Classification of Mammogram Images using Support Vector Machine

Karmilasari† **Suryarini Widodo**‡* **Lussiana ETP*** **Matrissya Hermita**** **Lulu Mawadah** ***

† ‡* ***Faculty of Computer Science and Information Technology,
Gunadarma University, Indonesia
*Study Program of Information System, School of STMIK Jakarta STI&K, Indonesia
**Faculty of Psychology,Gunadarma University, Indonesia
{karmila,srini,matrissya}@gunadarma.ac.id
lussiana@jak-stik.ac.id
lulu_chester91@student.gunadarma.ac.id

Abstract

Breast cancer is one of the main causes of cancer death in women. Early detection is eficiently performed by using Digital mammograms. This research will be developed methods that can be used to detect abnormality of breast cancer. Step by step to classify normal and abnormal classes on digital mammogram is image acquisition which images are taken from MIAS database, find the region of interest using morphology algorithm, feature extraction using GLCM (Gray Level Cooccurrence matrix) and the last is classification process using SVM (Support Vector Machine). The accuracy of abnormalities (normal or abnormal) on digital mammogram for each testing data used in classification process have a different value, that is 85.00% percentage were taken inside from training data, 60.00% were taken outside from training data and 67.5% were taken inside and outside from training data. From these results, testing data that used in classification process affect on percentage of abnormality accuracy. For the future, is expected in this system can determine the level of severity (benign and malignant) from classification results.

Keywords: Mammogram, GLCM, SVM

1. Introduction

Breast cancer is by far the most frequent cancer among women with an estimated 1.38 million new cancer cases diagnosed worldwide in 2008 (23% of all cancers), the number of deaths by 458 and ranks second overall (10.9% of all cancers) [3]. Indonesia Health Profile 2008, published by the Ministry of Health Republic Indonesia display the data in 2004 to 2007 the number of breast cancer cases occupies the top position, followed by cervical cancer [7].

The early symptoms of breast cancer is often not recognized or perceived by the patient. Thus, many patients who seek treatment in an already severe. Breast cancer is characterized by a lump in the breast. The lumps can be benign or malignant tumors. One way should be done by women to avoid breast cancer is early detection, such as breast self-examination.

Technological developments, especially in the field of medical imaging have made the detection of cancers such as breast cancer. In which the cancer can be more easily detected, one of them with mammogram. A mammogram is a tool that can generate two-dimensional, usually 8-bit gray scale image that obtained from the x-ray of the breast patients [10]. Visually, a doctor can identify breast abnormalities by looking at the characteristics of the mammogram image. These characteristics are the left and right breasts are not symmetrical, there is a lump, there is spread of breast tissue structures and contained microcalcifications [12].

From mammogram image, the doctor performs the conventional analysis or directly diagnose cancer cells with the naked eye [10]. However, sometimes there is a mistake when analyzing (due to fatigue and various kinds of human error) as well as let loose the important things from eyesight doctor.

Based on this background, it will be

developed methods that can be used to detect the presence of breast cancer. Some of the methods developed to detect the presence of breast cancer are recognizing abnormalities on mammography images and detection of breast tumors. Step by step to classify breast cancer is image acquisition, segmentation process or find the region of interest, feature extraction, and finally, classification process.

According to Liu, structural abnormalities in mammography images can be recognized through the existence of microcalcifications, a lump and distribution lesion or nodule. Normal characteristics are shown in mammogram image that have a lower density than abnormal mammography image, there are no stains that showed micro calcifications and there is no brighter areas (higher intensity), which showed a lump [11].

To diagnose breast cancer on mammography images, then performed region of interest selection process or the suspected area. In the research Basim Alhadidi, Mohammad H.Zu'bi and Hussam N.Suleiman, breast cancer can detect from mammogram image using image processing functions : threshold, edge based and watershed segmentation. From these method shows that threshold is the fastest. But the output image from threshold is not segmented clearly than other method, edge-base and watershed segmentation consuming long time than threshold but output image is better, also threshold only segment gray level images [5].

After found the region of interest, then do extraction features. In the research Pradeep N, Girisha H., Sreepathi B. and Karibasappa K., develop a computer aided method for mass or tumor classification based on extracted features from the Region Of Interest (ROI) in mammograms. ROI has to be segmented from the digital mammogram using the Segmentation techniques. Pattern recognition in image processing requires the extraction of features from regions of the image, and the processing of these features with a pattern recognition algorithm. They consider the feature extraction part of this processing, with a focus on the problem of tumor detection in digital mammography. Parameter used consists of 12 features [13].

