COMMUNICATIONS IN SCIENCE AND TECHNOLOGY
Homepage: cst.kipmi.or.id

Texture feature extraction for the lung lesion density classification on computed tomography scan image

Hasnely[a*], Hanung Adi Nugroho[a], Sunu Wibirama[a], Budi Windarta[b], Lina Choridah[b]

[a]*Department of Electrical Engineering and Information Technology, Universitas Gadjah Mada, Jl. Grafika No.2, Yogyakarta 55281, Indonesia*
[b]*Department of Radiology, Universitas Gadjah Mada, Jl. Farmako Sekip Utara, Yogyakarta 55281, Indonesia*

Article history:
Received: 25 May 2016 / Received in revised form: 28 May 2016 / Accepted: 28 May 2016

Abstract

The radiology examination by computed tomography (CT) scan is an early detection of lung cancer to minimize the mortality rate. However, the assessment and diagnosis by an expert are subjective depending on the competence and experience of a radiologist. Hence, a digital image processing of CT scan is necessary as a tool to diagnose the lung cancer. This research proposes a morphological characteristics method for detecting lung cancer lesion density by using the histogram and GLCM (Gray Level Co-occurrence Matrices). The most well-known artificial neural network (ANN) architecture that is the multilayers perceptron (MLP), is used in classifying lung cancer lesion density of heterogeneous and homogeneous. Fifty CT scan images of lungs obtained from the Department of Radiology of RSUP Dr. Sardjito Hospital, Yogyakarta are used as the database. The results show that the proposed method achieved the accuracy of 98%, sensitivity of 96%, and specificity of 96%.

Keywords: Classification; Density; CT Scan Image; Lung cancer

1. Introduction

Cancer is the most common cause of death in the world as revealed from WHO (World Health Organization) statistic data. In 2012, it was recorded that cancer induced 8.2 million of deaths and lung cancer was recorded contributing of 1.59 million deaths that this disease is the highest mortality rate compared with liver cancer, stomach cancer, colorectal cancer, breast cancer and oesophageal cancer [1]. As reported by the WHO in 2014, in Indonesia, 30.866 deaths were caused by lung cancer, with 8.390 and 22.476 deaths of female and male respectively. The lung cancer then became the most common cause of death towards male in Indonesia [2]. Lung cancer is an abnormal growth of the lung cells in body tissues and grow to be cancer cells [2] in one or both parts of them, which commonly caused by smoking [3]. It is divided into Small Cell Lung Cancer (SCLC) and Non-Small Cell Lung Cancer (NSCLC) [3].

The radiology examination of CT scan is one of early detection methods of the lung cancer to make the initial phase diagnosis in order to minimize the mortality rate. The image interpretation and assessment of lesion characteristic is subjective and various depend on the radiologist experience. Hence, the digital image processing of CT scan lungs image is necessary with an expectation that it can provide a second opinion diagnosis process. Based on morphological description on the CT Scan image there are number of criteria of the primary lung cancer diagnosis including ground glass opacity, irregular speculated margin, density, size of tumour, air bronchogram, lobulated, and enhancement [4]. This research is focused on the morphological characteristics detection of lung cancer lesion density. The lesion density is a description of the tissue density which can be further divided into heterogeneous density and homogeneous. A digital image processing is required for morphological characteristics of lung cancer lesion density detection.

A number of researches has been conducted for feature extraction of texture, such as a research which carried out by Uyun [5] about pattern density detection on the mammogram image by using feature extraction method of GLCM (Gray Level Co-occurrence Matrices). The results obtained a strong significance towards the determination of breast cancer. Furthermore, Devan, et al [6] in their research used the extraction method of texture feature of GLCM, GLRLM (Gray Level Run Length Matrices) and entropy to identify the characteristics of three lung tissues types including normal, fibrosis, and carcinoma. The research result showed that the features used can differentiate three types of lung tissues properly. The texture feature extraction by using the histogram and GLCM was also conducted by Patil, et al [7] to identify the benign cancer and malignant cancer by using X-ray image. Tun, et al [8] conducted Otsu segmentation and texture feature extraction to identify the lung cancer stages of stadium I, II, III, IV by using GLCM method. A number of other researchers have used the method of the classifications of MLP (Multilayer Perceptron). One of example is a research which conducted by Anand [9] in classifying the lung tumour as cancer and normal. The research conducted by Ahmad, et

* Corresponding author. Tel.: +62-274-552305; fax: +62-274-552305.
Email: hasnely.mti13@mail.ugm.ac.id.

al [10] and Mitrea, et al [11] also implemented MLP to classify the image of colorectal cancer while research conducted by Valarmathi, et al [12] was to classify the mammogram image.

In this research, the identification of the morphological characteristics of lung cancer lesion density is conducted by using the texture feature extraction method and classification of lesion density. The image processing phases are the pre-processing by cropped RoI (Region of Interest), segmentation process by using Otsu segmentation, morphological operation and the feature extraction of the texture based upon the histogram and GLCM. The result of the texture feature extraction would be used for the classification phase using the method of MLP.

2. Materials and Methods

This research is uses the data from the Department of Radiology of RSUP Dr.Sardjito Hospital, Yogyakarta, consist of 50 CT scan images of lung cancer. The aim of this research is to identify the morphological characteristics of lesion density from CT scan image in the case of primary lung cancer. Fig. 1 illustrates the block diagram of conducted research which include the pre-processing, segmentation, morphological operation, feature extraction and classification.

