Available *online* at : http://jitce.fti.unand.ac.id/
JITCE (Journal of Information Technology and Computer Engineering)
| ISSN (Online) 2599-1663 |

Increased Mammogram Image Contrast Using Histogram Equalization And Gaussian In The Classification Of Breast Cancer

Febri Liantoni[1], Coana Sukmagautama[2], Risalina Myrtha[3]

[1] *Department of Computer and Informatics Education, Sebelas Maret University, Indonesia*
[2,3] *Medical Doctor Profession, Faculty of medicines, Sebelas Maret University, Indonesia*

ARTICLE INFORMATION

Received: February 9th, 2020
Revised: March 04th, 2020
Available online: March 30th, 2020

KEYWORDS

Gaussian, histogram equalization, k-nearest neighbor, mammogram, naïve Bayes classifier.

CORRESPONDENCE

E-mail: febri.liantoni@staff.uns.ac.id

A B S T R A C T

Breast cancer is one of the most common disease among women in several countries. One of the most common method to diagnose breast cancer is mammography. In this study, we propose a classification study to differentiate benign and malignant breast tumors based on mammogram image. The proposed system includes five major steps, i.e. preprocessing, histogram equalization, convolution, feature extraction, and classification. Image is cropped using region of interest (ROI) at preprocessing stage. In this study, we perform image contrast quality enhancement of the mammogram to view the breast cancer better. Image contrast enhancement uses histogram equalization and Gaussian filter. Gray-Level Co-Occurrence Matrix (GLCM) is used to extract the mammogram features. There are five features used i.e. entropy, correlation, contrast, homogeneity, and variance. The last step is to classify using naïve Bayes classifier (NBC) and k-nearest neighbor (KNN). Based on the hypothesis, the accuracy of NBC method is 90% and the accuracy of KKN method is 87.5%. So, the mammogram image contrast enhancement is well performed.

INTRODUCTION

Breast cancer is one of the most feared diseases among women. Medical imaging is related to imaging techniques and processes of the human body for medical purposes by finding, examining or diagnosing diseases. In a narrow context, medical imaging is often equated to radiology. One way to identify the presence of breast cancer in the early stage is by interpreting mammogram images using a low-dose x-ray technique. Mammography screening is one of several known methods for detecting breast cancer at earlier stage. Mass and calcification are among findings that can be found on a mammogram. Calcification is calcium deposits in breast tissue. This sign can be observed as a white dot or spot on a mammogram but cannot be palpated during physical breast examination. Digital mammogram is one of the best methods for detecting breast cancer, but it requires skill and experience of radiologists to interpret this digital mammogram. Therefore, image processing techniques on digital mammograms are needed to help radiologists identify and obtain advice in diagnosing breast cancer. Image processing has been widely used in the medical field to help diagnose a disease including detection of disease in mammogram images [1], [2].

Several studies have been developed to help diagnose breast cancer. Research conducted by Tintu and Paulin discussed the classification of mammogram images into 2 classes, i.e. normal and identified cancers. This study used data from the Wisconsin Prognostic Breast Cancer (WPBC). The classification method used is Fuzzy C Means [3]. Vijayarajeswari classified breast cancer from mammogram images using SVM. The results showed that the proposed method could effectively classified abnormal mammogram classes [4]. Murat Karabatak carried out breast cancer detection using naïve Bayesian [5]. Celik made image improvements by equalizing two-dimensional histograms that effectively improve the quality of different image types [6]. Wang also performed contrast enhancement techniques based on local histogram equalization algorithm. The proposed technique was performed by segmenting the image into several sub-blocks using gradient values, this algorithm succeeded in increasing local contrast without adding noise to the image [7].

In this research, the mammogram image contrast is increased using equalization and Gaussian histogram. To analyze the texture of breast tissue, Gray Level Co-occurrence Matrices (GLCM) features are used. The GLCM feature provides information about homogeneity and roughness of image textures that are sometimes invisible to the eye. GLCM features are more effective in detecting calcifications than morphological-based features. In this study, features taken from GLCM are entropy, correlation, contrast, homogeneity, and variance. Those five GLCM features represent textural features that are interpreted in mammographic images, i.e. contrast between parts of the network

https://doi.org/10.25077/jitce.4.01.40-44.2020

image and the degree of graying homogeneity of the areas on the network. In the classification process of malignant and benign tumors, the Naïve Bayes Classifier (NBC) and K-Nearest Neighbor (KNN) methods are used.