Y.Ireaneus Anna Rejani, Dr.S.Thamarai Selvi, proposed system focuses on the solution of two problems. One is how to detect tumors as suspicious regions with a very weak contrast to their background and another is how to extract features which categorize tumors [15].

Based on the previous research, it will created a system that can assist doctors in the diagnosis of breast cancer. With these system, is expected to classify the mammogram image into the class of normal and abnormal (malignant or benign). So it can help the doctor or radiologist in the examine mammogram as a second opinion when perform a diagnosis. The system uses Support Vector Machine classification method, by perform classification of features value from digital mammogram. Where the features are generated from the segmentation and feature extraction.

This paper is limited to digital mammogram that used secondary data obtained from MIAS mammography database and limited to diagnoses abnormality of digital mammogram in the class of normal and abnormal.The paper aims: (1) find the region of interest or suspected area from mammogram image, (2) extract features from the mammogram image (3) classify normal and abnormal classes on mammogram image (4) determine the level of accuracy from classification results and (5) To find the influence of testing data on the level of accuracy from classification results

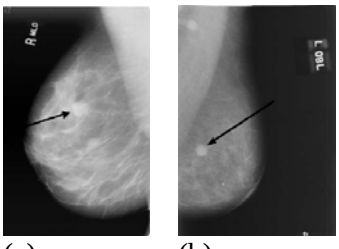
2. Literature Review

Breast cancer is a malignant neoplasm of the breast. A cancer cell has characteristics that differentiates it from normal tissue cells with respect to: the cell outline, shape, structure of nucleus and most importantly, its ability to metastasize and infiltrate. When this happens in the breast, it is commonly termed as 'Breast Cancer'. Cancer is confirmed after a biopsy (surgically extracting a tissue sample) and pathological evaluation or mammography exam [9]. Mammography is a specific type of imaging that uses a low-dose x-ray system to examine breasts. A mammography exam, called a mammogram, is used to aid in the early detection and diagnosis of breast diseases in women.

An x-ray (radiograph) is a noninvasive medical test that helps physicians diagnose and treat medical conditions. Imaging with x-rays involves exposing a part of the body to a small dose of ionizing radiation to produce pictures of the inside of the body. X-rays are the oldest and most frequently used form of medical imaging.

Two recent advances in mammography include digital mammography and computer-aided detection (1) Digital mammography, also called full-field digital mammography (FFDM), is a mammography system in which the x-ray film is replaced by solid-state detectors that convert x-rays into electrical signals. The electrical signals are used to produce images of the breast that can be seen on a computer screen or printed on special film similar to conventional mammograms. (2) Computer-aided detection (CAD) systems use a digitized mammographic image that can be obtained from either a conventional film mammogram or a digitally acquired mammogram. The computer software then searches for abnormal areas of density, mass, or calcification that may indicate the presence of cancer. The CAD system highlights these areas on the images, alerting the radiologist to the need for further analysis [2].

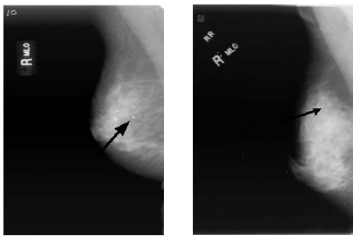
As we can see from the mammograms and diagrams, the breast tissue comprises of small intricate structures so pathology can easily be overlooked, especially if the film quality is not good. Not all structures are cancers. Many structures or macro (big) calcifications are often benign. Some basic forms of pathology and morphology presentations are : (A) Masses, differ from Densities because masses are seen on two views whereas densities are seen on one view only. Hence, two views of each breast to identify this abnormality. Masses with smooth rounded edge is generally a fluid-filled cyst that can be confirmed by an ultrasound and aspirated to relieve pain for the woman. A mass could be palpable (can feel it during a physical breast exam) depending on the size and proximity to the skin surface. Sometimes, it is very diffcult for radiologists to differentiate between a benign and a malignant mass off mammograms (as in Figure 1), so additional imaging modalities and/or biopsy may be required.



