Indonesian Journal of Electrical Engineering and Computer Science
Vol. 22, No. 2, May 2021, pp. 795~800
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v22.i2.pp795-800

Gray level co-occurrence matrix feature extraction and histogram in breast cancer classification with ultrasonographic imagery

Karina Djunaidi, Herman Bedi Agtriadi, Dwina Kuswardani, Yudhi S. Purwanto
Faculty of Telematics for Energy, Institut Teknologi PLN, Jakarta, Indonesia

Article Info

Article history:

Received Mar 6, 2020
Revised Jan 13, 2021
Accepted Feb 1, 2021

Keywords:

Breast cancer
Gray level co-occurrence
Matrix
Histogram
K-nearest neighbour
Ultrasonographic

ABSTRACT

One way to detect breast cancer is using the ultrasonography (USG) procedure, but the ultrasound image is susceptible to the noise speckles so that the interpretation and diagnosis results are different. This paper discusses the classification of breast cancer ultrasound images that aims to improve the accuracy of the identification of the type and level of cancer malignancies based on the features of its texture. The feature extraction process uses a histogram which then the results are calculated using the gray level co-occurrence matrix (GLCM). The results of the two extraction features are then classified using k-nearest neighbors (KNN) to obtain accurate figures from those images. The results of this study is that the accuracy in detecting cancer types is 80%.

Corresponding Author:

Karina Djunaidi
Faculty of Telematics for Energy
Institut Teknologi PLN
Menara PLN, Jl. Lingkar Luar Barat, Duri Kosambi, Cengkareng Jakarta Barat 11750, Indonesia
Email: karina@itpln.ac.id

1. INTRODUCTION

Breast cancer (carninoma mamae) is one of the most malignant types of cancer for women worldwide along with cervical cancer and ovarian cancer [1]. In developing as well as under-developing countries, breast cancer is the leading cause of mortality among women [2]. As it happened in Indonesia, breast cancer was the highest disease with 42.1 per 100,000 population with an average mortality rate of 17 per 100,000 population [3]. Breast cancer is a cancer of the breast tissue [4] and becomes a disease that really alarming for women, yet breast cancer is not only attacks women but also men [5].

Breast cancer detection is done by several screening procedures such as mammography, magnetic resonance imaging (MRI), biopsy, and ultrasonography (USG). The images from the examination results are very important for accurate diagnosis of breast cancer [6]. This research focuses on the classification of breast cancer using USG images. Ultrasonography is one of the imaging modalities that can detect and classify breast mass abnormalities [7]. USG can be applied to examine the body by utilizing ultrasonic waves to get image of body tissue which in this case: breast tissue. USG is considered cheaper, portable, and easier to find in health facilities than MRI [8]. But ultrasonographic images are very susceptible to noise speckle that appears at the time of acquisition or degradation of unclear edge characters [7], [8]. However, this matter depends on the operator at the time of examination, such as the examination techniques and module characteristics applied. Some of different interpretations and diagnosis results on the image can happen because of the experience or expertise of the USG operator.

The purpose of this study is to improve the accuracy of the identification of the type and degree of malignancy of breast cancer based on texture and classification features. The first process to do is extracting features. The feature extraction process is the images classifying based on the characteristics of the images [9] by calculating the value of the results of segmentation using a histogram. The results are then calculated using the gray level co-occurrence matrix (GLCM) method. The results of the value of the two extraction features are then recalculated in the classification by applying the KNN method to get an accurate number from the image. gray level co-occurrence matrix (GLCM) is one of the most famous texture analysis methods especially for irregular textures, by calculating frequent set of the pixels in a predetermined spatial connection and particular qualities transpire in an image [10], [11]. The way it works is by calculating the relationship between two adjacent pixels where the first pixel is a reference pixel and the second pixel is a neighboring pixel [6]. There are 5 stages in the GLCM method including: quantization, co-occurrence, symmetric, normalization and feature extraction [12]. In feature extraction, there are 3 parameters such as contrast, homogeneity, and energy [13] The GLCM analysis method can be applied in various fields including face retrieval system [13], Classification for chicken embryo recognition [14] as well as in breast cancer.