METHOD

In this research, several stages are carried out, i.e. mammogram image datasets collection, region of interest (ROI) extraction process, image enhancement using equalization and Gaussian histogram, features extraction using GLCM, and classification using NBC and KNN. The research process flowchart is shown in Figure 1.

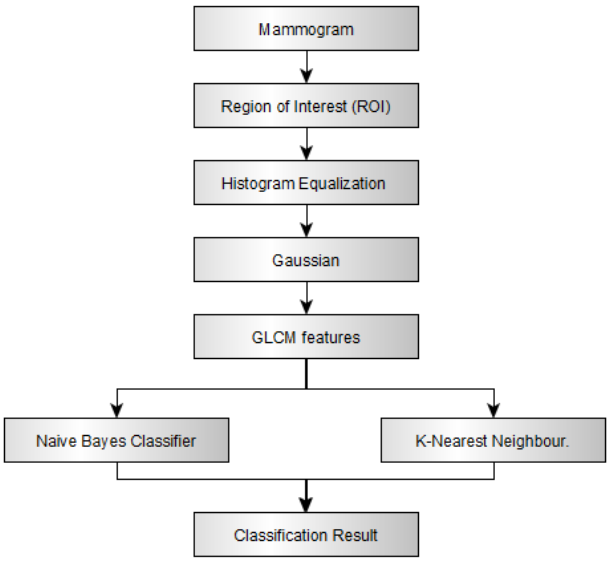


Figure 1. Research flowchart

Mammogram

In mammography, breast cancer is identified by the presence of a mass or calcification [8].
A mammogram is a special device that uses radiation or X-ray energy to observe the developments of breast tissue. A mammogram is used as a tool to examine and detect various forms of abnormalities in the breast tissue, such as breast cancer, tumors, breast cysts, or calcium deposition (calcification) of the breast tissue. Generally, it is used to detect breast cancer cells at earlier stage.

Based on its purpose, Mammograms consist of 2 types, i.e. screening mammogram and diagnostic mammogram. A screening mammogram is a test conducted to detect abnormalities in the breast even though there are no visible signs or abnormalities. A screening mammogram is performed to early detect breast cancer. While a diagnostic mammogram is a test conducted to identify changes that occur in the breast, such as pain, lumps, skin color changes around the breast, nipple thickening, or discharge from the nipple. A diagnostic mammogram is also used to evaluate abnormalities that were previously obtained during screening.

Histogram Equalization (HE)

Histogram equalization is a process of histogram equalization, where the distribution of gray degree values in an image is made flat. To perform histogram equalization requires a cumulative distribution function, cumulative distribution is a calculation of the histogram. The cumulative distribution function can be defined as follow.

$$f(k) = \frac{(N-1)}{M} \sum_{k=0}^{n} h(k); \quad n = 1,2,3,4 \ldots . N-1 \qquad (1)$$

M is the overall value of pixels in an image. N is the gray pixel value. And h (k) is a histogram of the gray value k.
Histogram equalization can be used to get information about the frequency of gray levels use in an image.

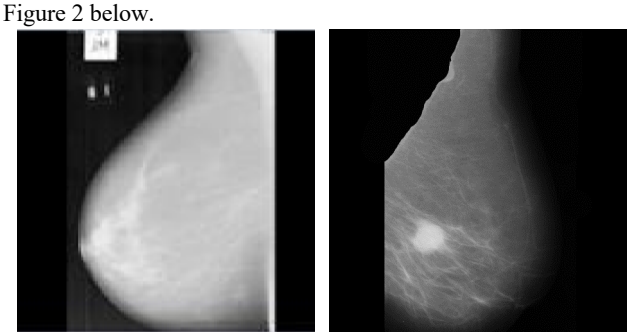
Gaussian filter includes in the H filter in the form of a matrix m x n, which has the same value for each element. This filter is LPF so the sum of all elements is one [9].