(a) (b)

Figure 1. Examples of (a) Benign Mass and (b) Malignant Mass

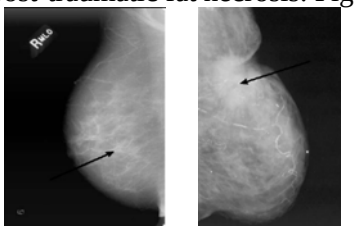
(B) Micro Calcifications, Calcifications are small calcium deposits that can be detected on a mammogram. Minute calcifications are called micro calcifications and bigger ones are called macro calcificiations. Artifacts on mammograms due to specs of dust may look like micro calcifications, but the difference is that these specs are bright and shiny whilst a micro calcification looks milky white.



(a) (b)

Figure 2. Examples of (a) Benign Calcification and (b) Malignant Calcification

(C) Spiculated lesions, This is by far the most definitive way to detect cancer. As a cancer cell proliferates, it shows up as a star-shaped or stellate lesion, with spiky lines radiating in all directions from a central region. A white star shape is characteristic of a malignant stellate lesion whereas the black star indicates a radial scar and post-traumatic fat necrosis. Figure 3.



(a) (b)

Figure 3: Examples of (a) Benign Spiculated and (b) Malignant Spiculated

The Mammographic Image Analysis Society (MIAS) is an organisation of UK research groups interested in the understanding of mammograms and has generated a database of digital mammograms. The database contains 322 digitised films and is available on 2.3GB 8 mm (ExaByte) tape. It also includes radiologist's "truth"-markings on the locations of any abnormalities that may be present. The database has been reduced to a 200 micron pixel edge and padded/clipped so that all the images are 1024x1024 [6]. In The MIAS database, mammogram image is divided into three classes,

that is glandular dense, fatty, and fatty glandular. In each class are subdivided into image of normal, benign and malignant. This is example of mammogram image from MIAS as shown in Figure 4.

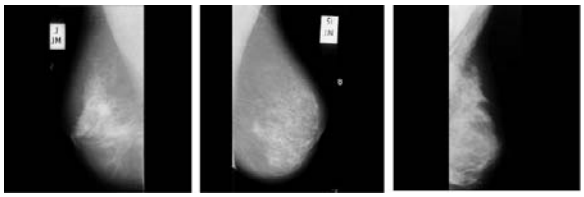


Figure 4. Mammogram Image From MIAS Dense Glandular (a) Benign (b) Malignant (c) Normal [6]

Graycomatrix creates the GLCM (Grey Level Co-occurrence Matriks) by calculating how often a pixel with gray-level (grayscale intensity) value i occurs horizontally adjacent to a pixel with the value j [1]. Graycomatrix calculates the GLCM from a scaled version of the image. By default, if I is a binary image, graycomatrix scales the image to two gray-levels. If I is an intensity image, graycomatrix scales the image to eight gray-levels. We can specify the number of gray-levels graycomatrix uses to scale the image by using the 'NumLevels' parameter, and the way that graycomatrix scales the values using the 'GrayLimits' parameter.

The basic concept of Support Vector Machine (SVM) is actually a harmonious combination of computational theories that have been there before, such as margin hyperplane (Duda and Hart 1973, Cover 1965, Vapnik 1964, etc). A support vector machine constructs a hyperplane or set of hyperplanes in a high or infinite-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

3. Proposed Method and Discussion

The stages of breast cancer classification both normal and abnormal, which is divided into two processes, namely training process and testing process. Training process and testing process have the same stage. The first stage is image acquisition and then the second stage is do segmentation by finding the region of interest (suspected area of cancer). The third stage is extraction the features using GLCM. The fourth stage is perform mammogram image classification and the last is the result. These stage can be seen in Figure 5.

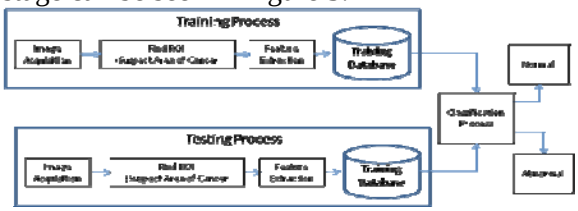


Figure 5. Architecture Design

3.1. Image Acquisition

The required data in this research are digital mammography image data that have normal case and abnormal case (benign breast cancer and malignant breast cancer). Based on these data requirements, the data used in this study is a standard data that are downloaded from the site abascus Mammography Database. In The MIAS database, mammogram image is divided into three classes, that is glandular dense, fatty, and fatty glandular. In each class are subdivided into image of normal, benign and malignant. Based on the data requirements that necessary to experiment, the sample data collected are as follows:

For training process, used 60 images. The image is consists of 30 normal images (10 dense glandular, 10 fatty, and 10 fatty glandular) and 30 abnormal images (15 benign and 15 malignant). Both of benign and malignant is taken 3 images from dense glandular, 3 images from fatty, and 4 images from fatty glandular.