The k-nearest neighbor (KNN) algorithm is one of the top algorithm in data mining and has been widely used in the fields of pattern recognition, regression, and other data mining [15]-[17]. The procedure of this algorithm requires the researcher to determine the input in the form of training data, test data, and k-values, then the training data are sorted by distance proximity by calculating the distance from the tested data and training data, and taken k-top training data to determine the dominant classification class [16]. In a training data, the working principle of the KNN algorithm is to find the closest distance between the data to be evaluated with the nearest K neighbor [18], [19]. KNN algorithm can be applied in various fields such as Wi-Fi fingerprint positioning [20], diagnosis of diabetes [21], as well as in predicting green consumption behavior of students [22].

2. RESEARCH METHOD

Figure 1 shows the research steps which will be covered. The discussion of this research focuses on the image feature extraction process using the histogram and the GLCM method, then the results of the extraction process are classified using the KNN algorithm to identify the type and degree of breast cancer's malignancy. The data used in the study were ultrasound images from the laboratory, i.e. 25 malignant cancer samples and 25 benign cancer samples. Malignant cancer is a group of fatal breast cancer while benign cancer is a group of mild breast cancer [23], [24]. The format of the data is jpg with an image size of 64 x 64 pixels. After entering the segmented image data, the data will be extracted with a histogram and GLCM. Histogram is representing a graph of the probability distribution of gray values in a digital image. Visualization of the histogram image can help to analyze the frequency of gray levels contained in the image [25]. The parameters to be calculated are mean, standard deviation, skewness, and entropy [26], [27], whereas the GLCM parameters calculated are contrast parameters, energy parameters, and homogeneity parameters. Then the results will be classified with the reference data using the KNN classification method to determine the type of breast cancer.

Figure 1. USG image identification process for breast cancer classification

2.1. Histogram

There are 4 features that are used in Histogram and taken as a reference for classification, namely: mean, skewness, standard deviation and entropy. The result of the histogram is a histogram feature which will then be classified at a later stage. The steps taken in the histogram process are as follows:

a) Input the image to be processed.
b) The first process is to change the photo (color image) into a grayscale image: $0.299R + 0.587G + 0.114B$
c) Calculating *mean*:

$$m = \sum_{n=0}^{255} f_n . p(n)$$

d) Calculating standard deviation value:

$$\sigma = \sqrt{\sum_{n=0}^{255}(f_n - m)^2 \cdot p(n)}$$

e) Calculating asymmetry value against the average intensity or *skewness* feature:

$$\frac{1}{\sigma^3}\sum_{n=0}^{255}(f_n - m)^3 \cdot p(n)$$

f) Calculating value of image complexity or entropy feature:

$$\sum_{n=0}^{255} p(n).\log_2 p(n)$$

Information: f_n = frequency at n[th] intensity; $p(n)$ = probability value at n[th] intensity; m mean value or mean intensity; σ = standard deviation value.

2.2. Gray level co-occurance matrix (GLCM)

In the GLCM process, the test data that has been entered will be calculated with GLCM parameters including energy, contrast and homogeneity. The following steps are taken in this process:

1) Input image to be processed
2) Change the photo (color image) to grayscale image using the formula: $0.299\mathrm{R} + 0.587\mathrm{G} + 0.114\mathrm{B}$
3) After the color image becomes a grayscale image, the next step is to convert the gray value of the original image from the range 0 - 255 into a new gray scale (in this study using a scale of 0 - 9) this means that each pixel value of the image will be smaller which aims to simplify the process of calculating GLCM elements and histograms in the form of the appearance of the gray value pairs of reference pixels and neighboring pixels.
4) The pixel value of the scale is made into a matrix, in this case the matrix obtained is a matrix of 100 × 100 because the size of the photo entered at the beginning has a size of 100 × 100 pixels which was changed into 64 × 64 pixels
5) Calculate the GLCM element that is the appearance of the gray value pairs of reference pixels and neighboring pixels at the distance and direction specified in the study using 1 pixel proximity and 0[o] direction.
6) The new matrix is obtained based on the calculation results
7) Calculate the features or parameters of the GLCM in this study. The GLCM parameters used are contrast, energy, and homogeneity. This parameter calculation uses the following formula:
   a) Contrast
      The function to produce contrast which is an amount of the distribution of elements in a matrix is defined in the following equation:

