# Breast Cancer Detection in Mammogram Images Exploiting GLCM, GA Features and SVM Algorithms

Elyas Palantei, Asma Amaliah, and Indrabayu Amirullah
*Department of Electrical Engineering, Faculty of Engineering,*
*Universitas Hasanuddin (UNHAS), Makassar, South Sulawesi, Indonesia.*
*elyas_palantei@unhas.ac.id*

***Abstract*—This paper presents the novel computing algorithms to maintain the quality of mammogram images for better quality of cancer detection. The advanced algorithms were incorporated with a cancer detection unit to allow an automatic and better accuracy of tumor determination and to better classify the existing normal and abnormal breast tissues. The proposed cancer detection method consists of several steps: The first stage of the Computer Aided Detection is to maintain the images and to show the location of the abnormal tissues. The pre-processing performed on the sampled image utilized the morphology algorithm and the multi threshold segmentation to provide the appropriate tissue classification. The use of the morphology algorithm was optimized to eliminate the presence of the mammogram image label. The textural features analysis was obtained by using Gray Level Coocurance Matrix (GLCM) of four different angles, i.e. 00, 450, 900, and1350, respectively. Genetic Algorithm (GA) was optimized to find the best GLCM features, and then the results were inserted in the Support Vector Machine (SVM) training. SVM with kernel radial basis function was used to classify the patient's images as normal or abnormal breast. SVM algorithm was very important during the data training and the data testing steps. Interesting results were generated during SVM classification, which include the sensitivity rate of 69%, the precision rate of 100% and the system classification accuracy of 88.2% were taken outside from the training data and 100 % were taken inside the training data.**

***Index Terms*—Breast Cancer; Mammogram Image; ROI; GLCM; Genetic Algorithm; SVM Algorithm.**

## I. INTRODUCTION

Breast cancer is the second cancer type that causes death to woman after cervic cancer, especially in Indonesia. The risk of breast cancer will increase directly to the aging of a woman [1-10]. Tremendous advancement of the medical technology, today has triggered more applications developed for the purposes of diagnosis and treatment without surgery. The image of the X-rays, CT Scan, Medical Resonance Image (MRI) or ultrasound (USG) have been able to show the anatomical structure of the body so that the anatomical abnormalities can be detected [1-10]. The mammogram image used for the breast cancer detection is efficient. In addition to its low cost, radiation emitted by mammogram machine is quite safe for the body. Mammogram image is commonly diagnosed manually by radiologists using lampbox tool. Visually, the presence of the breast tissue abnormalities can be detected by looking at the characteristics of a mammogram image, such as the presence or the absence of the micro classification, the bumps boundary and the breast tissue distribution [4]. If the image quality is not good enough, the information will be difficult to obtain and affect the medical diagnosis. The adaptation of image processing can help improving the quality of mammogram images. Several studies related to breast cancer detection based on an abnormalities breast tissue using mammogram images had previously been carried out by Karmilasari, et al [4]. In this research, they exploited the morphology algorithm, but the result of the pre-processing on the breast image still contains the unwanted image label. Moreover, there is a need to develop an approach to extract the feature the GLCM was used. There were five different features to be considered in the image processing including the energy, the correlation, the homogeneity, the contrast, and the entropy. Each was computed with the inclination angle 00. To classify the processed image, the SVM algorithm was applied. Nevertheless, the accuracy of the breast detection using the trained data was approximately 85% and using the testing data, the system accuracy obtained was 60 %. The similar related study was performed by J. Nagi, et al. [5] who implemented a number of algorithms to improve the image quality and to determine the Region of Interest (ROI) of mammogram images. They combined the powerful morphological and Seeded Region Growing (SRG) algorithms to perform segmentation. Furthermore, to determine image ROI, the Computer Aided Detection (CAD) algorithm was used. To recognize the breast tissue abnormalities on the mammogram image, the features extraction was executed.

Another related research activity to the GLCM features extraction was released by C.V. Angkoso, et al [6]. In the process of extracting panoramic x-ray image of the human teeth jaws, they used the features of histogram functions and GLCM matrix, i.e. mean, entropy, homogeneity, variance, correlation, energy and deviation standard. As the fully completed classification based on the feature extraction value was performed, the computing accuracy of 63.33% was obtained. Meanwhile, the research activity used SVM in a mammogram image classification was previously performed also by K. Baktiar, et al. [7]. They applied two stages of classification, i.e. the common SVM algorithm and the adaptive SVM algorithm. The adaptive SVM was, in fact, the previous SVM sequential output that had been optimized. The output performance of these sequential image classifications had recorded an accuracy of 63% [7].

