Feature Extraction and Classification of Tissue Mammograms Based on Grayscale and Gray Level Co-occurrence Matrix

[1st] Herwanto
Faculty of Engineering
Universitas Krisnadwipayana
Jakarta, Indonesia
herwanto@unkris.ac.id

[2nd] Ali Khumaidi
Faculty of Engineering
Universitas Krisnadwipayana
Jakarta, Indonesia
alikhumaidi@unkris.ac.id

[3rd] Harjono Padmono Putro
Faculty of Engineering
Universitas Krisnadwipayana
Jakarta, Indonesia
harjonopputro@unkris.ac.id

Abstract— Breast cancer is a major cause of death for women in the world. Breast cancer can be diagnosed by various means of examination, including mammography examination, which indicates abnormalities in the breast. Doctors need other information, such as a biopsy to detect breast cancer further. However removal of some tissue can cause bleeding, hematoma formation, and infection. A pattern recognition system is needed using mammogram images for breast cancer detection to avoid unnecessary biopsies. According to research conducted by experts from Kaiser Permanente in Oakland, California, breast tissue density can be one of the factors that determine whether a woman is at risk for breast cancer or not. Breast tissue density is always associated with cancer risk. The denser the breast, the more vulnerable it is to be attacked by cancer. This paper proposes a technique classification of breast tissue density into Glandular, Dense Glandular, or Fatty Glandular groups. The features used are mean, kurtosis, skewness, contrast, correlation, energy, and homogeneity. The proposed system consists of two main stages, namely (a) Performing feature extraction using Grayscale and Gray Level Co-occurrence Matrix (GLCM); (b) Compile transaction data and build a classification model. The evaluation results using the Tree and Random Forest algorithms are the accuracy rate is 92% (Tree), 95% (Random Forest).

Keywords— Breast cancer, GLCM, Region of Interest, Tree, Random Forest

I. INTRODUCTION

Breast cancer is a chronic disease, and a total cure is still very doubtful and requires a long treatment period and high costs. There are many ways to diagnose Breast cancer, including mammography, X-ray examination technique for soft tissue, which has proven effective indicating abnormalities of the breast [1]. Understanding of mammogram images to arrive at a diagnosis is a complicated thing because there are many steps that must be done, such as image processing, pattern recognition, segmentation, classification, and conclusions [2]. This process requires comprehensive knowledge in many fields, so it is interesting to study, primarily to obtain relevant features to breast cancer. A specialist can identify breast abnormalities visually by looking at the features seen on a mammogram. From the characteristics of the visually visible mammography image, expert doctors can classify breast tumors into two groups, namely benign tumors or malignant tumors [3]. Breast tissue density can be one of the factors that determine whether a woman is at risk for breast cancer or not [4]. Breast tissue density is always associated with cancer risk. The denser the breast, the more vulnerable it is to be attacked by cancer. The purpose of this study proposes a technique to classify breast tissue density into Glandular (G), Fatty Glandular (F), or Dense Glandular (D) groups [5] using texture feature extraction based on Gray Level Co-occurrence Matrix (GLCM). Figure 1. shows some types of breast tissue.

