International Journal of Technology (2016) 1: 71-77
ISSN 2086-9614

CLASSIFICATION OF DIGITAL MAMMOGRAM BASED ON NEAREST-NEIGHBOR METHOD FOR BREAST CANCER DETECTION

Anggrek Citra Nusantara[1*], Endah Purwanti[1], Soegianto Soelistiono[1]

[1]*Biomedical Engineering, Faculty of Science and Technology, Universitas Airlangga, Kampus C Universitas Airlangga, Surabaya 60115, Indonesia*

(Received: July 2015 / Revised: September 2015 / Accepted: January 2016)

ABSTRACT

Breast cancer can be detected using digital mammograms. In this research study, a system is designed to classify digital mammograms into two classes, namely normal and abnormal, using the k-Nearest Neighbor (kNN) method. Prior to classification, the region of interest (ROI) of a mammogram is cropped, and the feature is extracted using the wavelet transformation method. Energy, mean, and standard deviation from wavelet decomposition coefficients are used as input for the classification. Optimal accuracy is obtained when wavelet decomposition level 3 is used with the feature combination of mean and standard deviation. The highest accuracy, sensitivity, and specificity of this method are 96.8%, 100%, and 95%, respectively.

Keywords: Breast cancer; k-Nearest Neighbor; Mammogram; Wavelet transformation

1. INTRODUCTION

Breast cancer is the most frequently diagnosed cancer among women in the world. According to statistics published by the International Agency for Research on Cancer (IARC), there were 521,907 deaths from breast cancer in 2012 (Globocan, 2012). Early stage breast cancer detection is really helpful for reducing patient mortality (American Cancer Society, 2014). Breast cancer is usually detected using a mammogram after a physical examination is performed (Isource National Breast Cancer Centre, 2011). The mammogram technique is considered the most reliable technique for detecting breast cancer because it can display changes in the breast two years before a patient can sense the presence of cancer (Radiologyinfo, 2014).

Masses and microcalcifications are the two types of breast cancer indicators that can be seen on a mammogram (Michelle, 2010). Missed diagnoses can be caused by human factors such as subjectivity, distraction, or fatigue when reading a mammogram result. Thus, automatic classification using computer-aided diagnosis (CAD) is necessary. CAD can help a radiologist to interpret masses and microcalcifications because information from the mammogram images will be quantized. CAD can help to mark 77% of the cancer that cannot be detected by radiologists (Birdwell et al., 2001).

Feature extraction is an important step for detecting abnormality in images. Statistical properties, textures, and wavelet transformation are often used for extracting the features of an image. In CAD research studies on mammograms, texture is commonly used for image interpretation. The texture feature can be obtained by using the wavelet transform method.

*Corresponding author's email: anggrek.citra.n@gmail.com, Tel. +62-031-5922427, Fax. +62-031-5922427
Permalink/DOI: http://dx.doi.org/10.14716/ijtech.v7i1.1393

Wavelet transformation is a process used to analyze and reconstruct an image without losing information about the image (Putra, 2010).

Several schemes for mammogram analysis using wavelet transformation have been conducted by some researchers. Ferreira and Borges (2001) used the biggest wavelet coefficient as the input of the single nearest-neighbor classification method to classify mammogram images as benign or malignant. They obtained the 100 greatest coefficients of the decomposed image using wavelets Haar and Daubechies-4 in the first level of decomposition. Pratibha and Sadasivam (2010) also classified normal and abnormal mammogram images using the nearest-neighbor classifier. They compared feature extraction methods, ranging from those for gray-level co-occurrence matrix (GLCM) features to wavelet features and wavelet-based textural features. GLCM is used as a method to extract the second-order statistical texture feature. The researcher used the biggest wavelet coefficient (BWC), wavelet-transformed GLCM (WGLCM), and wavelet-transformed histogram statistics (WHS) to extract wavelet features. They found that the wavelet-based textural feature yielded the highest classification accuracy. Hamad et al. (2013) applied wavelet transformation to mammogram images to detect microcalcifications. The procedure consists of dimension reduction, feature extraction using one-dimensional (1-D) multi-resolution wavelet transformation, two dimensional (2-D) wavelet subband decomposition, and binary thresholding.

The approach in this paper differs from those of the abovementioned research studies. The objective is to obtain the best feature, one that gives the highest accuracy in classification using wavelet Haar in the feature extraction method to obtain the optimal level of wavelet decomposition with the k-Nearest Neighbor (kNN) as the classifier. The experiment is conducted using the Mammographic Image Analysis Society (MIAS) dataset. Firstly, the region of abnormality is cropped from each mammogram using the coordinates given in the ground truth file. Secondly, each of the mammogram images is decomposed using wavelet transformation in six different levels, and the set of wavelet coefficients of each mammogram is extracted. Finally, the kNN classifier is used to classify mammogram images to differentiate between normal and abnormal mammograms.