In this research activity, the applied morphology algorithm and the Region of Interest (ROI) detection method are expected to show the precise location of the abnormal breast tissue. To optimize the value of the feature selection, the output of the GLCM extraction step will be further processed using GA method. These sequential computing algorithms are required to maintain the classification process. The classification process itself was performed using Support Vector Machine (SVM) with kernel radial basis function. The breast cancer detection technique presented in this paper is extremely different from what has been published in [4]. In the published manuscript [4], there were only five different image features compute, but in this paper there are totally 16 different features to be computed.

## II. Image Processing for Feature Extraction and Selection

The CAD method applied in this study is described in Figure 1. The whole mammogram image processing system is basically constructed from five main processing functional units such as pre-processing, ROI detection, features extraction, features selection, and features classification. Mammogram images used is the result of x-ray image of the size 1024 x1024. The initial computing phase started by performing pre-processing and then continued by determining ROI. In the pre-processing, the morphological algorithm was implemented in 12 stages of image reconstructions. In determining ROI, threshold function was applied in the segmentation process. GLCM analyzes the texture image that passed the pre-processing step. GLCM texture analysis generates various different feature values such as contrast, correlation, energy and homogeneity. These values were then selected by the GA to find the best features value that can distinguish the characteristics of each image. The best feature values obtained by the GA were then processed by the SVM (Support Vector Machine) to be trained and tested to allow the generation of image detection decisions.

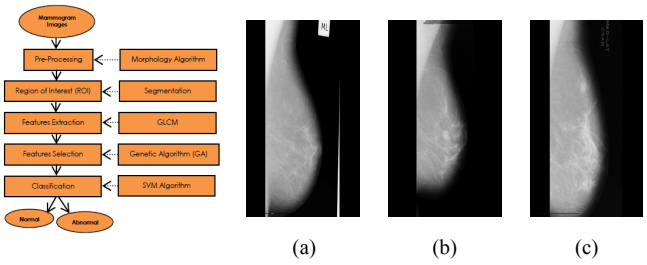


Figure 1: CAD algorithm and its corresponding three different conditions of woman breast images processed: (a) Normal (b) Benign (c) Malignant [9].

### A. Pre-Processing and Region of Interest (ROI)

Pre-processing and Region of Interest (ROI) detection is an early stage in the breast cancer detection system. In the pre-processing and ROI detection of mammogram image, there are 12 reconstruction stages must be executed. These include the original image acquisition, the foreground and background areas separation to generate the binary image, to search the largest image area of the white color presentation, the foreground (breast) area selection from the unwanted one, the object element structure fixed-up, the image reconstruction, the tumor area location searching, the unwanted parts elimination around the objects edge, masking of the tumor area, image segmentation, and the marking of the detected tumor area.

### B. Features Extraction on Mammogram Images

Texture analysis is an important part on detecting the presence of abnormalities tissue. Performance of the feature extraction algorithm in the vision machine when analyzing a variety of textures is compared to the performance of the human visual system as performing the same tasks. A vision machine system is assumed to be well running if it is able to perform a particular task as the same quality as of what human visual system do, or even it may outperform. Fig 1 above shows the normal breast, benign breast tumors and malignant breast tumors, respectively, that are difficult to distinguish through direct eye observation on the mammogram images.

In this research, GLCM was utilized for the features extraction. GLCM is a matrix. Its elements represent the number of pairs of pixels that have a certain level of brightness, where the pixel pair is separated by a distance $\boldsymbol{d}$, and has an inclination angle $\boldsymbol{\theta}$, i.e. $0^0$, $45^0$, $90^0$, and $135^0$, respectively. Co-occurence matrix is the probability of the gray levels $\boldsymbol{i}$ and $\boldsymbol{j}$ that exist from two pixels that are apart from each other at $\boldsymbol{d}$ distance and has an angle $\boldsymbol{\theta}$. Co-occurence could be defined as the joint distribution of the grayscale levels of two pixels apart at a certain distance and direction ($\Delta x$, $\Delta y$) [9].

The features that fully utilized in this research were generated by GLCM to analyze the textures such as contrast, correlation, energy, and homogeneity.