Classification

The K-Nearest Neighbor (K-NN) method is one of the nearest neighbor (NN) -based methods that can be used to compare histograms values [10]. The K value used here defines the number of nearest neighbors involved in determining the prediction of the class label in the test data. The working principle of the KNN method is by calculating the value of the nearest distance between the data to be evaluated and the nearest K neighbor in the processed data. Naïve Bayes classifier is a part of the probabilistic classifiers which is based on the application of the Bayes theorem using a strong assumption of independence [11]. The naïve bayes method is a popular method for categorizing texts, documents, spam detection based on word frequency as a feature.

RESULTS AND DISCUSSION

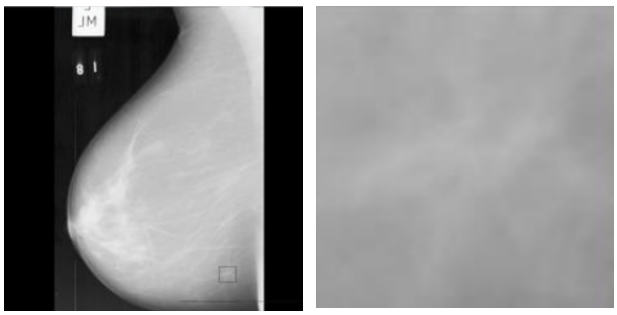
Test is performed to determine the level of success of the system in classifying malignant and benign tumors. This study uses 100 mammogram images that consist of 60 images of benign tumors and 40 images of malignant cancer. Next, those data are divided into 60 imagery training data and 40 imagery test data. In this test, 20 images of benign tumors and 20 images of malignant cancers are used. Examples of mammogram data used are shown in Figure 2 below.



a b

Figure 2. Mammogram, a) benign tumor, b) malignant cancer.

The initial phase of the research is the cropping process using ROI. This process is carried out to take part that will be observed in this study. An example of the result of the ROI process is shown in Figure 3 below.



a b

Figure 3. ROI cropping, a) initial image, b) ROI image result

The next process is carried out by histogram equalization. This process aims to improve the image quality by changing the gray level of the image distribution so that it is expected to obtain optimal feature values. The histogram equalization result is shown in Figure 4 below.

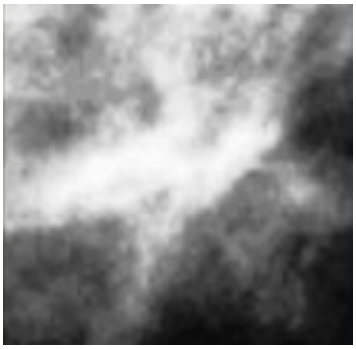


Figure 4. Histogram equalization results

After the histogram equalization process, a Gaussian filter is performed. This process aims to increase the image contrast of the mammogram. The image of the Gaussian filter result is shown in Figure 5.

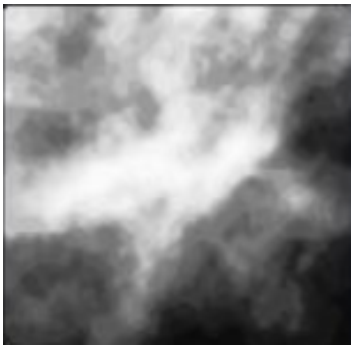


Figure 5. Gaussian filter image result

Image processing using histogram equalization and Gaussian filter is proven to increase the image contrast of the mammogram. So, the mass in breast cancer can be seen more clearly. Next, from the results of the Gaussian filter, feature extraction is performed using GLCM. The classification process of breast cancer is performed based on the features of entropy, correlation, contrast, homogeneity, and variance. Some examples of feature extraction results are shown in Table 1.