2. METHODOLOGY

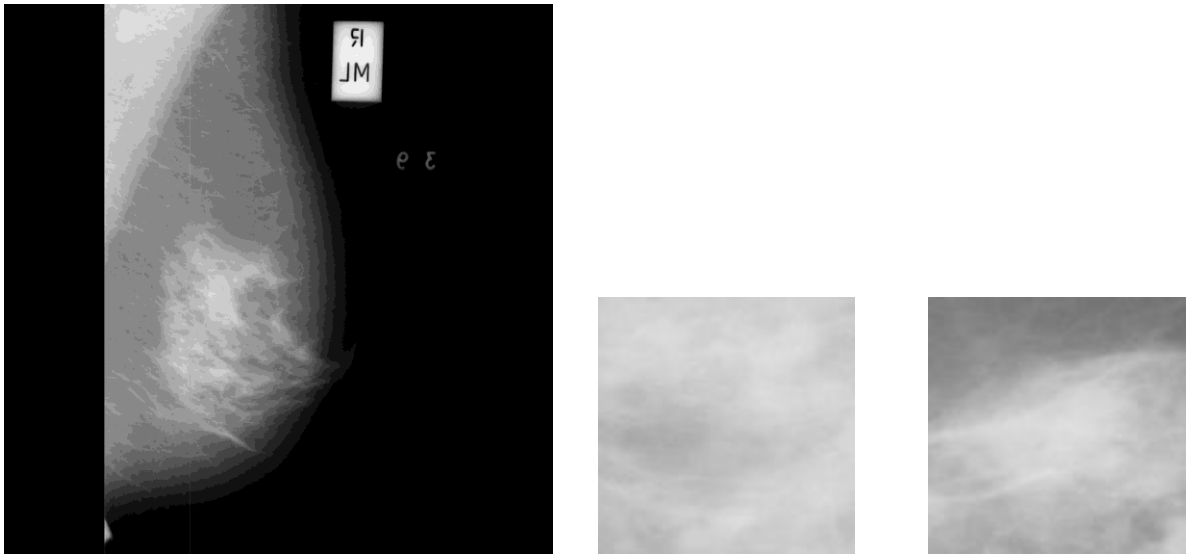
The proposed methodology comprises three main processes: mammogram acquisition and preprocessing, feature extraction, and classification. During the preprocessing step, the mammogram is cropped to obtain regions of interest (ROIs). Then, the ROI is decomposed using wavelet Haar in the feature extraction process to obtain the textural feature. Energy, mean, and standard deviation from wavelet coefficients are calculated in each wavelet decomposition level. Lastly, the feature sets are used in classification so that breast tissues can be discriminated using the kNN classifier into normal and abnormal.

2.1. Mammogram Acquisition and Preprocessing

Mammography data for this research are taken from the MIAS database (Suckling et al., 1994). It contains 322 mammogram images that are 1024×1024 pixels, 209 normal images, and 113 abnormal images that have been verified by radiologists. Every image contains ground truth information about the abnormalities, such as the type of cancer, severity of the diagnosis (benign or malignant), and coordinate location of the abnormality.

Each mammogram is cropped into an area of 128×128 pixels using the coordinates given in the ground truth file. Cropping is intended to obtain the ROI of that particular mammogram and to reduce the computational time required for processing the mammograms. Breast density is evaluated only in the fibroglandular disc; therefore, the entire breast area is not the ROI (Mustra et al., 2010).

The pixels of the ROI area vary between research studies. Ferreira and Borges (2001) cropped the images into 64×64 pixels, while Prathiba and Sadasivam (2010) used an ROI of 32×32 pixels. Lowis et al. (2015) and Karahaliou et al. (2008) conducted their research by using an ROI of 128×128 pixels. In this paper, an ROI with 128×128 pixels is used because it covers most of the abnormality in each image without showing the background of the mammogram.



(a) (b) (c)

Figure 1 (a) a mammogram image from the MIAS database (1024×1024 pixels); (b) an example of a normal ROI (128×128 pixels); (c) an example of an abnormal ROI (128×128 pixels)

2.2. Feature Extraction

Different classes in classification can be distinguished using suitable features. As discussed in the "Introduction", wavelet analysis is used to extract the features of mammogram images. The wavelet used in the experiments of this work was Haar. Feature extraction involves two steps: image decomposition and feature extraction from the wavelet coefficient.