#### a. Contrast

Contrast indicates the size of the elements distribution (moment of inertia) of the image matrix. If the element is located far away from the main diagonal, then the value of contrast will be greater. The calculation result contrasts with regard to the amount of gray intensity diversity contained in the sample image.

$$Contrast = \Sigma_i \Sigma_j \, (i-j)^2 \, P_d \, (i,j) \tag{1}$$

#### b. Correlation

Correlation represents the size of linear dependence of the degree of gray image. This provides an indication of the linear structures existed in the image.

$$Correlation = \sum_i \sum_j \frac{ijP_d(i,j) - \mu_x \mu_y}{\rho_x \rho_y} \tag{2}$$

#### c. Energy

The energy measures the image gray-scale that reflects the distribution of weight uniformity and its texture.

$$Energy = \sum_i \sum_j P(i,j)^2 \tag{3}$$

#### d. Homogeneity

Homogeneity describes the structure uniformity of the image grayscales.

$$Homogeneity = \sum_i \sum_j \frac{P_d(i,j)}{1 + |i-j|} \tag{4}$$

The use of the above four features on existing functions in MATLAB program is on graycoprops. The effect of those features on the computing process is the time required during the training and the testing becomes faster [8].

### *C. Features Selection on Mammogram Images*

GLCM matrix is capable to capture the texture properties, but it could not be used directly as an analytical tool, for example, comparing two textures. These data should be extracted again to obtain the appropriate values that can be used to classify the texture. The next steps required to perform after the feature extraction is the feature selection. In this study, the features selection was performed using Genetic Algorithm based Kernel Neural Network. GA will select the overall value of the features, i.e. contrast, correlation, energy and homogeneity, by selecting the best features of the existing features values to be used as input in the SVM classification.

The selection method used was roulette machines that are well-known as stochastic sampling with replacement [9]. The computing selection algorithm using GA is illustrated in Figure 2. The detailed selection mechanism is explained below:

1. Calculate the fitness value of each individual ($f_i$ where $i$ is the individual number that start from the $1^{st}$ until the $n^{th}$.
2. Calculate the total fitness of all individuals.
3. Calculate the probability of each individual.
4. Based on step-3, calculate the quota of each individual started from 1 to 100.
5. Generate the random numbers between 1 and 100.

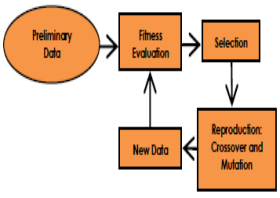


Figure 2: GA Flowchart

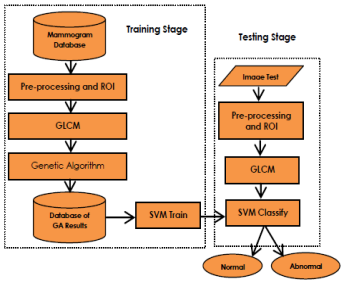


Figure 3: Block diagram of classification stages

## III. SVM CLASSIFICATION

The method of features classification applied in this study is the Support Vector Machine (SVM). The whole processes are described in Fig.3. In this classification process, the training stage is required. The process uses the extracted image features as an input data. The learning process is then generated a decision function to classify the mammogram image.

The image data exploited in this research project was obtained from the MIAS (mammographic Image Analysis Society) database [10]. The mammogram image adopted is in the form of 8-bit gray level image with the size 1024 x 1024 pixels. The pixel size is 200 microns. The image format is in the form of Portable Graymap (.pgm).

## IV. RESULTS AND DISCUSSION

The breast cancer detection mechanism was initially performed by separating the mammogram data of the left and right breasts. In order to maintain the breast image quality, each data mammogram was fed for the pre-processing step using morphological algorithm. The ROI was obtained from the segmentation process using otsu thresholding, as well as normalizing the size of all mammogram image data to 512 x 512 pixels. The significant outcome obtained from the pre-processing step is the elimination of the mammogram image label as well as the improvement of the image quality from the contaminated noise.

### *A. Application of Pre-Processing and Region of Interest (ROI)*

There are two types of images generated from the pre-processing stage, i.e. the gray and RGB images. These are illustrated in Figure 4. As clearly seen in Figure 4, the interesting information regarding the woman breast was presented. The image (a) shows the original image that has not been processed. The image (b) shows the image that marked the tumor location. This mark is the tumor coordinates obtained from the MIAS database information. The image (c) shows the gray image obtained from the segmentation results. The use of the multithreshold otsu level 8 exhibited an excellent result on distinguishing each tissue layer of the breast structure. Image (d) shows the RGB image with details that is very clearly distinguished between the healthy tissue and the infected tumor tissue. The red color is marked as the location of the abnormal tissue and the green color is characterized as the normal tissue location.