      $$\sum_i \sum_j (i-j)^2 P[i,j]$$

   b) Energy
      The function to produce *energy* is defined in the following equation:

      $$\sum_i \sum_j P^2[i,j]$$

   c) Homogeneity
      The function to produce *homogeneity* or gray level similarity level is defined in the following equation:

      $$\sum_i \sum_j \frac{P[i,j]}{1+|i-j|}$$

      Where (i, j) are the values in row (i) and column (j) and P is the GLCM matrix.

2.3. K-Nearest neighbour

Figures 2 explain KNN process which in this process training images and testing images are included. The parameters needed are the features of the histogram calculation and the features of the GLCM calculation. They are then be sought for the euclidean distance value in the image to be tested with all the

training images that already exist. After getting the euclidean distance value from the test image with all available training images, the euclidean distance is sorted from the smallest value. After that, 3 data were taken with the smallest euclidean distance which they will be classified. If there are more than 3 data from the benign data, then the tested image can be classified into the benign groups. Meanwhile, if there are more than 3 malignant data, then the tested image can be classified into a malignant group.

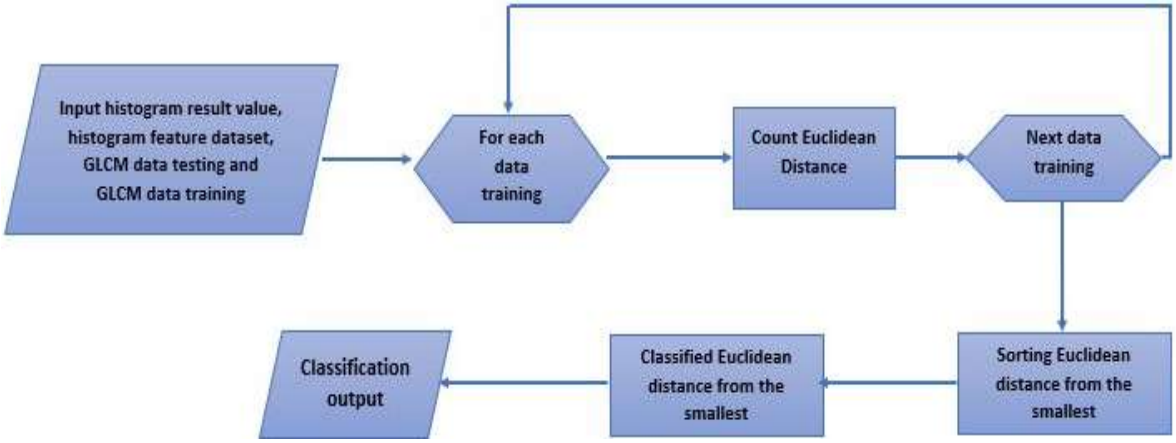
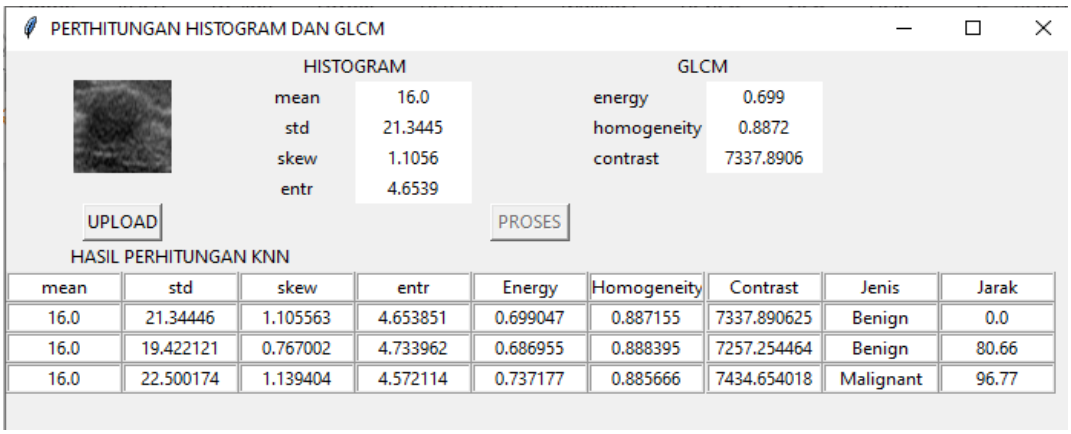