In the image decomposition process, the proposed method differs from that of the references. Ferreira and Borges (2001) used level 1 wavelet decomposition, while Pratibha and Sadasivam (2010) used level 3 decomposition and Hamad et al. (2013) performed wavelet decomposition up to level 4. This study applies the wavelet Haar up to level 6 to determine the optimal level of wavelet decomposition.

Decomposing an image using wavelet is carried out by applying the convolution of low-and high-pass filters on the images. The image can be decomposed into specific sets of coefficients in every level of decomposition. They are: low-frequency coefficients, which are the approximation of the original image; horizontal low-frequency coefficients, or the horizontal edge detail of the image; vertical high-frequency coefficients, or the vertical edge detail of the image; and diagonal high-frequency coefficients, or the diagonal edge detail of the image (Putra, 2010).

After the four sets of coefficients are obtained, relevant information for representing the original mammogram is extracted. The energy, mean, and standard deviation from each wavelet coefficient in every decomposition level are calculated to simplify the features and are used as the classification input.

2.3. Classification

In this step, the kNN is used to classify the extracted features. The classification is designed using Euclidean distance as a metric between the features of the testing data and the reference data as shown in Equation 1.

$$D_{Euclidean} = \sqrt{\sum(A(i,j) - M(i,j))^2} \quad (1)$$

Here, A is a vector of the testing data features, and M is a vector of the reference data in a class. The kNN method requires an integer k, a set of labeled references, and a metric of closeness. The selection of the k value becomes a problem to solve by trying various values of k and then selecting the k value that gives the best classification accuracy (Duda et al., 2000). The class of sample data will be determined by the majority of the class, which has the minimum distance within the k subset.

3. RESULTS AND DISCUSSION

The data set is divided into two groups; 189 normal images and 102 abnormal images as reference data, and 20 normal images and 11 abnormal images as testing data. In this work, we use the Haar wavelet function to decompose the image up to level six as seen in Figure 2. The wavelet energy feature, mean, and standard deviation are used as a reduction to simplify the wavelet coefficients. In the classification step, every level of decomposition along with a possible set of features from each textural feature extraction is analyzed to search the best performance in classifying data into normal and abnormal.

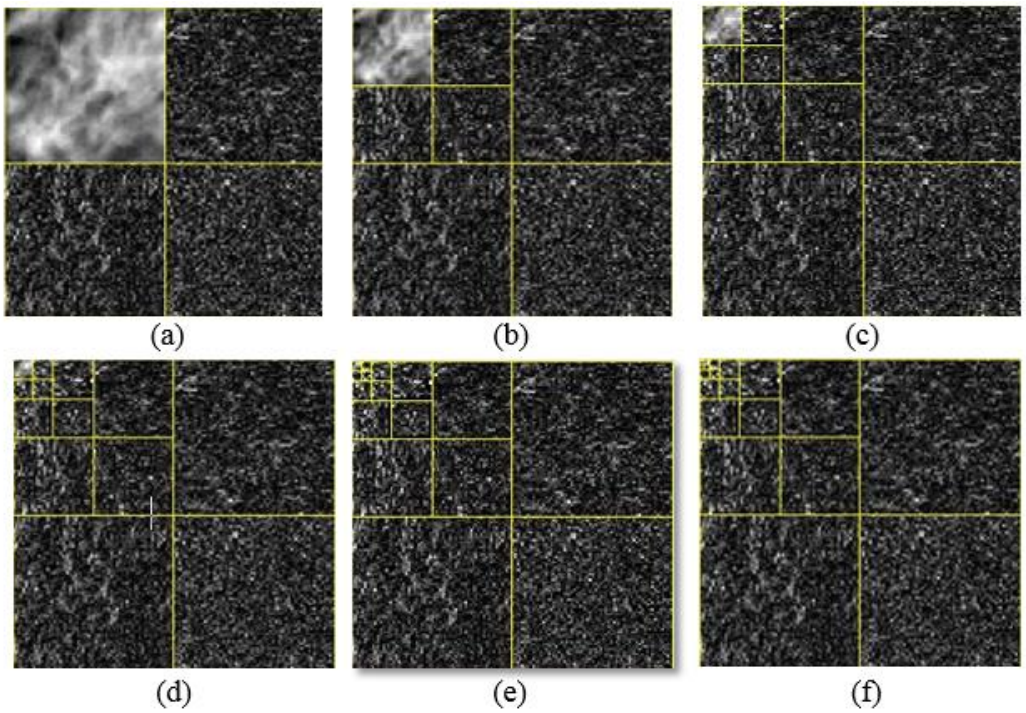


Figure 2 Wavelet decomposition in levels: (a) one; (b) two; (c) three; (d) four; (e) five; and (f) six