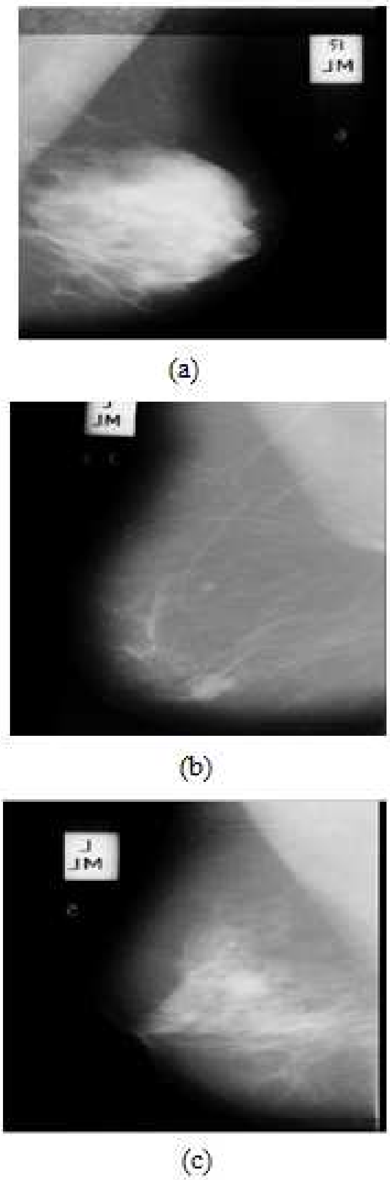

Fig 1. Types of breast tissue: (a) Grandular, (b) Fatty, (c) Dense

Several previous studies have discussed mammogram analysis using algorithms. The algorithm is able to extract features from the image for the desired region and is able to classify the malignancy from the mammogram. GLCM is capable of extracting features on mammograms [6][7][8]. Mammogram image research using the GLCM method to extract features, features with GLCM using 4 directions (0° ,45° ,90° ,135°) and distance = 1 can be used to distinguish between cystic mass and non-cystic mass including myoma images and solid tumor images on ultrasound images. The methods compared are histogram intensity, GLCM, and intensity based on features. From these results, feature extraction using GLCM is the best extraction method [9].

Several studies have examined the classification in 2 classes. In this study will classify into 3 classes. The preprocessing stage carried out is the conversion of the original image to grayscale, interpolation for resample images, prices cropping, image enhancement and adaptive thresholding. GLCM methods and statistical analysis are used to get the value of the features used as parameters. The classification stage uses the Tree and Random Forest algorithms because it is able to classify very well and explore data and be able to find hidden relationships.

II. Material and Method

The data used in this study were 322 images containing position information, individual mass size and microcalcifications, abnormal class types, and tissue type mammograms obtained from the Mammogram Image Analysis Society database.

Figure 2 is a classification stage which includes preprocessing, feature extraction using gray scale and GLCM, building classifier and evaluation model.

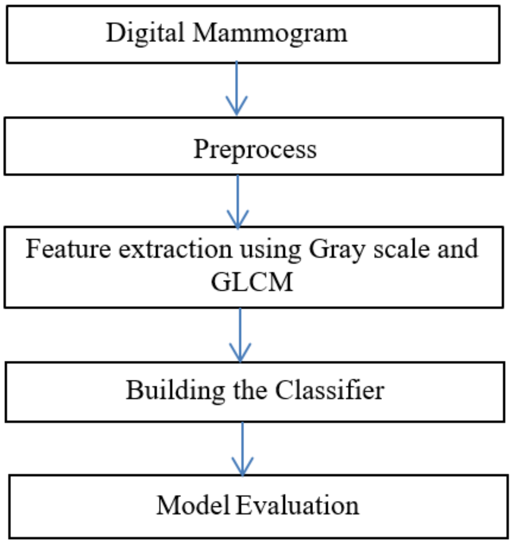

Fig. 2. Tissue Classification Method

A. Preprocess

Preprocess is the initial stage in breast cancer detection, this process is more about improving the quality of the mammogram image by increasing the intensity of the image between the image area and the object through highlighting features and reducing the effects of being too dark and light [10]. In addition, cropping, histogram equalization, median filter method are also options [11]. In this preprocessing, prices cropping, image enhancement and adaptive thresholding are carried out. The purpose of this preprocessing is to obtain more accurate segmentation results. At this stage it does not generate the type of tissue.

TABLE I. Sample Database

Tisue	Mean	Kurtosis	Skewness	Contrast	Correlation	Energy	Homogeneity
G	188.42	4.76	0.94	0.16	0.95	0.15	0.91
F	129.07	51.42	1.20	0.34	0.74	0.23	0.83
F	163.68	2.51	0.09	0.20	0.88	0.27	0.89
F	133.22	2.57	-0.51	0.14	0.95	0.17	0.92
G	197.59	4.33	-0.77	0.19	0.92	0.17	0.90
G	162.86	2.75	-0.01	0.26	0.91	0.12	0.86
G	179.92	4.66	-0.97	0.21	0.95	0.11	0.89
F	123.65	3.45	0.28	0.32	0.88	0.12	0.84
G	196.60	2.64	-0.58	0.23	0.93	0.15	0.88
G	193.75	2.84	0.008	0.33	0.86	0.13	0.83
D	181.18	6.54	-0.77	0.11	0.96	0.18	0.94
F	135.04	5.80	-0.13	0.23	0.90	0.15	0.88