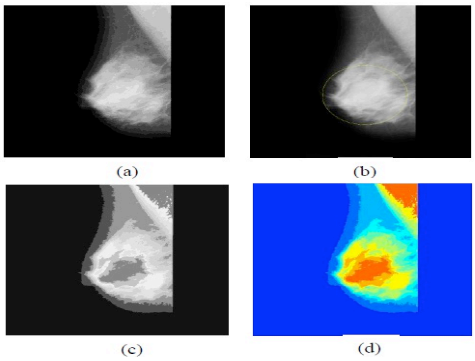


Figure 4: Various breast image representations: (a) original image (b) the image with sign of tumor location based information MIAS (c) gray image segmentation (d) RGB image segmentation.

Figure 5 shows the difference between the detection of the abnormal and normal mammogram images based on the results of the pre-processing. In Figure 5 (a) of the abnormal image, it is clearly shown the differences of the healthy tissue and unhealthy one. The healthy tissue is characterized with a bright blue color. However, the indicated tumor tissue/cancer is marked with the red color. In contrary, Figure 5 (b) of the normal image illustrates that the whole breast tissue structures are homogeneously filled with only one color, the green color. This color represents

the healthy breast tissue (there are no tumors/cancers). For the red color that appeared on the top edge of the processed mammogram image frame is regarded as the bone structure.

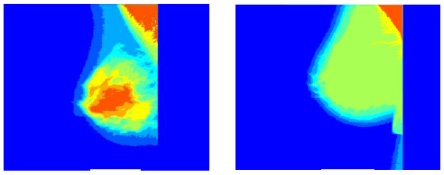


Figure 5: Mammogram image detections: (a) Abnormal image (b) Normal image (Healthy Breast).

### B. *Application of Features Extraction*

The gray image was generated from the segmentation using the multithreshold otsu algorithm then to be inputted to GLCM computing algorithm for further extraction. The output of the sequential processing generated 4 image features of four different inclination angles θ (i.e. 0°, 45°, 90°, and 135°, respectively). Therefore, a total of 16 image features must be computed sequentially. The results of the texture analysis the interval values of each features were tabulated in Table 1.

From Table 1, it can be seen that the range values obtained from the extraction overlap between the normal and abnormal images for each type of features. This, in practice, may affect the difficulty on distinguishing between the normal and abnormal tissues if the diagnosis only relied on the eye observation alone. This also means that the texture feature values almost identical on the extracted image.

### C. *Application of Features Selection*

During the computing of the mammogram image to be analyzed for the detection of the potential cancer suffered in the woman breast, the value of the extracted features GLCM is inputted to GA processing. This procedure was performed to find the best features to be used later in SVM classification.

From Figure 6, it can be seen that from the selection results generated from GA, the best fitness is found at the value 0.0123367 for the total generation of 55. The data in column 6, 8 and 12 on GLCM were the best feature that can describe the differences/ special characteristic of the overall features existed. Thus, the data presented in column 6, 8 and 12 of the GLCM matrix acted as the input for SVM classification. The meaning of the columns 6, 8 and 12 are the feature correlation at two inclination angles (45° and 135°) and also the energy feature at the inclination angles 135°, respectively.

Table 1
The Range of GLCM Features Extraction

| Features | Normal | Abnormal |
|---|---|---|
| Contrast | 0.0246 - 0.6154 | 0.0219 – 0.8826 |
| Correlation | 0.8592 – 0.9724 | 0.8415 – 0.9778 |
| Energy | 0.7250 – 0.9775 | 0.7434 – 0.9758 |
| Homogeneity | 0.9733 – 0.9993 | 0.9705 – 0.9993 |

Table 2
Confusion Matrix

| The result of Classification | Target | | FPV | TPV |
|---|---|---|---|---|
| | Abnormal | Normal | | |
| Abnormal | 9 (TP) | 4 (TN) | 0.69 | - |
| Normal | 0 (FP) | 21 (FN) | - | 1.00 |