Table 1. GLCM feature extraction results

Filename	Correlation	Contrast	Homogeneity	Variance	Entropy
mdb124.pgm	17.1624743	0.982251132	0.859799	51.83921	1.510753
mdb125.pgm	17.27066671	1.760561576	0.801248	49.79566	1.603934
mdb130.pgm	17.9355759	0.515040964	0.860306	56.91917	1.399553
mdb134.pgm	16.02431226	0.374126912	0.870941	50.89106	1.480934
mdb141.pgm	16.18568254	0.244398313	0.905521	53.13463	1.425794
mdb195.pgm	16.75808725	1.168007016	0.861902	50.838	1.501385
mdb198.pgm	15.80976847	0.451895048	0.880806	50.85548	1.48161
mdb199.pgm	17.29396528	0.459412442	0.86675	53.55148	1.497371
mdb204.pgm	18.02743188	0.575863085	0.859619	57.387	1.363372
mdb207.pgm	19.75806677	0.572758841	0.874934	59.92591	1.417134

The classification testing process uses the NBC and KNN methods. The tests are carried out for 40 test data images that consist of 20 benign tumor images and 20 malignant cancer images. Based on the testing for benign tumors, data are obtained as shown in Table 2.

 https://doi.org/10.25077/jitce.4.01.40-44.2020

Table 2. Results for benign cancer class

Filename	Class		
	ACTUAL	NBC	KNN
Mdb195.Pgm	BENIGN	MALIGNANT	MALIGNANT
Mdb198.Pgm	BENIGN	BENIGN	BENIGN
Mdb199.Pgm	BENIGN	BENIGN	BENIGN
Mdb204.Pgm	BENIGN	BENIGN	BENIGN
Mdb207.Pgm	BENIGN	BENIGN	BENIGN
Mdb212.Pgm	BENIGN	BENIGN	BENIGN
Mdb218.Pgm	BENIGN	BENIGN	BENIGN
Mdb219.Pgm	BENIGN	BENIGN	BENIGN
Mdb222.Pgm	BENIGN	BENIGN	BENIGN
Mdb223.Pgm	BENIGN	BENIGN	BENIGN
Mdb226.Pgm	BENIGN	BENIGN	BENIGN
Mdb227.Pgm	BENIGN	BENIGN	BENIGN
Mdb236.Pgm	BENIGN	BENIGN	BENIGN
Mdb240.Pgm	BENIGN	BENIGN	BENIGN
Mdb244.Pgm	BENIGN	BENIGN	BENIGN
Mdb248.Pgm	BENIGN	BENIGN	BENIGN
Mdb252.Pgm	BENIGN	BENIGN	BENIGN
Mdb290.Pgm	BENIGN	BENIGN	BENIGN
Mdb312.Pgm	BENIGN	BENIGN	BENIGN
Mdb314.Pgm	BENIGN	BENIGN	BENIGN

Table 2 shows the test results using the NBC method and there is a classification error in the filename "mdb195.pgm". The same error also occurs when testing is carried out using the KNN method. Whereas data tests for the malignant cancer class are obtained as shown in Table 3.

Table 2. Results for malignant cancer class

Filename	Class		
	ACTUAL	NBC	KNN
Mdb144.pgm	MALIGNANT	BENIGN	BENIGN
Mdb148.Pgm	MALIGNANT	MALIGNANT	MALIGNANT
Mdb155.Pgm	MALIGNANT	MALIGNANT	MALIGNANT
Mdb158.Pgm	MALIGNANT	MALIGNANT	BENIGN
Mdb170.Pgm	MALIGNANT	MALIGNANT	MALIGNANT
Mdb171.Pgm	MALIGNANT	MALIGNANT	MALIGNANT
Mdb178.Pgm	MALIGNANT	MALIGNANT	MALIGNANT
Mdb179.Pgm	MALIGNANT	BENIGN	BENIGN
Mdb181.Pgm	MALIGNANT	MALIGNANT	MALIGNANT
Mdb184.Pgm	MALIGNANT	BENIGN	BENIGN
Mdb186.Pgm	MALIGNANT	MALIGNANT	MALIGNANT
Mdb202.Pgm	MALIGNANT	BENIGN	BENIGN
Mdb206.Pgm	MALIGNANT	BENIGN	BENIGN
Mdb209.Pgm	MALIGNANT	MALIGNANT	MALIGNANT
Mdb213.Pgm	MALIGNANT	MALIGNANT	MALIGNANT
Mdb231.Pgm	MALIGNANT	MALIGNANT	MALIGNANT
Mdb238.Pgm	MALIGNANT	BENIGN	BENIGN
Mdb239.Pgm	MALIGNANT	MALIGNANT	MALIGNANT
Mdb241.Pgm	MALIGNANT	MALIGNANT	MALIGNANT
Mdb249.Pgm	MALIGNANT	MALIGNANT	MALIGNANT

Table 3 shows the test results using NBC, there are 6 errors, while there are 7 classification errors using KNN. Based on the overall test results on the breast cancer classification process using NBC method, 33 test data are obtained in accordance with the actual data. While the overall test results using the KNN method obtain 32 test data in accordance with the actual data.