B. Feature extraction

This study uses a second-order texture analysis that applies second-order statistical feature extraction using a co-occurrence matrix, which is an intermediate matrix that represents the neighboring relationship between pixels in the image in various orientations and spatial distances. In GLCM for the second order statistic to determine the texture, entity contrast, correlation, energy, homogeneity is used [11], while in the first order the mean, skewness and kurtosis are used [12]. In addition, the average of the seven orientations is also used as an additional feature. Figure 3 explains of the seven features calculated in the 256 x 256 region of interest then the feature extraction results are stored in a transactional database which can be seen in Table 1. Table 2 shows average value every features. Figure 4 shows GLCM Feature Extraction.

The following is the calculation for the GLCM feature[13]:

Contrast is a measure of the gray level of pixels, calculated by the formula:

$$\sum_{i,j} |i-j|^2 p(i,j)$$

Correlation is a measure of the dependence of the gray level on pixels, calculated by the formula:

$$\sum_{i,j} \frac{(i-\mu i)(j-\mu j)p(i,j)}{\sigma_i \sigma_j}$$

Energy is a measure that expresses the distribution of pixel intensity over the range of gray levels, calculated by the formula:

$$\sum_{i,j} p(i,j)^2$$

Homogeneity is used to measure homogeneity, calculated by the formula:

$$\sum_{i,j} \frac{p(i,j)}{1+|i-j|}$$

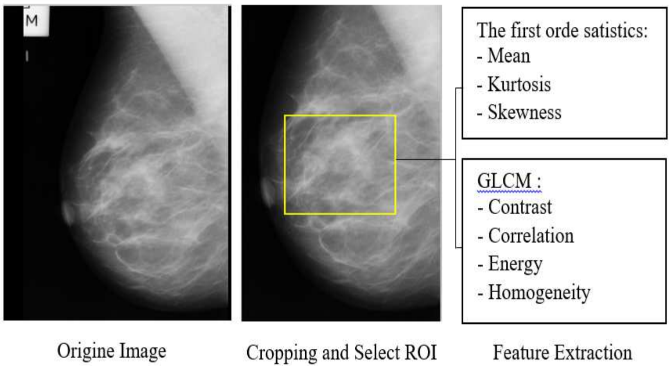

Fig 3. Feature extraction phase

C. Building Classifier

The basis of a decision tree is to make a decision rule from a data set. Decision trees are able to break down complex decision-making processes into simple ones, making it easier to interpret solutions. Tree is able to explore data and find hidden relationships between a number of variables. Tree combines data modeling and exploration makes for a great first step [14]. This decision tree can overlap, especially when the class and criteria used are very large, of course it can increase the decision-making time according to the amount of memory needed. In terms of accumulation, decision trees also often experience error problems, especially in large numbers. In addition, there are also difficulties in designing an optimal decision tree. Moreover, considering that the quality of decisions obtained from the decision tree method is very dependent on how the tree is designed [15]. So we need a Random Forest to overcome the overlap above.

Random forest is a classification consisting of several decision trees. Each decision tree is constructed using random vectors. The basis of a random forest is to create a random collection of trees from an attribute, with the aim of making tree creation and analysis faster. Thus the tree that is created will only have a few attributes. The accuracy of random forest will logically improve from Tree, this is because the classification results are generated from several trees and do not depend on only one tree [16]. A random collection of trees is generated by a random forest in a tree-like manner. Then in the determination using a voting model selected from all trees [17]. Random forest is a combination of each good tree which is then combined into one model. Random Forest depends on a random vector value with the same distribution in all trees where each decision tree has a maximum depth. A random forest is a classifier consisting of a classifier in the form of a tree {h(x, k), k = 1, . . .} where k is an independently distributed random vector and each tree in a unit will choose the most popular class on input x. Following are the characteristics of accuracy in random forest: Focusing on random forest, Strength and Correlation, Random Forest using random input selection, Random Forest using a linear combination of inputs.

