrence matrix is computed based on two parameter, which are the relative distance between the pixel pair d measured in pixel number and their relative orientation ϕ . Normally, ϕ is quantized in four directions (horizontal: 0° , diagonal 45° , vertical: 90° and anti-diagonal 135°). In this experiment, for each d , the resulting values for the four directions are averaged, and counting the number of the pixel pair that occurred in the whole image at relative distance d and direction ϕ . Table 4 shows the GLCM for $d=1$ and $\phi=0^\circ$. Table 5 shows the homogeneity matrix, where the resulting homogeneity is diagonal symmetry. Homogeneity matrix is computed with the formula 4:

The method computes weight matrix using Matlab is as follow :

```

mw = zeros(p,q);
[glcm,SI] = graycomatrix(input,'offset',[0
Dthresh],'G',[],'symmetric',true);
mglcm = glcm/sum(sum(glcm));
for j = 1 : p
for k = 1 : q
mw(j,k) = mglcm(j,k)/(1+abs(j-k));
hmat(j,k) = 1/(1+abs(j-k));
end
end
    
```

```

mw = mw;
mwt = triu(mw);
[row col] = find(hmat >= Lthresh);
position = [row col];
position1 = sortrows(position);
[uk bar] = size(position1);
sum = 0;
for su = 1 : uk
sum = sum + mw(position1(su,1),position1(su,2));
end
end
    
```

Table 3. Input Matrix

198	194	194	193	190	189	187	189
197	196	196	197	195	193	189	188
199	199	200	201	200	196	190	188
199	198	201	201	201	198	194	193
197	197	199	200	200	197	194	194
196	196	196	198	200	197	192	192
195	193	193	195	196	194	190	188
193	191	191	193	194	190	187	186

Table 4. GLCM Matrix

2	2	1	0	0	0	0	0
2	2	3	1	0	0	0	0
1	3	2	3	2	1	0	0
0	1	3	4	6	1	0	0
0	0	2	6	4	4	2	0
0	0	1	1	4	12	2	3
0	0	0	0	2	2	4	5
0	0	0	0	0	3	5	10

Table 5. Homogeneity Matrix

1.00	0.50	0.33	0.25	0.20	0.17	0.14	0.13
0.50	1.00	0.50	0.33	0.25	0.20	0.17	0.14
0.33	0.50	1.00	0.50	0.33	0.25	0.20	0.17
0.25	0.33	0.50	1.00	0.50	0.33	0.25	0.20
0.20	0.25	0.33	0.50	1.00	0.50	0.33	0.25
0.17	0.20	0.25	0.33	0.50	1.00	0.50	0.33
0.14	0.17	0.20	0.25	0.33	0.50	1.00	0.50
0.13	0.14	0.17	0.20	0.25	0.33	0.50	1.00

It is shown in Table 5 that the homogeneity values increase when becoming closer to the diagonal. The

proposed technique computed the homogeneity value based on the threshold distance p main diagonal. This means that the computation only involving pixel pair within the threshold distance greater than or equals to p . From Table 3, if the threshold distance equals to 0.2, the homogeneity value is 0.5804, but if it is not use this technique, the homogeneity value is 0,6443. Table 6 presents the accuracy result after modifying homogeneity feature.

Table 6. Improvement Result

No	Rule	L.A
1	IF Mean between 185.74 and 203.74 Then Positive Microcalcification	0.83
2	IF Homogeneity between 0.5 and 0.56 Then Positive Microcalcification	0.82
3	IF Mean between 167.53 and 184.53 Then Positive Microcalcification	0.69
4	IF Mean between 140.82 and 165.82 Then Positive Mass	0.68
5	IF Homogeneity between 0.6 and 0.66 Then Positive Mass	0.73

6. Concluding Remarks

Mammography is one of the best methods in breast cancer detection. The emphasis of using mammography is to define an area with a different appearance from other areas in the breast. The segmentation of microcalcification and mass will help radiologist to judge the likelihood of the present of cancer and to recommend whether a biopsy is necessary. This paper has proposed a method that could be used for a specific application in detection microcalcification and mass in mammogram.