For testing process, mammogram images performed in 3 cases. The first case is data used are taken inside from training data. The second case, data used are taken outside from training data and the third case, data used are taken inside and outside from training data. The data used for each case is 40 mammogram images. It consists of 20 normal images and 20 abnormal images (10 benign,10 malignant).

Image size used is 1024 x 1024 that means the image consists of 1024 columns and 1024 rows of pixels. So the total image consisted of 50873423 pixels, and the image has a PGM format that requires 8 bits per pixel with the

number of possible is 28 = 256 colors and possible color 0 (min) to 255 (max). The results can be seen in Figure 6. It is the original image from mammogram data (mdb028).

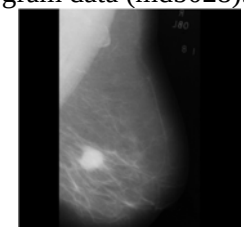
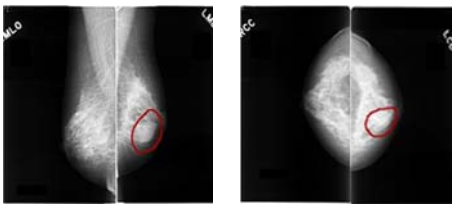


Figure 6. Original Mammogram Image (malignant mdb028.pgm)

3.2. Finding Region Of Interest (ROI)

A radiologist or doctor visually identify breast abnormalities by comparing mammography image taken from various points of view of image acquisition. Abnormalities were identified based on the characteristics of image, such as the left and right breast is not symmetrical. In addition, there is a similar shape with a circle that looks like a lump or there are areas where the image appears brighter that show the area has a higher intensity than the surrounding area as shown in Figure 7.



(a) (b)

Figure 7. Abnormal Mammography Image That Identified by Doctors

Visually, doctor recognize the left breast had abnormal characteristics. Therefore, doctor manually marked the areas that have abnormal characteristics as the suspected area of cancer (see Figure 7). In this research, The process of finding region of interest or suspected area is done by using morphology algorithms as shown in Figure 8. Morphology algorithm is used because it is able to separate abnormal area from normal area. Morphology operation used in this research is creation of structural elements operation, imtophat, imbothat, imsubtract, imcomplement, imextended, and imimposed.

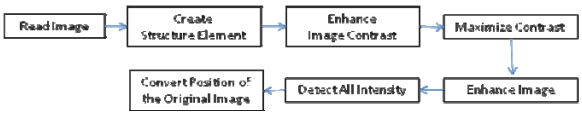


Figure 8. The Stages of Finding The ROI

3.3.2 .Create Structure Element (STREL)

The next step is create structure element, because it will apply dilation and erosion operations and it is an essential part of doing this is to create the structuring element that is used to probe the input image.

The structure element is a matrix consisting of only 0's and 1's that can have any arbitrary shape. Dilation and erosion functions is used to accept structure element objects and this objects is called STRELs. And strel function is used to create disk shape. Structural element is an operation that determines the shape and size of the addition and subtraction of pixels. Disk shape selected with consideration, every edge and curve of the object can be detected without none have passed. The size enlarged and shrinking expressed are determined by 10 pixels, because it can minimize the appearance of new information or missing information. Creation of structural elements are required for the next step segmentation using morphological operations.

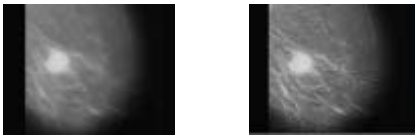
3.3.3. Enhance The Image Contrast

After structure of element is created, then performed the operation using imtophat and imbothat functions to enhance the image contrast. The operation was performed in order to displays detailed information from the invisible image.

3.3.4. Maximize The Contrast Between The Object and The Gaps

Next, to maximize the contrast between the objects and the gaps that separate them from each other by using subtract functions that added the top-hat image to the original image and then subtracts the bottom-hat image.

To maximize the contrast between the objects and the gaps that separate them from each other by using subtract functions that added the top-hat image to the original image and then subtracts the bottom-hat image. The result of maximize the contrast between the object and the gaps can be seen in Figure 9.