The feature sets for input classification are divided into seven categories: energy; mean; standard deviation; a combination of energy and standard deviation; a combination of energy and mean; a combination of mean and standard deviation; and a combination of energy, mean, and standard deviation in every decomposition level. Accuracy, sensitivity, and specificity are calculated to evaluate the performance of the classifier. Accuracy is the ability to differentiate

between normal and abnormal cases correctly, sensitivity is the ability to determine the abnormal cases correctly, and specificity is the ability to determine the normal cases correctly (Baratloo et al., 2015). The accuracy of every set of feature has been analyzed to obtain the best feature combination and the best decomposition level in classification performance.

In decomposition level 1, the feature combination of the mean and standard deviation along with the feature combination of energy, mean, and standard deviation from wavelet coefficients give better accuracy, sensitivity, and specificity than other features as input. The kNN is tried with various values, where for k = 1, the classifier performance is optimal. The optimum accuracy, sensitivity, and specificity are 87.1%, 63.6%, and 100%, respectively. At this level of decomposition, the sensitivity rate is not good enough to detect abnormality in the mammogram.

The classification result from the feature extraction in decomposition level 2 gives a better result than that of decomposition level 1. In this level, the best accuracy is also obtained from sets of features that are the same as those of level 1; a combination of mean and standard deviation and a combination of energy, mean, and standard deviation. The optimum k value of the kNN method in this level is k = 4. The optimum accuracy, sensitivity, and specificity obtained are 83.9%, 72.7%, and 90%, respectively.

As the classification input, the features of wavelet decomposition level 3 give better performance than do those of level 1 and level 2. The feature sets that give the best performance are also the same as those of the two levels before: mean-standard deviation and mean-energy-standard deviation. The optimum k value of this method is k = 2. The accuracy, sensitivity, and specificity of the input criteria mentioned are 96.8%, 100%, and 95%, respectively. This level gives the best performance for detecting abnormalities from mammogram images. Table 1 shows the accuracy rate in every k value and every possible input feature in decomposition level 3.

Table 1 Accuracy rate of classification for combination of selected features of wavelet decomposition level 3 (in percentage)

k Value of kNN Method	Energy	Mean	Standard Deviation	Energy-Standard Deviation	Energy-Mean	Mean-Standard Deviation	Energy-Mean-Standard Deviation
1	67.7	71.0	80.6	71.0	80.6	93.5	93.5
2	64.5	64.5	61.3	61.3	61.3	96.8	96.8
3	61.3	74.2	77.4	74.2	77.4	83.9	83.9
4	54.8	71.0	61.3	67.7	61.3	77.4	77.4
5	64.5	71.0	61.3	71.0	61.3	80.6	80.6
6	51.6	74.2	64.5	74.2	61.3	77.4	77.4
7	61.3	74.2	67.7	71.0	67.7	77.4	77.4
8	58.1	77.4	71.0	77.4	71.0	77.4	74.2
9	61.3	67.7	74.2	67.7	74.2	77.4	77.4

In level 4, the best classification result is obtained in the same set of features as that in level 3. The optimum k value in this level is k = 1. The optimum accuracy, sensitivity, and specificity obtained are 90.3%, 81.8%, and 95%, respectively. The accuracy and sensitivity performance in this level are better than that of level 1 and 2 but is not the best among all of the levels.

In level 5, the optimum accuracy is obtained using a combination of mean and standard deviation as the input feature with k = 5. The accuracy, sensitivity, and specificity obtained in

the feature mentioned are 83.9%, 54.5%, and 100%, respectively. At this level, the sensitivity to abnormalities in a mammogram is very low, probably because the image has been decomposed into a very small approximation image, so the breast density is not well represented. Thus, the classifier cannot distinguish between a normal and an abnormal image properly.

In level 6, the best classification result is obtained with a feature combination of energy and standard deviation, with k = 3. This result is different with level 5, before which the best feature is always shown in a feature combination of mean and standard deviation. This result is probably because high-level decomposition makes an approximation image become too small and not representative enough to distinguish breast tissue classes. The highest accuracy, sensitivity, and specificity obtained from this level are 83.9%, 72.7%, and 90%, respectively.

From the explanation above, the highest accuracy is obtained with mean-standard deviation and mean-energy-and-standard deviation as the input feature. Mean-standard deviation is considered the best feature set because it is simpler than the other for the sake of computational time. It is also shown that energy features from wavelet decomposition gives no significant result for classification performance.