TABLE II. AVERAGE VALUE EVERY FEATURES

Feature	Fatty (F)	Dense (D)	Grandular (G)
Mean	147,753	169,617	161,293
Kurtosis	12,121	6,725	7,073
Skewness	0,374	0,101	0,053
Contrast	0,265	0,151	0,181
Correlation	0,876	0,951	0,940
Energy	0,176	0,190	0,181
Homogeneity	0,871	0,925	0,910

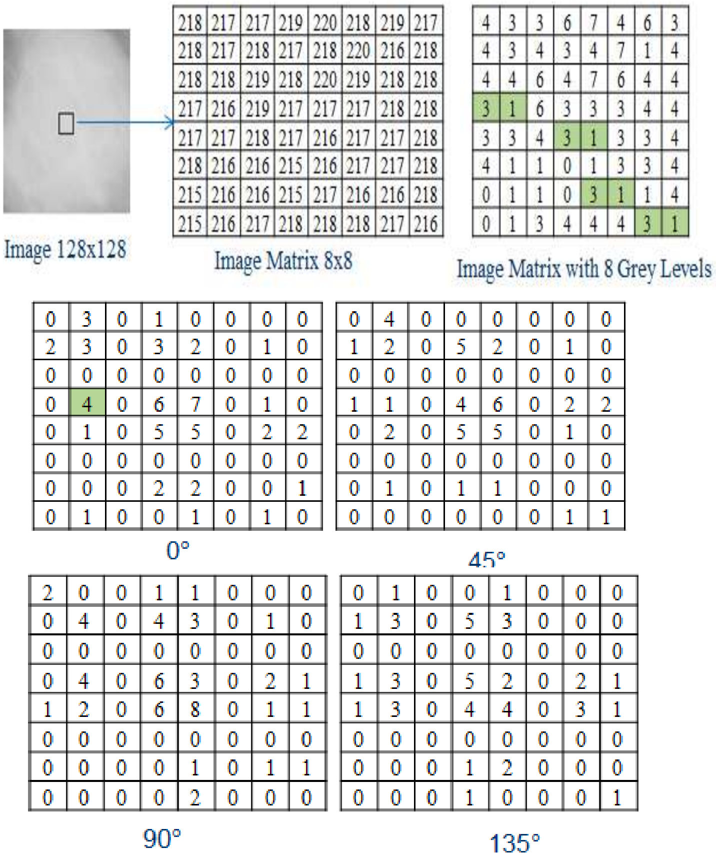

Image Matrix 8x8

218	217	217	219	220	218	219	217
218	217	218	217	218	220	216	218
218	218	219	218	220	219	218	218
217	216	219	217	217	217	218	218
217	217	218	217	216	217	217	218
218	216	216	215	216	217	217	218
215	216	216	215	217	216	216	218
215	216	217	218	218	218	217	216

Image Matrix with 8 Grey Levels

4	3	3	6	7	4	6	3
4	3	4	3	4	7	1	4
4	4	6	4	7	6	4	4
3	1	6	3	3	3	4	4
3	3	4	3	1	3	3	4
4	1	1	0	1	3	3	4
0	1	1	0	3	1	1	4
0	1	3	4	4	4	3	1

0°

0	3	0	1	0	0	0	0
2	3	0	3	2	0	1	0
0	0	0	0	0	0	0	0
0	4	0	6	7	0	1	0
0	1	0	5	5	0	2	2
0	0	0	0	0	0	0	0
0	0	0	2	2	0	0	1
0	1	0	0	1	0	1	0

45°

0	4	0	0	0	0	0	0
1	2	0	5	2	0	1	0
0	0	0	0	0	0	0	0
1	1	0	4	6	0	2	2
0	2	0	5	5	0	1	0
0	0	0	0	0	0	0	0
0	1	0	1	1	0	0	0
0	0	0	0	0	0	1	1

90°

2	0	0	1	1	0	0	0
0	4	0	4	3	0	1	0
0	0	0	0	0	0	0	0
0	4	0	6	3	0	2	1
1	2	0	6	8	0	1	1
0	0	0	0	0	0	0	0
0	0	0	0	1	0	1	1
0	0	0	0	2	0	0	0