So the precision, recall and accuracy of the system is calculated as shown in Table 4, Table 5 and Table 6.

Table 4. Confusion matrix of algoritma NBC

Prediction Class	True Yes	True No	Precision Class
Prediction: Yes	33	2	94.28%
Prediction: No	2	3	60%
Recall Class	94.28%	60%	

Table 5. Confusion matrix of algoritma KNN

Prediction Class	True Yes	True No	Precision Class
Prediction: Yes	32	3	91.42%
Prediction: No	2	3	60%
Recall Class	94.11%	50%	

Table 6. Comparative results of data mining algorithms

Parameter	NBC	KNN
Accuracy	90%	87.5%

These results indicate that the NBC algorithm has a better performance in building a model used to find cancer classifications.

CONCLUSIONS

Based on the hypotheses that have been tested and discussed, it can be concluded that the equalization of the histogram and the Gaussian filter can improve the image contrast of the mammogram better. The accuracy obtained in the classification using the naïve Bayes classifier is 90%, while the accuracy using the k-nearest neighbor is 87.5%. This shows that the process of increasing the image contrast of mammogram using histogram equalization and Gaussian filter works well. These results also prove that the naïve Bayes classifier method and k-nearest neighbor can be used as a reference for classifying breast cancer.

REFERENCES

[1] M. M. Jadoon, Q. Zhang, I. U. Haq, S. Butt, and A. Jadoon, "Three-Class Mammogram Classification Based on Descriptive CNN Features," *Biomed Res. Int.*, vol. 2017, pp. 1–11, 2017.

[2] D. Daye *et al.*, "Mammographic Parenchymal Patterns as an Imaging Marker of Endogenous Hormonal Exposure," *Acad. Radiol.*, vol. 20, no. 5, pp. 635–646, May 2013.

[3] Tintu and Paulin, "Detect Breast Cancer using Fuzzy C means Techniques in Wisconsin Prognostic Breast Cancer (WPBC) Data Sets," *Int. J. Comput. Appl. Technol. Res.*, pp. 614–617, Sep. 2013.

[4] R. Vijayarajeswari, P. Parthasarathy, S. Vivekanandan, and A. A. Basha, "Classification of mammogram for early detection of breast cancer using SVM classifier and Hough transform," *Measurement*, vol. 146, pp. 800–805, Nov. 2019.

[5] M. Karabatak, "A new classifier for breast cancer detection based on Naïve Bayesian," *Measurement*, vol. 72, pp. 32–36, Aug. 2015.

[6] T. Celik, "Two-dimensional histogram equalization and contrast enhancement," *Pattern Recognit.*, vol. 45, no. 10, pp. 3810–3824, 2012.

[7] Y. T. Chang, J. T. Wang, W. H. Yang, and X. W. Chen, "Contrast Enhancement in Palm Bone Image Using Quad-Histogram Equalization," pp. 1091–1094, 2014.

[8] S. Timp and N. Karssemeije, ""Interval change analysis to improve computer aided detection in mammography," *Med. Image Anal.*, vol. 10, no. 1, pp. 82–95, 2006.

[9] R. Gonzales and R. Wood, *Digital Image Processing*. Prentice-Hall, Inc., United State, America, 2007.

[10] F. S. Mohamad, A. A. Manaf, and S. Chuprat, "Nearest Neighbor For Histogram-based Feature Extraction," *Procedia Comput. Sci.*, vol. 4, pp. 1296–1305, 2011.

[11] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3 edition. Upper Saddle River: Pearson, 2009.

 https://doi.org/10.25077/jitce.4.01.40-44.2020