Figure 2. KNN process

3. RESULTS AND ANALYSIS

The GLCM and histogram in the previous stage were applied to software built using python in identifying breast cancer. The results of the histogram parameter calculation of the breast cancer image data and the results of the identification of the type of cancer based on the K-NN classification available in the software. There are two buttons on this display, the upload button to upload an ultrasound image and the process button to display the calculation results by the KNN method as shown in Figure 3.



mean	std	skew	entr	Energy	Homogeneity	Contrast	Jenis	Jarak
16.0	21.34446	1.105563	4.653851	0.699047	0.887155	7337.890625	Benign	0.0
16.0	19.422121	0.767002	4.733962	0.686955	0.888395	7257.254464	Benign	80.66
16.0	22.500174	1.139404	4.572114	0.737177	0.885666	7434.654018	Malignant	96.77

Figure 3. Calculation result display

After the calculation is done, the values are seen with the smallest distance, so it will match the type of test data as shown in the display above. Figure 3 shows the value that has the proximity of the top 3 neighbors, so the results of the classification of test data that shows the type of benign cancer. After the results of identification of the test data are known, then the reference data processing process ends.

3.1. Confusion matrix evaluation result and analysis

Table 1 is a confusion matrix table [28] used to test the accuracy of the results of GLCM calculations and histograms in the previous stage.

Table 1. The confusion matrix breast cancer types test result

Actual	Prediction	
	Benign	Malignant
Benign	9	1
Malignant	3	7

$$\text{Accuracy} = \frac{BB+MM}{BB+BM+MB+MM}$$

Explanation: B = Benign; M = Malignant
Based on the table above the accuracy presentation value obtained is:

1. Overall accuracy: $\frac{(9+7)}{(9+1+3+7)}\ x\ 100\% = 80\%$

3. M Accurary (Malignant): $\frac{7}{(3+7)}\ x\ 100\% = 70\%$

2. B Accuracy (Benign):

$$\frac{9}{(9+1)}\ x\ 100\% = 90\%$$

This software is generally made to apply a histogram that can be used to analyze color features in cancer images, the GLCM method to be able to analyze textures in breast cancer image data and the k-nearest neighbor (KNN) to classify test data with reference data so that it can recognize the types of the cancer. The features used in the histogram are the mean, standard deviation, skewness and entropy. While the GLCM parameters used to analyze the texture of breast cancer are energy, contrast and homogeneity. Before the GLCM parameter is calculated, the first step taken is that the gray image from the range 0-255 is changed to range 0-9 so that the system's load in processing data becomes lighter. The pixel value of the scale results is made into the original matrix and then the next step is to create a co-occurrence matrix by counting the number of occurrences of the reference pixel and neighboring pixel's gray values in the direction of 45° with proximity of 1 pixel. Each parameter of reference data that is calculated will be stored in a semi-database. The KNN method is used to classify each GLCM feature and histogram data that has been tested in the previous stage so that the type of breast cancer from the classification results can be identified. This study uses 50 breast cancer image data used as reference data consisting of 25 benign type cancer image data and 25 malignant type cancer image data.

4. CONCLUSION

Based on the testing results on the suitability of the histogram results and GLCM calculations from test data and reference data classified by the K-NN method, there was an identification error in 10 image data tested, namely 1 error in benign breast cancer type and 1 error in malignant breast cancer type. So, from each type of breast cancer resulted 80% accuracy of the total; 90% from benign breast cancer type and 70% from malignant breast cancer type. Future work for the development of this research is to apply or to combine this method with other methods so that it can improve the accuracy in detecting types of cancer.

ACKNOWLEDGEMENTS

This research is supported and fully funded by Institut Teknologi PLN.