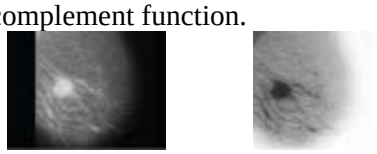
(a) (b)

Figure 9. (a) Original Image and (b) Maximize The Contrast Between The Object and The Gaps

3.3.5 Enhance The Image

The next step is convert object to photographic negative. It is particularly useful for enhancing white or gray detail embedded in a large, predominantly dark region.

The result of enhance the image can be seen in Figure 10. Enhance the image do with convert objects of Interest by watershed transform that detects intensity valleys in the image and also do enhanced the image by highlighting the intensity valleys using the imcomplement function. The negative of an image can be obtained with imcomplement function.



(a) (b)

Figure 10 (a) Maximize The Contrast and (b) Complement of Enhanced Image

3.3.6 Detects All The Intensity Valleys with Extended-minima Transform

After that, detects all the intensity valleys deeper than a particular threshold with the imextendedmin functions. This function used to mark areas that have a region minima and region maxima. The output of this function is a binary image.

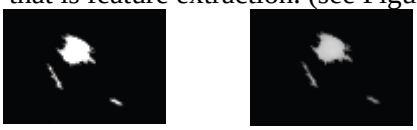
This is the result (see Figure 11) of detects all the intensity valleys deeper than a particular threshold with the imextendedmin functions. The output of this function is a binary image. The result shows that region minima marked with black color and region maxima marked with white color.

(a) (b)

Figure 11. (a) Complement of Enhanced Image and (b) Extended Minima Image

3.3.7 Convert The Position of The Original Image Intensity

Program above describe part of code to convert pixel values of Iemin (which has been done before) with binary values of 0 and 1. Conversion of binary value is conducted by changing the pixel value matrix 1 to be 0 and if the pixel is not 1 then converted into 1. Furthermore, the value is stacked into the original image. Where the position of the value 0 is converted into pixel value of original image and if the pixel is 1 then converted into 0. Therefore, the position of region of interest or segmentation found. This is the result segmentation mammogram image or Region of Interest (ROI). The segmentation results will be used for the next step that is feature extraction. (see Figure 12).



(a) (b)

Figure 12. (a) Extended Minima Image and (b) Region Of Interest

4.3 Feature Extraction

Feature extraction is very important part of pattern classification to identify texture in image, performed modeling texture as a two dimensional array gray level variation. This array called Gray Level Co-occurrence matrix. Five statistical measures such as contrast, correlation, energy, entropy, and homogeneity are computed based on GLCM. This research used direction 00 and distance 2. Feature extraction results below are training data by using 60 data (normal and abnormal). This data is a combination character of the background tissue, namely fatty, fatty glandular, and dense glandular.

The results of range between normal and abnormal mammogram image is obtained by finding the minimum and maximum values of all the data, then find range by maximum value minus minimum value and divide by the number range to be made. This research will be made 3 ranges, it because total data from all range can more easily be seen most clearly the difference. After getting the range value, then create the first range by adding the range value with the minimum value. Then, create the second range by

adding the first value with range value. The third value created by adding the second range value with range value. After that, from all the feature extraction data that have been obtained seen how the highest number of the first range, the second range and the third range. The highest number of each range used to see a significant difference between normal and abnormal from mammogram data.

Table 1. The Results Of Range Between Normal And Abnormal Mammogram Image

Features	Normal	Abnormal
Contrast	0.040565-0.238626	0.232185-0.439165
Correlation	0.896396-0.983610	0.928927-0.980383
Energy	0.868832-0.987895	0.745674-0.858751
Entropy	0.050317-0.304292	0.342372-0.593927
Homogeneity	0.989306-0.998193	0.980197-0.989456

SVM (Support Vector Machine) is a machine learning method that works on the principle of structural risk minimization in order to find the best hyperplane that separates two classes (normal and abnormal). The data used for this SVM is training data and testing data. In this research, testing data are divided into 3 groups. The first group, testing data were taken inside from training data. The second group, testing data were taken outside from training data. And the third group, testing data were taken inside and outside from training data. Testing data used are 40. Grouping is performed to see the accuracy from each group. The process of classification is performed to classify category of normal and abnormal from mammogram image.