The most important quality of this classification is the sensitivity rate because it shows the number of abnormal mammogram images that are classified correctly. The highest sensitivity rate is obtained in decomposition level 3, which is 100%, while the other levels show average performances in detecting mammogram abnormality. Abnormality detection in a mammogram image is important for taking further action for the patient. The obtained results indicate that the proposed algorithm for feature extraction and classification is a promising technique for helping to diagnose breast cancer.

4. CONCLUSION

In this paper, texture properties of ROI as well as the performance of the kNN classification technique are analyzed in classifying digital mammograms as normal or abnormal. The energy, mean and standard deviation from the wavelet decomposition coefficient are used as input for the classification. The best feature set, one that gives the greatest accuracy and sensitivity, is the feature combination of mean and standard deviation from wavelet decomposition level 3. The highest accuracy, specificity, and sensitivity obtained are 96.8%, 95%, and 100%, respectively.

5. REFERENCES

American Cancer Society, 2014. Detailed Guide: Breast Cancer. Available online at: http://www.cancer.org/cancer/breastcancer/detailedguide/, Accessed on 12 November 2014

Baratloo, A., Hosseini, M., Negida, A., Ashal, G.E., 2015. Simple Definition and Calculation of Accuracy, Sensitivity and Specificity. *Emergency*, Volume 3(2), pp. 48–49

Birdwell, R.L., Ikeda, D.M., O'Shaughnessy, K.F., Sickles, E.A., 2001. Mammographic Characteristics of 115 Missed Cancers Later Detected with Screening Mammography and the Potential Utility of Computer-aided Detection. *Radiology*, Volume 219(1), pp. 192–202

Duda, R.O., Hart, P.E., Stork, D.G., 2000. *Pattern Classification* (2nd edition), Wiley-Interscience, USA

Ferreira, C.B.R., Borges, D.L., 2001. Automated Mammogram Classification using a Multi-resolution Pattern Recognition Approach. *IEEE*, USA

Globocan, 2012. Fact Sheets Population Estimated Cancer Incidence Mortality and Prevalence Worldwide. Available online at: http://globocan.iarc.fr/Pages/fact_sheets_population.aspx, Accessed on 11 November 2014

Hamad, N.B., Taouil, K., Bouhlel, M.S., 2013. Mammographic Microcalcification Detection using Discrete Wavelet Transform. *International Journal of Computer Applications*, Volume 64(21), pp. 17–22

Isource National Breast Cancer Centre, 2011. *Clinical Practice Guidelines for the Management of Early Breast Cancer*, National Health & Medical Research Council, Australia

Karahaliou, A.N., Boniatis, I.S., Skiadopoulos, S.G., Sakellaropoulos, F.N., Likaki, E., Panayiotakis, G.S., Costaridou, L.I., 2006. A Texture Analysis Approach for Characterizing Microcalcifications on Mammograms. In: *the Proceeding of IEEE International Special Topic Conference on Information Technology Applications in Biomedicine 2006*, 26-28 October, Greece

Lowis, X., Hendra, Y., Lavinia, Z., 2015. The Use of Dual-tree Complex Wavelet Transform (DTCWT) Based Feature for Mammogram Classification. *International Journal of Signal Processing, Image Processing, and Pattern Recognition*, Volume 8(3), pp. 87–96

Michell, M., 2010. *Breast Cancer: Contemporary Issues in Cancer Imaging*, Cambrigde University Press, UK

Mustra, M., Grgic, M., Delac, K., 2010. Feature selection for automatic breast density classification. In: *the Proceedings of the 52nd International Symposium Elmar 2010*, 15-17 September, Croatia

Prathibha, B.N., Sadasivam, V., 2010. Multi-resolution Texture Analysis of Mammograms using Nearest Neighbor Classification Techniques. *International Journal of Information Acquisition*, Volume 7(2), pp. 109–118

Putra, D., 2010. *Pengolahan Citra Digital*, Penerbit Andi, Yogyakarta. (in Bahasa)

Radiologyinfo, 2014. Mammography, Available online at: www.radiologyinfo.org/en/info.cfm?pg=mammo, Accessed on 12 November 2014

Suckling, J., Parker, J., Dance, D.R., Astley, S., Hutt, I., Boggis, C.R.M., Ricketts, I., Stamatakis, E., Cerneaz, N., Kok, S-L, Taylor, P., Betal, D., Savage, J., 1994, *Digital Mammography*, pp. 375–378