135°

0	1	0	0	1	0	0	0
1	3	0	5	3	0	0	0
0	0	0	0	0	0	0	0
1	3	0	5	2	0	2	1
1	3	0	4	4	0	3	1
0	0	0	0	0	0	0	0
0	0	0	1	2	0	0	0
0	0	0	1	0	0	0	1

Fig 4. GLCM feature extraction

D. Model Evaluation

After modeling, it is necessary to carry out the process of evaluating or validating the model. This process is needed to choose the best model. In this paper, the technique used to measure the performance of the model uses a confusion matrix. The confusion matrix is a predictive analytic tool that displays and compares the actual value or the actual value with the predicted model value that can be used to generate evaluation metrics such as Accuracy (accuracy), Precision, Recall, and F1-Score or F-Measure. [18]. Table 3 below is a confusion matrix with four different combinations of predicted values and actual values. There are four terms as a representation of the results of the classification process in the confusion matrix. The four terms are True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The confusion matrix formed measures the model's performance, namely accuracy, precision, and recall.

Accuracy is the ratio of correct predictions (positive and negative) to the overall data. Precision is the ratio of positive correct predictions to the overall positive predicted results. Recall is the ratio of true positive predictions compared to the total number of true positive data.

TABLE III. CONFUSION MATRIX

		Actual Values	
		True	False
Prediction	True	TP Correct result	FP Unexpected result
	False	FN Missing result	TN Correct absence of result

III. RESULT AND DISCUSSION

After preprocessing, the next step is to build a classification model using Tree and Random Forest. The results of the three models are evaluated by measuring the success of the classification results based on the Accuracy, Precision, and Recall parameters.

The distribution of training data and testing data with proportions of 70 and 30 and the results of modeling the training data with the Tree algorithm obtained the value of accuracy = 0.93, precision = 0.94, and recall = 0.98 with 61 dense glandular predicted correctly and 1 incorrectly predicted, 55 fatty glands predicted correct and 10 predicted incorrectly, as well as a total of 68 predicted correct for glandular and 2 incorrect predictions. The results of modeling the training data with Random Forest obtained the value of accuracy = 0.93, precision = 0.92, and recall = 0.92 with 60 dense glandular which were predicted correctly and 5 were predicted incorrectly, 52 fatty glands were predicted to be correct and 1 were predicted to be incorrect, and a total of 73 were predicted to be correct for glandular and 7 were predicted to be incorrect.

TABLE IV. RESULTS OF TRAINING DATA

			Actual			**Result**		
			D	F	G	Accuracy	Precision	Recall
Prediction	Tree	D	61	2	2	0.93	0.94	0.98
		F	1	55	0			
		G	0	8	68			
	Random Forest	D	60	1	4	0.93	0.92	0.92
		F	1	52	3			
		G	4	0	73			

IV. CONCLUSION

Based on the test results, machine learning succeeded in classifying mammogram tissue into three categories, namely Glandular, Fatty Glandular, and Dense Glandular. The Tree algorithm has the same accuracy value as Random Forest with an accuracy value of 0.93, but the precision and recall values are higher with the Tree algorithm. The precision value for the Tree is 0.94 and the Random Forest is 0.92, the recall value for the Tree is 0.98 and the Random Forest is 0.92. The performance of the model on the training data built using the Tree algorithm is better than the Random Forest.

REFERENCES

[1] L. Heck and J. Herzen, “Recent advances in X-ray imaging of breast tissue: From two- to three-dimensional imaging,” Phys. Medica, vol. 79, pp. 69–79, Nov. 2020, doi: 10.1016/j.ejmp.2020.10.025.

[2] K. L. Kashyap, M. K. Bajpai, and P. Khanna, “Globally supported radial basis function based collocation method for evolution of level set in mass segmentation using mammograms,” Comput. Biol. Med., vol. 87, pp. 22–37, Aug. 2017, doi: 10.1016/j.compbiomed.2017.05.015.

[3] Y. Shen et al., “Artificial intelligence system reduces false-positive findings in the interpretation of breast ultrasound exams,” Nat. Commun., vol. 12, no. 1, p. 5645, Dec. 2021, doi: 10.1038/s41467-021-26023-2.