From the classification results using testing data were taken inside from training data, percentage accuracy abnormalities value (normal or abnormal) is 85.00 %. From the classification results using testing data outside from training data, percentage accuracy abnormalities value (normal or abnormal) is 60.00%. From the classification results using testing data were taken inside and outside from training, percentage accuracy abnormalities value (normal or abnormal) is 67.50%. The accuracy of abnormalities (normal or abnormal) on mammogram data for each testing data have a different value. The first group have accuracy percentage of abnormalities by 85.00%. The second group have accuracy percentage of abnormalities by 60.00%. The third group have accuracy percentage of abnormalities by 67.5%.From the three groups, abnormalities accuracy values close to 100 percent is a group which using testing data were taken inside from training data by 85.00%.

4. Conclusion

In this research, diagnosis of Breast cancer in mammogram images using feature extraction and support vector machine is developed. The conclusion that the author retrieve from this research are: (1) Region of interest on mammogram image can be found using morphology algorithm. ROI or suspected area of cancer will be used to segmentation process. (2) Texture features have been proven to be useful in differentiating masses and normal breast tissues. The feature extraction method used is statistical method such as Grey level co-occurence matrix (GLCM). This research concentrate on statistical descriptors that include contrast, energy, homogeneity, entropy, and correlation of gray level values. (3) SVM (Support Vector Machine) is used to classify normal and abnormal classes on mammogram image. (4). The accuracy of abnormalities (normal or abnormal) in digital mammogram for each testing data have a different value. The first group have accuracy percentage of abnormalities by 85.00%. The second group have accuracy percentage of abnormalities by 60.00%. The third group have accuracy percentage of abnormalities by 67.5%.From the three groups, abnormalities accuracy values close to 100 percent is a group which is using testing data taken from training data by 85.00%.

This research is still necessary to development and improvement in the system. For the future, is expected to improve segmentation process (find the region of interest) by removing pectoral muscle and removing text noise from digital mammogram. Addition of GLCM features necessary to improve percentage of abnormality. Beside that, system can determine the level of severity (benign and malignant) from classification results. So it can help a doctor in diagnose breast cancer easily.

5. References

[1] Anonim, "Mammography", http://www.radiologyinfo.org/en/info.cfm?pg=mammo, June 2011.

[2] Anonim, "E Radiography", http://www.e-radiography.net/articles/mammo/mammo_introduction.htm, 2012

[3] A. Indrati, "Pengenalan Karakteristik Bentuk dan Batas Tepi Tumor Payudara," Ph.D. Dissertation, Gunadarma University, 2009.

[4] A. O. Malagelada, "Automatic Mass Segmentation in Mammographic Images," Ph.D. Dissertation, Universitat de Girona, April 2007.

[5] C.-I. I. Database, "Minimammographic Database", http://abacus.ee.cityu.edu.hk/imagedb/cgi-bin/ibrowser/ibrowser.cgi?folder=Medical_Image/mammogram/dense-glandular/, 2012.

[6] D. B. K. P. Rabi Narayan Panda and D. M. R. Patro, "Feature Extraction for Classification of Microcalcification and Mass Lession in Mammograms," 2009.

[7] D. S. Y. Ireaneus Anna Rejani, "Early Detection of Breast Cancer using SVM Classifier Technique," ser. 1, Pages 127-130, Publisher International Journal on Computer Science and Engineering, vol. 1(3), 2009.

[8] Dhika, "Apa Kabar Trend Kanker Payudara di Indonesia?", http://dhikatuy.blogdetik.com/2011/05/13/apa-kabar-trend-kanker-payudara-di-indonesia-2/, 2011.

[9] G, "Breast Cancer Incidence and Mortality Worldwide in 2008," http://globocan.iarc.fr/factsheets/cancers/breast.asp, 2008.

[10] M. H. Basim Alhadidi and H. N.Suleiman, "Mammogram Breast Cancer Image Detection using Image Processing Functions," ser. 2), Pages 217-221, Publisher Information Technology Journal, Vol. 6, 2007.

[11] P. P. D.Narain Ponraj, M.Evangelin Jenifer and J. Manoharan, "A Survey on The Preprocessing Techniques of Mammogram for The Detection of Breast Cancer," ser. 12, Pages 656-663, Month December, Publisher Journal of Emerging Trends in Computing and Information Sciences, Vol. 2, 2011.

[12] S. Basha and D. Prasad, "Automatic Detection of Breast Cancer Mass in Mammograms using Morphological Operators and Fuzzy C Means Clustering," pp. 704_709, 2009.

[13] S. B. Pradeep N, Girisha H. and K. K, "Feature Extraction of Mammograms," ser. 1, Pages 241-244, Publisher International Journal of Bioinformatics Research, Vol. 4, 2012.