Sample K-Means Clustering Method for Determining the Stage of Breast Cancer Malignancy Based on Cancer Size on Mammogram Image Basis

Karmilasari, Suryarini Widodo, Matrissya Hermita,
Nur Putri Agustiyani, Yuhilza Hanum
Dept. of Information System
Faculty of Computer Science &
Information Technology
Gunadarma University
Depok, Indonesia

Lussiana ETP.
Dept. of Information Systems
STMIK Jakarta STI&K Jakarta,
Jakarta, Indonesia

***Abstract*—Breast cancer is a disease that arises due to the growth of breast tissue cells that are not normal. The detection of breast cancer malignancy level / stage relies heavily on the results of the analysis of the doctor. To assist the analysis, this research aims to develop a software that can determine the stage of breast cancer based on the size of the cancerous tissue. Steps of the research consist of mammogram image acquisition, determining the ROI (Region of Interest), using Region growing segmentation method, measuring the area of suspected cancer, and determine the stage classification of the area on the mammogram image by using Sample K-Means Clustering method. Based on 33 malignant (abnormal) mammogram sample images taken from the mini mammography database of MIAS, the proposed method can detect stage of breast cancer is in malignant group.**

Keywords—classification; staging; breast cancer; mammogram; k-means clustering

I. Introduction

Cancer is a group of diseases that cause cells in the body to change and grow out of control. Most types of cancer cells eventually form a lump or masses called a tumor, and are named after the part of the body where the tumor originates. Breast cancer begins in breast tissue, which is made up of glands for milk production, called lobules, and the ducts that connect lobules to the nipple. The remainder of the breast is made up of fatty, connective, and lymphatic tissue [1].

Breast cancer is leading cause of cancer deaths among women. In 2013, an estimate 232.340 new cases of invasive breast cancer will be diagnosed among women, as well as an estimated 64.640 additional cases of in situ breast cancer, and appoximately 39.620 women are expected to die from breast cancer [1]

Detection and diagnosis of breast cancer in its early stage increases the chances for successful treatment and complete recovery of the patient. Screening mammography is currently the best available radiological technique for early detection of breast cancer [2]. It is an x-ray examination of the breasts in a woman who is asymptomatic. Mammography detects around 80% to 90% of breast cancers [3]. Masses or abnormalities detection at early stage is quite possible with the usage of mammography. Mammography is used as a primary tool for detecting breast cancer[4].

There are several stages to breast image processing. The first stage, breast image acquisition through mammography. The next stages are pre-processing image, segmentation, feature extraction, feature selection and classification [5]. With technique digital mammography, characteristics calcification, circumscribed, speculated and other ill defined masses can be diagnosed [6]

Breast cancer is the type of silent cancer because there is no typical symptoms suffered. Most people find this disease after entering the level of high stage of malignancy. Breast cancer stage is used to describe the condition of cancer, namely its location, its size, where it spreads and the extent of its influence on other organs .

In general , the level of breast cancer stage is stage I , II , III and IV [7]. In fact, determining the level of breast cancer stage is not easy. Many factors differ between the levels of the stadium .

The aim of this paper is to propose a method to determine the stage of breast cancer malignancy based on cancer size on mammogram image based on cancer size. This work is organized as follows. In Section 2, literature review that related work are presented. In Section 3, we present the proposed method includes the process of segmentation and classification. Next, in Section 4, the results are shown. Finally, Section 5 presents some concluding remarks.

II. Literature Review

Breast cancer detection can be carried out by using a variety of techniques. For successful treatment of the patient the breast cancer has to detected in its early stage and thus the patient can be recovered quickly. For breast cancer detection, the mammogram images will be collected in the first stage, after the image acquisition stage preprocessing will be performed. Next stage will be the image enhancement in which in the resultant image the finer details will be more clearer than the original image, the image will be segmented to extract the microcalcification part or cancer detected area.

Various technique used in breast cancer detection is described below:

A. Segmentation Technique

Segmentation is the process of partitioning a digital image into multiple segments. By segmentation technique it is easy to change the representation of an image so it will be easier to analyze and it is easy to locate objects and boundaries in images. In this technique image can be segmented and the set of segments will cover the entire image. Segmentation can be carried out using any of the standard techniques like Local Thresholding, K–Means Clustering, Otsu Segmentation Technique [8].

Thresholding is a way to change an image that has a level of grayscale or true color into an image with fewer color levels, in this case bilevel color is used. Bilevel image is a color image which is divided into two colors, 0 (black) and 1 (white). Simplification of color using thresholding is widely used for pattern recognition by eliminating color complexity into simple color so that an observed image has a color pattern which characteristics are easily grouped.

