The application of statistic image analysis for classification of breast cancer based on mammograms

D R Sulistyaningrum¹,a, B Setiyono², D B Utomo³ and B A Sanjoyo⁴

¹,²,³,⁴Department of Mathematics, Institut Teknologi Sepuluh November Campus ITS, Sukolilo, Surabaya, 60111, Indonesia

E-mail: a dwiratna@matematika.its.ac.id

Abstract. Cancer is a disease characterized by the ability of abnormal cells to grow uncontrollably. In the medical field, detection of breast cancer is done using a mammogram. Examination of the mammogram image is still done manually by the doctor / radiologist, so it is necessary to use technology as supporting information. In this research, mammogram image classification based on gray-level co-occurrence (GLCM) matrix and gray-tone difference matrix (GTDM) has been done with backpropagation method. The stages of the mammogram image classification process include the process of image acquisition, pre-processing, feature extraction with GLCM and GTDM and classification using backpropagation. The pre-processing process carried out is gray-scaling, contrast enhancement, image segmentation with Otsu thresholding, edge detection process, and image thickening process with widening morphology method. The highest performance results for accuracy are 85% and precision is 85.7%. This result was obtained when using the GLCM and GTDM feature extraction methods.

1. Introduction
Cancer is one of the leading causes of death worldwide. Lung, liver, stomach, colorectal, and breast cancer are the biggest causes of cancer deaths every year. Breast Cancer is one of the diseases suffered by most of the female population in Indonesia.

In the medical field, the detection of breast cancer is done using mammography which has become a commonly used method for early detection of breast cancer although mammograms have low contrast so that sometimes mammogram examination is not easy. Examination of the image of a mammogram that is done manually by doctors or radiologists often has little accuracy. Therefore a technology (system) that can help detect the presence of cancer in the breast through the image of a mammogram by classifying the image is needed.

Classification of breast cancer based on mammogram images can be done based on statistical characteristics of the image. The Gray-Level Co-occurrence Matrix (GLCM) and Gray-Tone Difference Matrix (GTDM) methods can be used to determine the features of a mammogram image. GLCM is a method used to determining features of the image based on spatial relationships between gray degrees of pixels, while GTDM is a method that is used to determine the feature of an image based on the difference between the gray level of a pixel and its neighboring pixel.

Both GLCM and GTDM have been used in several studies on feature extraction techniques. For example, in the study of hybrid extraction feature methods on the classification of war scene using
support vector machine [1]. GLCM is combined with several other methods such as Zernike moments, statistical moments, Haar wavelets and Daubechies wavelets to improve process performance. Research conducted by R. N. Dhivya et al focus on the hybrid approach for the detection and classification of fabric errors in the textile industry using Wavelet Transformation and GLCM Techniques [2]. Another study is the research conducted by B. Arthy et al focus on the various representations of perceptual features for texture classification by using GLCM and GTDM as a method for feature extraction process [3].

Based on this review, in this study we investigated breast cancer classification using statistical features namely GLCM and GTDM based on mammogram images with backpropagation method.

2. Methods

Figure 1 shows the proposed methods to classification breast cancer using statistical image analysis. It consists of several processing units: First, image acquisition. Second, pre-processing to enhancement image quality. Third, feature extraction with GLCM and GTDM. Next, classification using backpropagation. A detailed description of the scheme is described in the next discussion.

![Figure 1. Scheme of the proposed method](image_url)

2.1. Image Enhancement using Mathematical Morphology

The Mathematical morphology is largely concerned with the mathematical theory that describes the shape using the set. In image processing, mathematical morphology is used to investigate the interaction between an image and a selected element structure such as lines, circles or hexagons. The application of mathematical morphology are object identification, reducing image noise, and others [4].
Mathematical morphology consists of two simple morphological operations, namely erosion and dilation. These operations are the basis for morphological processing. By combining one with another can obtain different image processing operations.

The dilation is a process that is done to thicken the object in a binary image $A$ using the structuring element $B$. The dilation of the image $A$ by structuring element $B$ is defined by:

$$ A \oplus B = \{ z | (\hat{B}) \cap A \neq \emptyset \} $$

Erosion is a process that is done to attenuate object in a binary image $A$ using a structuring element $B$. The erosion of the image $A$ by structuring element $B$ is defined by:

$$ A \ominus B = \{ z | (B) \subseteq A \} $$

Opening and closing are processes that are carried out to soften the outline of an object, remove narrow parts, and remove ridge objects in binary images using structure elements $B$ to control the thickening process.

2.2. Feature Extraction using GLCM and GTDM

The purpose of feature extraction is to get the main characteristics that will provide important information for the classification process. In this study, the Gray-Level Co-occurrence Matrix (GLCM) and Gray-Tone Difference Matrix (GTDM) methods were used for feature extraction.