[4] K. Kerlikowske, D. L. Miglioretti, and C. M. Vachon, “Discussions of Dense Breasts, Breast Cancer Risk, and Screening Choices in 2019,” JAMA, vol. 322, no. 1, p. 69, Jul. 2019, doi: 10.1001/jama.2019.6247.

[5] D. Christopher and P. Simon, “A Novel Approach for Mammogram Enhancement using Nonlinear Unsharp Masking and L0 Gradient Minimization,” Procedia Comput. Sci., vol. 171, pp. 1848–1857, 2020, doi: 10.1016/j.procs.2020.04.198.

[6] V. Nagarajan, E. C. Britto, and S. M. Veeraputhiran, “Feature extraction based on empirical mode decomposition for automatic mass classification of mammogram images,” Med. Nov. Technol. Devices, vol. 1, p. 100004, Mar. 2019, doi: 10.1016/j.medntd.2019.100004.

[7] S. P. A. Kirubha, M. Anburajan, B. Venkataraman, and M. Menaka, “A case study on asymmetrical texture features comparison of breast thermogram and mammogram in normal and breast cancer subject,” Biocatal. Agric. Biotechnol., vol. 15, pp. 390–401, Jul. 2018, doi: 10.1016/j.bcab.2018.07.001.

[8] 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, doi: 10.1016/j.measurement.2019.05.083.

[9] A. Mohd. Khuzi, R. Besar, W. Wan Zaki, and N. Ahmad, “Identification of masses in digital mammogram using gray level co-occurrence matrices,” Biomed. Imaging Interv. J., vol. 5, no. 3, Jul. 2009, doi: 10.2349/biij.5.3.e17.

[10] E. B. Cole et al., “The Effects of Gray Scale Image Processing on Digital Mammography Interpretation Performance1,” Acad. Radiol., vol. 12, no. 5, pp. 585–595, May 2005, doi: 10.1016/j.acra.2005.01.017.

[11] D. Tohl and J. S. J. Li, “Contrast enhancement by multi-level histogram shape segmentation with adaptive detail enhancement for noise suppression,” Signal Process. Image Commun., vol. 71, pp. 45–55, Feb. 2019, doi: 10.1016/j.image.2018.10.011.

[12] J. Xiang, E. Maue, H. Fujiwara, F. T. Mangano, H. Greiner, and J. Tenney, “Delineation of epileptogenic zones with high frequency magnetic source imaging based on kurtosis and skewness,” Epilepsy Res., vol. 172, p. 106602, May 2021, doi: 10.1016/j.eplepsyres.2021.106602.

[13] M. Yogeshwari and G. Thailambal, “Automatic feature extraction and detection of plant leaf disease using GLCM features and convolutional neural networks,” Mater. Today Proc., May 2021, doi: 10.1016/j.matpr.2021.03.700.

[14] T. Lan, H. Hu, C. Jiang, G. Yang, and Z. Zhao, “A comparative study of decision tree, random forest, and convolutional neural network for spread-F identification,” Adv. Sp. Res., vol. 65, no. 8, pp. 2052–2061, Apr. 2020, doi: 10.1016/j.asr.2020.01.036.

[15] J. Olivier and C. Aldrich, “Use of Decision Trees for the Development of Decision Support Systems for the Control of Grinding Circuits,” Minerals, vol. 11, no. 6, p. 595, May 2021, doi: 10.3390/min11060595.

[16] M. C. E. Simsekler, A. Qazi, M. A. Alalami, S. Ellahham, and A. Ozonoff, “Evaluation of patient safety culture using a random forest algorithm,” Reliab. Eng. Syst. Saf., vol. 204, p. 107186, Dec. 2020, doi: 10.1016/j.ress.2020.107186.

[17] J. Xia et al., “Adjusted weight voting algorithm for random forests in handling missing values,” Pattern Recognit., vol. 69, pp. 52–60, Sep. 2017, doi: 10.1016/j.patcog.2017.04.005.

[18] S. Ruuska, W. Hämäläinen, S. Kajava, M. Mughal, P. Matilainen, and J. Mononen, “Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle,” Behav. Processes, vol. 148, pp. 56–62, Mar. 2018, doi: 10.1016/j.beproc.2018.01.004.