Otsu's method is used to automatically perform clustering-based image thresholding, or, the reduction of a graylevel image to a binary image [9]. The algorithm assumes that the image to be thresholded contains two classes of pixels or bi-modal histogram (e.g. foreground and background) then calculates the optimum threshold separating those two classes so that their combined spread (intra-class variance) is minimal. The extension of the original method to multi-level thresholding is referred to as the Multi Otsu method.

B. Edge Detection Method

Detection of edges in an image is a very important step towards understanding image features. Since edges often occur at image locations representing object boundaries, edge detection is extensively used in image segmentation when images are divided into areas corresponding to different objects. This can be used specifically for enhancing the tumor or cancer area in mammographic images. Different methods are available for edge detection like Roberts, Sobel, Prewitt, Kirsch and Laplacian of Gaussian edge operators [10].

C. Region Growing

Region Growing is a procedure that classifies the pixels or sub-regions into larger regions based on predefined criteria. The approach basically starts from the beginning of the set of points, then the area is enlarged by adding each neighboring pixel point that has properties similar to those points (for example the range of intensity or color specification).

The selection of similar criteria, in addition to depending on the problem at hand, also depends on the type of image data available, for example descriptor. Examples of descriptors include moment and texture. Region-growing segmentation provides the clear edges of the images. [11]. Region growing segmentation can be implemented to breast cancer detection [12].

D. Clustering K-Means

K-Means is a technique that is quite simple and quick clustering data. The main principle of this technique is to develop a k prototype/center of mass (centroid)/average (mean) of n-dimensional data set. This technique requires that the value of k is already known in advance (a priori). K-Means algorithm begins with the formation of initial prototype cluster. Then the prototype cluster is improved iteratively to converge (no significant changes to the prototype cluster). This change is measured using an objective function J which is generally defined as the sum / average distance to the centroid of each group of data. K-Means algorithm can be implemented to masses detection on digitized mammogram [13]

III. PROPOSED METHOD

The method proposed in this paper is to classify the stages of malignancy of breast cancer based on the mammogram image, through segmentation by sample K-Means clustering method.

Stages of the proposed method are outlined in Fig 1.

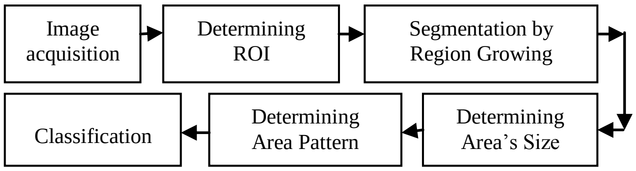

Fig. 1. Proposed method

The initial step is the image acquisition to get the data in the form of mammogram digital images that are required in the research. Mammography Image Analysis Society (MIAS) database used in this research [14]. Data is in the form of image format PGM (Portable GrayMap). PGM format is used by many medical image as a PGM is a lossless type image format where at the time of data compression, no parts are removed so that the details of the image will remain intact and not lost. Each mammogram image has a resolution of 1024x1024 pixels and the average file size of 1MB. MIAS database have been grouped into three categories, namely: (1) Dense-Glandular is the mammogram image of breasts that are dense and have many glands by nature (2) Fatty is the mammogram image of breasts that are not dense by nature because they contain a lot of fat, and (3) Fatty -glandular is a mammogram image of breasts that are not dense because they contain a lot of fat and have many glands.

Each of the three categories is further grouped into three sections, namely: (1) Normal, is a collection of normal mammogram images that are not affected by breast cancer, (2) Benign, is a collection of abnormal mammogram images that have been affected by benign breast cancer on breast tissue, and (3) malignant, is a collection of abnormal mammogram images that have invasive breast cancer. Figure 2 shows hierarchy MIAS database. In this research, 33 malignant (abnormal) mammogram images are used as test data and training data. Figure3 shows one abnormal image used in this research.