A. Gray-Level Co-occurrence Matrix (GLCM)

GLCM is one of the most well-known statistical methods for image processing introduced by Haralick[5]. Gray level co-occurrence matrix (GLCM) is one of the statistical methods that can be used for texture analysis. Co-occurrence matrix is a matrix that describes the relationship between neighboring pixels in an image with a certain orientation and distance. GLCM is a matrix $P$ that contains the relative frequency of two pixels, one with a gray level value $i$ and another with a gray level value $j$ separated by a distance $d$ at an angle $\theta$. The neighboring pixels ($i$ and $j$) that has a distance between the two can be located in 8 different directions both horizontally, vertically and diagonally. Figure 2 shows the co-occurrence matrix of GLCM.

![Figure 2. Gray-Level Co-occurrence Matrix](image-url)
Co-occurrence matrix elements can be used to calculate the measure of the image texture component. Table 1 shows the mathematical formulas of some of the statistical features of GLCM that can be obtained from images[6].

Table 1. Mathematical Formula of GLCM Features

| Texture features | Mathematical Formula |
|------------------|----------------------|
| Contrast         | $Contrast = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} P_{i,j}(i-j)^2$ |
| Energy           | $Energy = \sqrt{\sum_{i=0}^{N-1} \sum_{j=0}^{N-1} P_{i,j}^2}$ |
| Entropy          | $Entropy = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} P_{i,j}(-\ln P_{i,j})$ |
| Homogeneity      | $Homogeneity = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \frac{P_{i,j}}{1 + (i-j)^2}$ |
| Dissimilarity    | $Dissimilarity = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} P_{i,j}|i-j|$ |
| Correlation      | $Correlation = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} P_{i,j} \frac{(i-\mu_i)(j-\mu_j)}{\sqrt{(\sigma_i^2)(\sigma_j^2)}}$ |

B. Gray-Tone Difference Matrix (GTDM)

Gray-Tone Difference Matrix (GTDM) is an algorithm for determining texture features that correlate with human perception. A GTD matrix is a column vector consisting of G elements. Vector element values are calculated based on the difference between the intensity level of a pixel and the average intensity of a neighbor[7]. Suppose the intensity level of an image $f(x,y)$ at the location $(x, y)$ is $i$, where $i=1, 2, \ldots, G-1$. The average intensity in a window with the center $(x,y)$ is:

$$\bar{f}_i = \bar{f}(x,y) = \frac{1}{W-1} \sum_{m=-k}^{k} \sum_{n=-k}^{k} f(x+m, y+n)$$

where $k$ is the window size and $W = 2(2k + 1)$.

The value of the first element of the GTDM matrix is represented by $S(i)$ and formulated as:

$$S(i) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} |i - \bar{f}(x,y)|$$
for all pixels that have intensity $i$ and $S(i) = 0$ for the others. Mathematical formulas for GTDM features are given in Table 2.

**Table 2. Mathematical Formula of the GTDM Feature**

| Texture features | Mathematical equations |
|------------------|-----------------------|
| Coarseness       | $Coarseness = \left( \varepsilon + \sum_{i=0}^{G-1} P_i S(i) \right)^{-1}$ |
| Busyness         | $Busyness = \frac{\sum_{i=0}^{G-1} P_i S(i)}{\sum_{i=0}^{G-1} P_i \sum_{j=0}^{G-1} \lvert iP_i - jP_j \rvert}$ |
| Complexity       | $Complexity = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} \frac{|i - 1|}{np_i + np_j} [P_iS(i) + P_jS(j)]$, $P_i$ dan $P_j \neq 0$ |
| Texture Strength | $Texture\ Strength = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} P_i + P_j (i - j)^2 \varepsilon + \sum_{i=0}^{G-1} S(i)$ |

2.3. Backpropagation algorithm for classification

Backpropagation is a systematic method of artificial neural networks using supervised learning algorithms. Usually the backpropagation algorithm uses multi layers to change the weights in hidden layers. This pattern of weight adjustment is done to achieve the minimum error value between the output of the predicted results and the actual output [8].

The backpropagation algorithm has two computational stages, namely feedforward and backward calculations. In each iteration, the network will correct the weight values and bias all neurons in the network. Figure 3 shows the scheme of the network of multi-layer perceptron from the backpropagation algorithm with 10 input layers and 3 output layers.

![Figure 3. The network of multi-layer perceptron](image-url)
3. Experiments Design

The purpose of this study is to classify mammogram images into 3 classes, namely the image of a normal mammogram, benign cancer and malignant cancer. The initial step of this research is that pre-processing that aims to strengthen the features of the mammogram. The feature extraction process uses two statistical methods, namely GLCM and GTDM. The purpose of feature extraction is to determine the characteristics that distinguish between one and the other mammogram images. There are 10 feature values obtained from a mammogram image, namely 8 characteristic values of GLCM and 4 characteristic values of GTDM. The characteristics of GLCM include contrast, dissimilarity, energy, entropy, homogeneity and correlation. While the characteristics of GTDM include coarseness, busyness, complexity, and texture strength.