REFERENCES

[1] M. Ghoncheh, Z. Pournamdar e H. Salehiniya, "Incidence and Mortality and Epidemiology of Breast Cancer in the World," *Asian Pacific Journal of Cancer Prevention,* vol. 17, pp. 43-46, 2016, doi: 10.7314/APJCP.2016.17.S3.43.
[2] S. U. Khan, N. Islam, Z. Jan, I. U. Din e J. J. Rodrigues, "A novel deep learning based framework for the detection and classification of breast cancer using transfer learning," *Pattern Recognition Letter,* vol. 125, pp. 1-6, 2019.
[3] Indonesian ministry of health, "World cancer Day", January 2019. [Online]. https://www.kemkes.go.id/article/view/19020100003/hari-kanker-sedunia-2019.html
[4] H. Goto, *et al.,* "Adipose-derived stem cells enhance human breast cancer growth and cancer stem cell-like properties through adipsin," *Oncogene,* vol. 38, no. 6, pp. 767-779, 2019.

[5] D. M. Vo, N.-Q. Nguyen e S. W. Lee, "Classification of breast cancer histology images using incremental boosting convolution networks," *Information Sciences*, vol. 482, pp. 123-138, 2019, doi: 10.1016/j.ins.2018.12.089.
[6] R. S e N. S, "Breast Cancer Detection and Classification Using Ultrasound and Ultrasound Elastography Images," *International Research Journal of Engineering and Technology (IRJET)*, vol. 4, no. 7, pp. 596-601, 2017.
[7] M. Rahmawaty, H. A. Nugroho, Y. Triyani, I. Ardiyanto e I. Soesanti, "Classification of Breast Ultrasound Images based on Texture Analysis," em *1st International Conference on Biomedical Engineering (IBIOMED)*, Yogyakarta, pp. 1-6, 2016, doi: 10.1109/IBIOMED.2016.7869825.
[8] H. A. Nugroho, Y. Triyani, M. Rahmawaty e I. Ardiyanto, "Performance Analysis of Filtering Techniques for Speckle Reduction on Breast Ultrasound Images," *International Electronics Symposium*, pp. 450-454, 2016, doi: 10.1109/ELECSYM.2016.7861048.
[9] D. C. R. Dian Candra Rini Novitasari, A. Lubab, A. Sawiji e A. H. Asyhar, "Application of Feature Extraction for Breast Cancer using One Order Statistic, GLCM, GLRLM, and GLDM," *Advances in Science, Technology and Engineering Systems Journal*, vol. 4, no. 4, pp. 115-120, 2019, doi: 10.25046/aj040413.
[10] F. Albregtsen, "Statistical texture measures computed from gray level coocurrence matrices," *Image processing laboratory, department of informatics, university of oslo*, vol. 5, no. 5, 2008.
[11] F. E. Batool, *et al.*, "Offline signature verification system: a novel technique of fusion of GLCM and geometric features using SVM," *Multimedia Tools and Aplications*, doi: 10.1007/s11042-020-08851-4, 2020.
[12] M. Hall-Beyer, "Gray Level Co-occurrence Matrix," Maret 2017. [Online]. Available: http://www.fp.ucalgary.ca/mhallbey/the_glcm.html. [Acesso em 28 February 2020].
[13] S. A. Alazawi, N. M. Shati e A. H. Abbas, "Texture features extraction based on GLCM for face retrieval system," *Periodicals of Engineering and Natural Sciences*, vol. 7, no. 3, pp. 1459-1467, 2019, doi: 10.21533/pen.v7i3.787.
[14] W. Lumchanow e S. Udomsiri, "Combination of GLCM and KNN Classification for Chicken Embryo Development Recognition," em *Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON)*, Nan Thailand, 2019, doi: 10.1109/ECTI-NCON.2019.8692272.