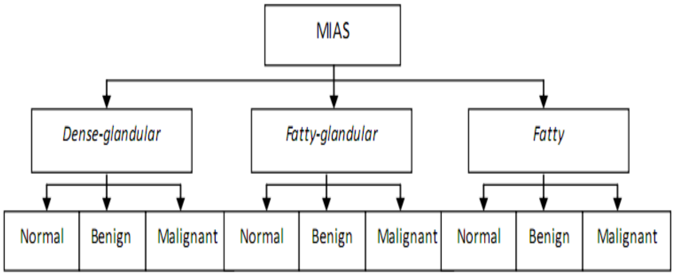

Fig. 2. Hierarchy MIAS database

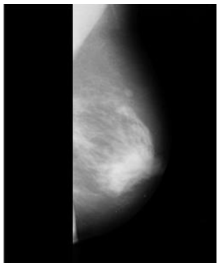

Fig. 3. Example malignant (abnormal) mammogram in MIAS

After image acquisition, the next step is the determination of the Region of Interest (ROI). The details of ROI determination step are shown in Figure 4.

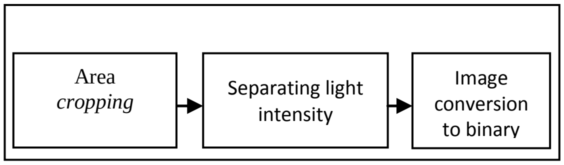

Fig. 4. Determining ROI in detail.

Region of Interest (ROI) allows for coding differently in certain areas of the digital image so as to have a better quality than the surrounding area. Determining ROI is a very important step if there is a certain part of the digital image that is more important than others. The first part in determining ROI is the are cropping process. This step aims to reduce the size of the image to be processed so that the result will be more accurate. Figure 5 is the result of cropping.

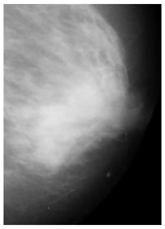

Fig. 5. Image after cropping

Next is to separate the light intensity with the Otsu method. Otsu thresholding separate the background and foreground by getting the value of each gray level variance. In this research, we used Otsu = 5. This method is more optimal than the Global thresholding method because of the way it works is to maximize the variance between classes.

The variance between these classes is suitable to statistically analyze class discriminant. The results of this phase is shown in Figure 6.

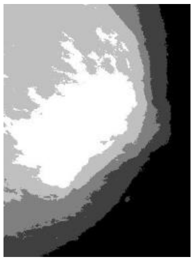

Fig. 6. Image after the separation of light intensity with Otsu=5

Then take an area that has estimated abnormalities by whitening the area which is regarded as normal. Black areas are considered as areas that are suspected of being cancerous abnormalities. This step is also called converting the image to binary. Figure 7 is the result of converting the image to binary phase.

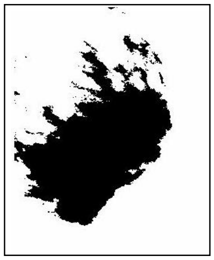

Fig. 7. Image after the phase of converting the image to binary.

The next step is segmentation. The purpose of image segmentation is to divide the image into a number of regions (areas) and separate an area that is estimated to show abnormalities. The estimated area is the one that has a stronger / brighter white color intensity than its surrounding, has a nearly uniform density, has a regular shape with various sizes, and its boundaries are blurred. Segmentation used in this study is the Region Growing segmentation method based on the resulted image from the determination of the ROI (cropping) step. Region Growing Segmentation is a segmentation method based on region. The principle of this method starts from a set of seed points, and initiate an initial region of the seed. This region will continue to grow from seed points to a collection of points adjacent to each other according to criteria. The result of the segmentation process can be seen in Figure 8.

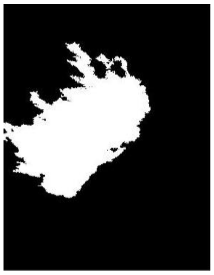

Fig. 8. Image after segmentation step.

Then calculate the area size of suspected abnormality found on the mammogram image resulted from segmentation with Region Growing method. At the time of the segmentation process, the image produced is a binary image of 0 and 1. Area 0 describe the normal area, marked with black. Area 1 illustrates the suspected area of cancer, marked with white. Calculation of area size 0 and 1 is generated in pixels. Values obtained from this process is based on the unit pixels.

After that, set the area pattern which aims to determine the pattern of suspected abnormalities to distinguish areas with suspected cancer from normal area. The next process is performed to detect the edge of the area, detecting the presence of edges of the suspected cancer area. This detection is useful for displaying the area borderline more clearly. Edge detecting method performed in this study is the Canny method, and the result is as shown in Figure 9. This method was chosen because it is able to produce a boundary edge more detailed than any other method.

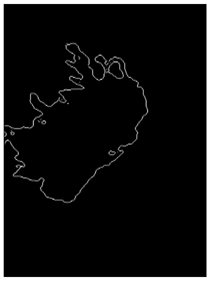

Fig. 9. Image after edge detection (thin line).