In this study, the mammogram image used in the experiment was obtained from MIAS Mini Mammographic Database. The number of images on the dataset is 320 images. The training data were 22 malignant cancer mammogram images, 32 benign cancer mammogram images, and 170 normal mammogram images. Tests were carried out on 4 different angles namely 00, 450, 900 and 1350 with a distance of 1.

The performance assessment of the method is done through calculating the value of accuracy, error, and precision. These values are determined using the following equation[9]:

\[
Accuracy = \frac{TP + TN}{P + N}
\]

\[
Error = \frac{FP + FN}{P + N}
\]

\[
Precision = \frac{TP}{TP + FN}
\]

TP (True Positive) is a mammographic image diagnosed and detected by cancer, both benign and malignant cancer. TN (True Negative) is a mammographic image that is diagnosed and detected as normal. FP (False Positive) is a mammographic image of malignant cancer that is detected by benign cancer or the image of benign cancer detected by malignant cancer. FN (False Negative) is a mammographic image of malignant cancer and benign cancer that is detected as normal. P (Positive) is the number of mammographic images of malignant cancer and benign cancer. N (Negative) is the number of normal mammographic images.

4. Results and discussions

The purpose of the pre-processing stage is to obtain a better quality mammogram image. In the pre-processing stage, there are four main sub-processes, namely contrast strengthening, thresholding, edge detection and morphology. Figure 4 show examples of images obtained from each stage.

![Figure 4](image-url)

**Figure 4.** (a) Mammogram Image. (b) Result of using contrast strengthening. (c) Result of using thresholding. (d) Result of using edge detection. (e) Result of using morphology
The experimental results in each test are shown by the values of accuracy, error, and precision presented in Figure 5 and Figure 6. The level of accuracy, error, and precision in experiments using the GLCM and GTDM feature extraction methods are shown in Figure 5. The highest accuracy and precision is achieved when the angle is 90 degrees, as well as the lowest error reached at 90 degrees. Figure 6 shows the level of accuracy, error, and precision in the experiment using the GLCM feature extraction method. The level of accuracy, error and precision in the extraction of GTDM features is 57.5, 37.5 and 53.3.

**Figure 5.** The level of accuracy, error and precision when using the GLCM and GTDM

**Figure 6.** The level of accuracy, error and precision when using only the GLCM

Based on the experiments, there are several conditions that cause low performance. The main cause of the low value of accuracy, error, and precision:

1. Image quality after pre-processing is less good. In some images, non-cancer cells are also extracted by their features
2. The amount of data in the training process is still lacking
3. The number of features used as input for the classification process is too little.
5. Conclusions
This study has succeeded in classifying breast cancer types based on mammogram images using statistical analysis and backpropagation algorithms. The highest performance results for accuracy are 85% and precision is 85.7%. This result was obtained when using the GLCM and GTDM feature extraction methods.

References
[1] T. Nadu and T. Nadu, “EFFICIENT HYBRID FEATURE EXTRACTION METHODS IN WAR SCENE CLASSIFICATION USING SUPPORT VECTOR MACHINES,” vol. VII, no. ii, 2016.
[2] N. Dhivya, R. A. Lakshmi, and D. Ramalakshmi, “A Hybrid Approach for Detection And Classification of Fabric Faults In Textile Industries Using Wavelet Transform And GLCM Technique,” no. 4, pp. 121–124, 2013.
[3] G. Tamilpavai and S. Tamil Selvi, “Multiple Representations of Perceptual Features for Texture Classification and Retrieval,” Int. J. Comput. Appl., vol. 48, no. 20, pp. 5–11, 2012.
[4] Z. Diao, C. Zhao, G. Wu, and X. Qiao, “Review of application of mathematical morphology in crop disease recognition,” IFIP Int. Fed. Inf. Process., vol. 294, pp. 981–990, 2009.
[5] N. Zayed and H. A. Elnemr, “Statistical Analysis of Haralick Texture Features to Discriminate Lung Abnormalities,” Int. J. Biomed. Imaging, vol. 2015, 2015.
[6] M. Castro, C. Lo, D. Dicastillo, and J. Manuel, “Article pubs.acs.org/JAFC,” pp. 1–14, 2013.
[7] M. Amadasun and R. King, “Textural features corresponding to textural properties,” EEE Trans. Syst. Man, Cybern., vol. 19, no. 5, pp. 1264–1274, 1989.
[8] L. Fausett, Fundamentals of Neural Networks: Architectures, Algorithms And Applications, 1st ed. Pearson, 1993.
[9] J. Han, J.,Kamber, M.,Pei, Data Mining : Concepts and Techniques, Third Edition, 3rd ed. Elsevier, 2011.