[15] J. Gou, W. Qiu, Z. Yi, Y. Xu, Q. Mao e Y. Zhan, "A Local Mean Representation-based K-Nearest Neighbor Classifier," *ACM Transaction on Intelligent System and Technology*, vol. 10, no. 3, pp. 1-25, 2019, doi: 10.1145/3319532.
[16] Y. Pan, Z. Pan, Y. Wang e W. Wang, "A new fast search algorithm for exact k-nearest neighbors based on optimal triangle-inequality-based check strategy," *Knowledge Based System Journal*, vol. 189, 2020, doi: 10.1016/j.knosys.2019.105088.
[17] W. Zhang, C. Xiaohui, L. Yueqi e X. Qian, "A Distributed Storage and Computation k-Nearest Neighbor Algorithm Based Cloud-Edge Computing for Cyber-Physical-Social Systems," *IEEE Access*, vol. 8, 2020, doi: 10.1109/ACCESS.2020.2974764.
[18] S. Zhang, Z. Deng, D. Cheng, M. Zong e X. Zhu, "Efficient kNN Classification Algorithm for Big Data," *Neurocomputing*, vol. 195, pp. 143-148, 2016, doi: 10.1016/j.neucom.2015.08.112.
[19] A.-J. Gallego, J. C. Zaragoza, J. J. Valero-Mas e J. R. Rico-Juan, "Clustering-based k-nearest neighbor classification for large-scale data with neural codes representation," *Patern Recognition*, vol. 74, pp. 531-543, 2018, doi: 10.1016/j.patcog.2017.09.038.
[20] B. Wang, X. Gan, X. Liu, B. Yu, R. Jia, L. Huang e H. Jia, "A Novel Weighted KNN Algorithm Based on RSS Similarity and Position Distance for Wi-Fi Fingerprint Positioning," *IEEE Access*, vol. 8, pp. 30591-30602, 2020.
[21] A. Ali, M. A. T. Alrubei, L. F. M. Hassan, M. A. M. Al-Ja'afari e S. H. Abdulwahed, "Diabetes Diagnosis Based On KNN," *IIUM Engineering Journal*, vol. 21, no. 1, pp. 175-181, 2020.
[22] H. Tang, Y. Xu, A. Lin, A. A. Heidari, M. Wang, H. Chen, Y. Luo e C. Li, "Predicting Green Consumption Behaviors of Students Using Efficient Firefly Grey Wolf-Assisted K-Nearest Neighbor Classifiers," *IEEE Access*, vol. 8, pp. 35546-35562, 2020, doi: 10.1109/ACCESS.2020.2973763.
[23] J. Zhou, *et al.*, "Diagnosis of Benign and Malignant Breast Lesions on DCE-MRI by Using Radiomics and Deep Learning With Consideration of Peritumor Tissue," *Journal of Magnetic Resonance Imaging*, vol. 51, no. 3, pp. 798-809, 2019, doi: 10.1002/jmri.26981.
[24] J.-S. Huang, H.-B. Pan, T.-L. Yan, B.-H. Hung, C. L. Chiang, M.-Y. Tsai e C.-P. Chou, "Kinetic patterns of benign and malignant breast lesions on contrast enhanced digital mammogram," *Plos One*, 2020, doi: 10.1371/journal.pone.0239271.
[25] N. Salem, H. Malik e A. Shams, "Medical image enhancement based on histogram algorithms," *International Learning & Technology Conference*, pp. 300-311, 2019.
[26] K. Nagasaka, H. Satake, S. Ishigaki, H. Kawai e S. Naganawa, "Histogram analysis of quantitative pharmacokinetic parameters on DCE-MRI: correlations with prognostic factors and molecular subtypes in breast cancer," *Breast Cancer*, vol. 26, no. 1, pp. 113-124, 2019, doi: 10.1007/s12282-018-0899-8.
[27] N.-N. Sun, X.-L. Ge, X.-S. Liu e L.-L. Xu, "Histogram analysis of DCE-MRI for chemoradiotherapy response evaluation in locally advanced esophageal squamous cell carcinoma," *La radiologia medica*, vol. 125, no. 2, pp. 165-176, 2020, doi: 10.1007/s11547-019-01081-1
[28] A. Luque, A. Carrasco, A. Martín e A. l. Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognition*, vol. 91, pp. 216-231, 2019, doi: 10.1016/j.patcog.2019.02.023.