Resulted image of Canny edge detection has a thin line. This causes a line that is not clear and dotted. To make it thicker conducted dilation process as shown in Figure 10.

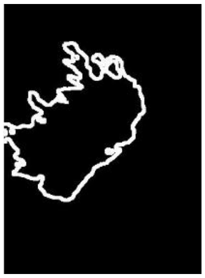

Fig. 10. Image after dillation process.

The next step is the incorporation of the image of the cropped area with the image resulted from dilation, and blacken the line resulted from edge detection. Both aim for the clear vision of the position of suspected areas with abnormalities. The results of these two processes can be seen in Figure 11 (a) and (b).

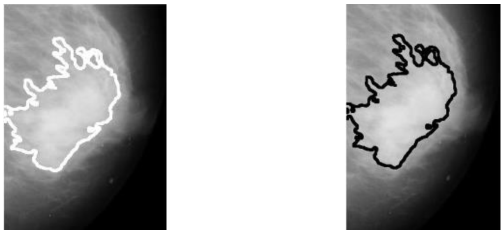

Fig. 11. (a) Image after the joining step of the cropped and dilation results. (b) Image after the blackening step of the edge.

As the last step is the mammogram image classification. This step aims to classify the mammogram image , whether the image is suspect of breast cancer stage I, II. or III . Stage IV is not used as in the basis of the theory used, patients in stage IV have cancers that vary in size and has spread to several parts of the body so that further examination is needed. In this research, the classification method used is based on the results of the mammogram image segmentation of 33 samples with Region Growing method.

Segmentation results from this method produce an area suspected of cancer. The area size is calculated in units of pixels. The measures are grouped into 3 major groups using K - Means clustering method. The clustering results reflect the 3 group stage, i.e. stage I, II, and III. Each group size has a lower limit and upper limit that are used as a reference to determine the stage of cancer in which the application is made .

IV. RESULT

In this research, 33 malignant mammogram images from MIAS database are used, where 12 images are from the group of malignant mammogram dense-glandular, 7 malignant mammogram images derived from fatty group and 14 malignant mammogram images derived from fatty-glandular groups.

After the process of segmentation, object area suspected breast cancer malignant from each group malignant mammogram will be obtained. K-Means Clustering is used to classify the object area suspected breast cancer into 3 stages. Stage 1 has size of the area between 3000 to 35000 pixel. Stage 2 has size of the area between 35000 to 85000 pixel. Stage 3 has size of the area between 85000 to 250000 pixel. Table 1 show the result of determining cancer stadium sample malignant mammogram image

TABLE I. The Results of Determining Cancer Stadium From Sample Malignant Mammogram Image.

No.	Mammogram Image	Group	Area Size (Pixels)	Stadium
1	mdb023.pgm	Dense-grandular	20052	Stadium I
2	mdb028.pgm	Dense-grandular	6850	Stadium I
3	mdb058.pgm	Dense-grandular	67754	Stadium II
4	mdb072.pgm	Dense-grandular	24205	Stadium I
5	mdb090.pgm	Dense-grandular	58623	Stadium II
6	mdb092.pgm	Dense-grandular	25521	Stadium I
7	mdb095.pgm	Dense-grandular	30294	Stadium I
8	mdb105.pgm	Dense-grandular	137669	Stadium III
9	mdb110.pgm	Dense-grandular	74595	Stadium II
10	mdb111.pgm	Dense-grandular	30065	Stadium I
11	mdb115.pgm	Dense-grandular	46132	Stadium II
12	mdb117.pgm	Dense-grandular	9333	Stadium I
13	mdb120.pgm	Fatty	48937	Stadium II
14	mdb124.pgm	Fatty	58808	Stadium II
15	mdb130.pgm	Fatty	83938	Stadium II
16	mdb134.pgm	Fatty	4772	Stadium I
17	mdb171.pgm	Fatty	135946	Stadium III
18	mdb178.pgm	Fatty	3470	Stadium I
19	mdb184.pgm	Fatty	19414	Stadium I
20	mdb202.pgm	Fatty-grandular	3558	Stadium I
21	mdb206.pgm	Fatty-grandular	20445	Stadium I
22	mdb209.pgm	Fatty-grandular	34093	Stadium I
23	mdb211.pgm	Fatty-grandular	37749	Stadium II
24	mdb213.pgm	Fatty-grandular	29616	Stadium I
25	mdb216.pgm	Fatty-grandular	156059	Stadium III
26	mdb233.pgm	Fatty-grandular	38599	Stadium II
27	mdb239.pgm	Fatty-grandular	154402	Stadium III
28	mdb241.pgm	Fatty-grandular	48011	Stadium II
29	mdb249.pgm	Fatty-grandular	49609	Stadium II
30	mdb253.pgm	Fatty-grandular	209699	Stadium III
31	mdb265.pgm	Fatty-grandular	47321	Stadium II
32	mdb267.pgm	Fatty-grandular	14508	Stadium I
33	mdb271.pgm	Fatty-grandular	4998	Stadium I