References

[1] A. Baraldi, F. Parmiggiani. "An Investigation Of The Textural Characteristics Associated With GLCM Matrix Statistical Parameters", *IEEE Trans. on Geoscience and Remote Sensing*, vol. 33, no. 2, pp. 293-304, March 1995.

[2] A Mohd. Khuzi, BEng, R Besar, PhD, WMD Wan Zaki, MEngSc, NN Ahmad, PhD, "Identification of masses in digital mammogram using gray level co-occurrence matrices", *Biomedical Imaging and Intervention Journal*, 2009.

[3] Anna N Karahaliou, Ioannis S Boniatis, Spyros G Skiadopoulos, Filippo N, "A Texture Analysis Approach for Charaterizing Microcalcifications on Mammograms", School of Medicine, University of Patras, Greece, 2006.

[4] Aswini Kumar Mohanty, Sukanta Kumar Swain, Pratap Kumar Champati, Saroj Kumar Lenka, "Image Mining for Mammogram Classification by Association Rule Using

Statistical and GLCM features", *International Journal of Computer Science Issue*, 2011.

[5] Brijesh Verma, "A Neural Network Based Technique to Locate and Classify Microcalcification in Digital Mammograms", School of Information Technology, Griffith University, Gold Coast Campus, Australia, 1998.

[6] Coenen F, "The LUCS-KDD Implementations of CPAR (Classification Based on Predictive Association Rules)", Department of Computer Science The University of Liverpool, 2004.

[7] Gonzalez Rafael C, Richard E.Woods, "Digital Image Processing", New Jersey, Prentice-Hall, Inc, 2002.

[8] Gunilla Svane, E. James Potchen, Arlene Sierra, Edward Azavedo, "Screening Mammography Breast Cancer Diagnosis in Asymptomatic Women", St. Louis, Missouri 63146.

[9] Han Zhang, Say Wei Foo, S.M. Krisman & Choon Hua Thng, "Automated Breast Masses Segmentation in Digitized Mammograms", School of Electrical and Electronics Engineering, Nanyang Technological University, National Cancer Center, Singapore, 2004.

[10] Jay Bhaskar & Poonam Rani Gupta, "Design of an Automated Cancer Detection System", *Proceeding of ASCNT-2009, CDAC, India*, pp.9-5, 2009.

[11] Luiza M, O. Zaiane, A Coman, "Application of Data Mining Techniques for Medical Image Clasiification", Department of Computing Science University of Alberta, Canada, 2001

[12] Luiza M, O. Zaiane, A Coman, "Mammography Classification by an Association Rule-based Classifier", Department of Computing Science University of Alberta, Canada, 2002.

[13] Juha Koikkalainen, "Image Database in Medical Applications", Department of Engineering Physics and Mathematics, Hensinki University of Technology, 2006.

[14] Lanyi M, "Diagnosis and differential diagnosis of breast calcification", Springer – Verlag, Berlin, 1986.

[15] Le-Gal M, Chavanne G, Pellier D, "Diagnostic value of clustered microcalcifications discovered by mammography (apropos of 227 cases with histological verification and without a palpable breast tumor)", *Bull Cancer* 71(1): 57 – 64, 1984.

[16] M. M. Mokji, S.A.R. Abu Bakar, "Adaptive Thresholding Based On Cooccurrence Matrix Edge Information", *Journal of Computers*, Vol. 2, No. 8, 2007.

[17] R. Haralick, K. Shanmugam, I. instein, "Texture Features For Image Classifi-cation", *IEEE Transaction on SMC*, vol.3, no. 6, pp. 610-621, 1973.

[18] Robert J.Pizzutiello, Jr., M.S. & John E. Cullinan, R.T. (R), "Introduction to Medical Radiographic Imaging", Health Science Division, Eastman Kodak Company, 1993.