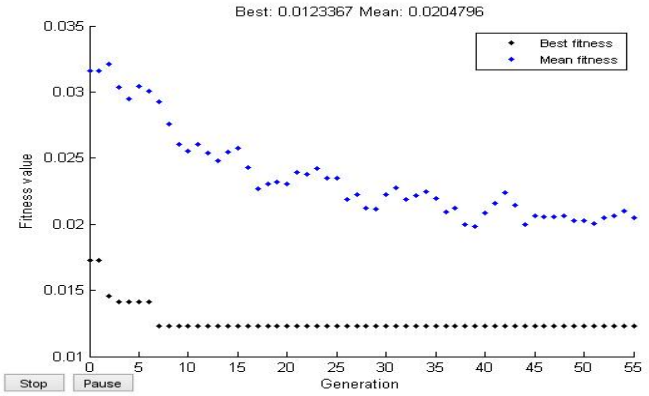


Figure 6: Graph Selection of Genetic Algorithm

### D. *Application of SVM Classification*

In the testing step after the SVM classification to the tested data, the computation to find the system precision and sensitivity was performed using a confusion matrix. For the abnormal image data, from the total of 13 existing data, only 9 images that can be classified properly. Furthermore, for as many as 21 images of the normal data, the system can detect and classify all normal image correctly with the True Positive Rate (TPV) 1:00 and the value False Positive Rate (FPV) of 0.69 was achieved, respectively. Another performance metric value, such as the precision, sensitivity and accuracy were computed using the formulas: precision = TP / (TP + FP) and sensitivity = TP / (TP + FN). The computed results using those formula exhibited 100% and 69%, respectively. Meanwhile, the accuracy of system performance of 88.2% was achieved.

## V. CONCLUSION

The systematic and very attractive computation method to process the recorded mammogram images to determine the presence of the unwanted tissue structure on the woman breast was demonstrated. The applied morphology algorithms and ROI detection of the mammogram image can significantly improve the performance of the tumor detection existed at certain area. The breast detection system performance, such as the precision and sensitivity, were computed using a confusion matrix. It has been demonstrated that for the abnormal image data, from the 13 existing data, only 9 images could be classified properly. For the total numbers of 21 normal images data, the system can detect and classify all normal image correctly. The result of the SVM classification has obtained the sensitivity rate of 69 %, the precision rate of 100 % and the accuracy of system classification of 88.2 %, respectively.

## ACKNOWLEDGEMENT

We are grateful for the generous financial research funds granted from the Ministry of Communication and Information, the Republic of Indonesia through the advanced R&D program of the telecommunication product to the Department of Electrical Engineering, UNHAS research group.

## REFERENCES


[1] Liu, S., Babbs C.F. and Delp E.J. 1998. Normal Mammogram Analysis and Recognition in Image Processing. ICIP 98, Chicago, IL, USA. Malagelada.

[2] Spandana, Rao, Kunda., Rao, Prabhakar., Jwalasrikala., 2013 .Novel Image Processing Techniques for Early Detection of Breast Cancer, Matlab and Lab view implementation. 2013 *IEEE Point-of-Care Healthcare Technologies (PHT),* Bangalore, India, 16 - 18 January.

[3] Tomar, Ranjeet., Singh, Tripty., Wadhwani, Sulochana and Badhoria, Sarita. 2009. Analysis of Breast Cancer Using Image Processing Techniques. *Third UKSim European Symposium on Computer Modeling and Simulation.*

[4] Karmilasari, Suryarini Widodo, Lussiana ETP, Matrissya Hermita, and Lulu Mawaddah, 2013. Classification of Mammogram Images Using Support Vector Machine, presented *in Asian Conference on Information System (ACIS 2013)*.

[5] Nagi, Jawad. et all. 2010. Automated Breast Profile Segmentation for ROI Detection Using Digital Mammograms Universiti of Malaya, Kuala lumpur.

[6] Angkoso, Cucun, et all. 2011. Analisa Tekstur Untuk Membedakan Kista Dan Tumor Pada Citra Panoramik Rahang Gigi Manusia, Institut Teknologi Surabaya. Surabaya.

[7] Karisma, Baktiar. Purwitasari, Diana, Yuniarti, Anny. Implementas Adaptive Support Vector Machine untuk Membantu Identifikasi Kanker Payudara. *Teknik Informatika*, ITS. Surabaya.

[8] Tony Wijanarko dan Adi Putra. 2013. Pengenalan Wajah Dengan Matriks Kookurensi Aras Keabuan dan Jaringan Syaraf Tiruan Probabilistik.Pascasarjana Universitas Diponegoro, Semarang.

[9] Stefan Lessmann, Robert Stahlbock, Sven F. Crone. 2006. Genetic Algorithms for Support Vector Machine Model Selection. *International Joint Conference on Neural Networks.* Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada. July 16-21.

[10] Mini-MIAS (Mammography Image of Analysis Society) database http://peipa.essex.ac.uk/info/mias.html