Based on the test of 33 samples of the malignant mammogram image, showed that: 48.5% are detected as stage I, stage II 36.37% detected and detected stage III 15.15%. In the dense-grandular group, 58.3% are detected as stage I, stage II 33.3% and detected stage III 8.33%. In the fatty group, 42,9% are detected as stage I, stage II 42.9% and detected stage III 14.2%. In the fatty-grandular group, 42.3% are detected as stage I, stage II 35.7% and detected stage III 21.4%.

V. Conclusion

The paper presented sample k-means clustering method for determining the stage of breast cancer malignancy based on cancer size on mammogram image basis. Previously done the process of determining ROI with Otsu method and segmentation with region growing. The method is tested on 33 mammograms in 3 groups of malignant in MIAS database. The result, system can determine the stage of breast cancer based on the size of the area of the suspected object. The further work may be develop stage of breast cancer based on patern of mallignant

References

[1] American Cancer Society, "Breast Cancer Facts & Fig. s, 2013-2014", American Cancer Society, Inc, 2013.

[2] Acha, B., Rangayyan, R.M., Desautels, J.E.L., "Detection of Microcalcifications in Mammograms", In: Suri, J.S., Rangayyan, R.M. (eds.) Recent Advances in Breast Imaging, Mammography, and Computer-Aided Diagnosis of Breast Cancer. SPIE, Bellingham, 2006.

[3] Highnam R and Brady M, "Mammographic Image Analysis", Kluwer Academic Publishers, British Journal of Radiology, 74(887), 2001.

[4] Maitra, I.K., Nag S., Bandyopadhyay S.K., "Identification of Abnormal Masses in Digital Mammography Images", International Journal of Computer Graphics, 2(1), 2011.

[5] Bozek, J., Mustra, M., Delac, K., and Grgic, M., "A Survey of Image Processing Algorithms in Digital Mammography", Multimedia Signal Processing and Communications Studies in Computational Intelligence Volume 231, 2009, pp 631-657.

[6] Yasmin, M., Sharif, M., and Mohsin, S., "Survey Paper on Diagnosis of Breast Cancer Using Image Processing Technique", Research Journal of Recent Sciences, Vol. 2(10), 88-98, October 2013.

[7] American Joint Committee on Cancer, "Breast Cancer Staging", AJCC 7th Edition Staging Posters, 2009.

[8] Pradeep, N., Girisha, H., Sreepathi, B., and Karibasappa, K., "Feature Exctraction of Mammograms",International Journal of Bioinformatics Research, Volume 4, Issue 1, 2012, pp.-241-244.

[9] Sezgin, M., and Sankur, B., "Survey over image thresholding techniques and quantitative performance evaluation". Journal of Electronic Imaging ,13 (1): 146–165. 2004.

[10] Maitra, I.K., Nag S., Bandyopadhyay S.K., "A Novel Edge Detection Algorithm for Digital Mammogram", International Journal of Information and Communication Technology Research, Vol 2 No.2, February 2012.

[11] Kamdi, S.," Image Segmentation and Region Growing Algorithm", International Journal of Computer Technology and Electronics Engineering (IJCTEE), Vol 2 Issue no.1, October 2011.

[12] Priya, D.S., and Sarojini, B., "Breast Cancer Detection In Mammogram Images Using Region-Growing And Contour Based Segmentation Techniques", International Journal of Computer & Organization Trends, Vol.3 Issue 8, September 2013.

[13] Martins, L.d.O., Junior, G.B., Silva, A.C., Paiva, A.C. and Gattass, M., "Detection of Masses in Digital Mammograms using K-Means and Support Vector Machine", Electronic Letters on Computer Vision and Image Analysis 8(2) : 39-50, 2009.

[14] Suckling, J., "The Mammographic Image Analysis Society Digital Mammogram Database", Exerpta Medica, International Congress 1069, pp 375-378, 1994.