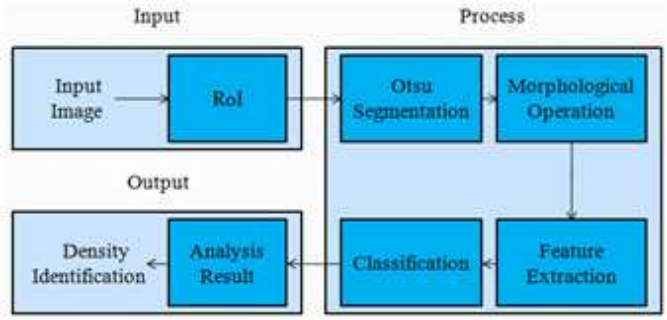


Fig. 1. Block diagram of the identification of lesion density morphology

2.1. Pre-processing

At the pre-processing stage, the image is cropped on the RoI as initial step to make the research focus on the lesion part. This process is conducted manually to facilitate the process of identifying the morphology of lesion density. The cropping result of RoI is shown in Fig. 2.

2.2. Otsu Segmentation

Otsu segmentation is a process of classification of pixel to differentiate two parts: object and background [13][14] by calculating the threshold values automatically based on the input image [15]. At first, the main principle of Otsu is to determine the probability of intensity value of i in the histogram which is calculated by (1). N states is the total number of all pixels in the image and n_i states is the number of pixels with the intensity i.

$$p(i) = \frac{n_i}{N}, p(i) \geq 0, \sum_1^{256} p(i) = 1 \quad (1)$$

The weighting in two classes: object and background are calculated with the equation (2) and (3) in which L states is the number of grey level.

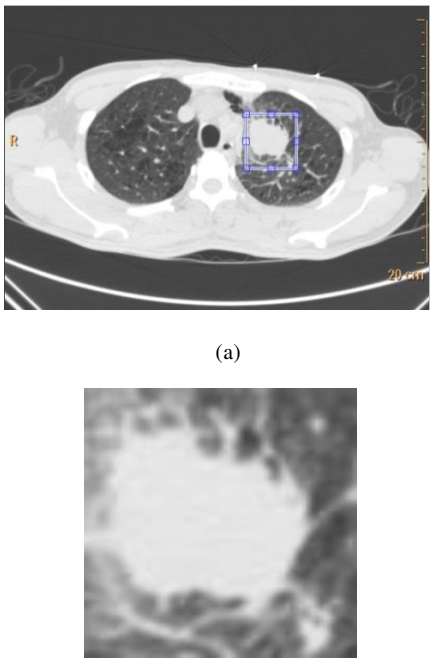
$$w_1(t) = \sum_{i=1}^{t} p(i) \quad (2)$$

$$w_2(t) = \sum_{i=t+1}^{L} p(i) = 1 - w_1(t) \quad (3)$$

The mean of object and background was calculated with the Equation (4) and (5).

$$m_1(t) = \sum_{i=1}^{t} i \,.\, p(i)/w_1(t) \quad (4)$$

$$m_2(t) = \sum_{i=1}^{t} i \,.\, p(i)/w_2(t) \quad (5)$$



(a)

(b)

Fig. 2. RoI cropping process of image: (a) Original image; (b) Image as the result of RoI cropping

Equation (6) states σ_B^2 called as between-class variance (BVC). The total means are calculated with equation by using equation (7). The optimum threshold value was obtained by maximizing BVC and more less computation time [13].

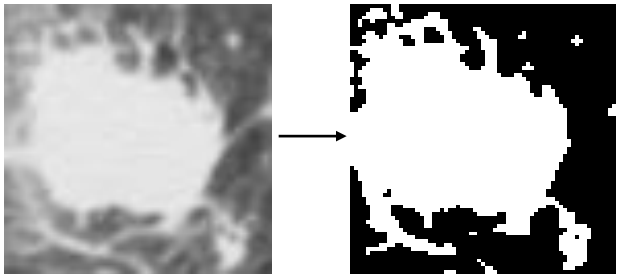
$$\sigma_B^2(t) = w_1.[m_1(t) - m_T]^2 + w_2.[m_2(t) - m_T]^2 \quad (6)$$

$$m_T = \sum_{i=1}^{N} i \,.\, p(i) \quad (7)$$

Fig. 3 shows the segmentation process by using Otsu method. The result of the Otsu segmentation showed the fixed form of lesion object separated from the background.

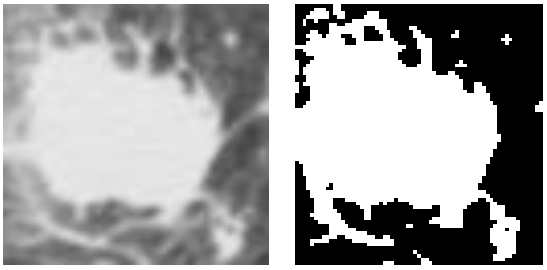
2.3. Morphological Operation

The output of Otsu segmentation is binary image that will be used as a template to get the lesion area by applying a simple morphological operation such as AND, OR and NOT. AND operation is used in this research. Fig. 4 shows the process of morphological operation.

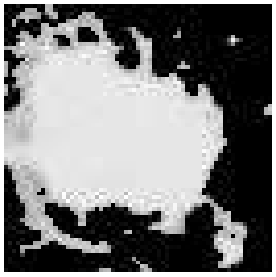


(a) (b)

Fig. 3. Process of Otsu segmentation: (a) Image of RoI; (b) Segmentation image



(a) (b)



(c)

Fig. 4. Process of morphological operation: (a) RoI image; (b) Segmentation image; (c) Morphological operation image

2.4. Feature Extraction

The texture extraction method consisted of three groups: statistic method, structural method and spectral method. In this research, the statistic method is used of the first order based upon the histogram and the second order with the base of GLCM as it can identify the density of constituent tissues by using the intensity of grey level with the highest performance in a number of previous researches.

2.4.1 Histogram-Based Texture

The simplest extraction method of the statistic properties for the texture is the order one which based on the histogram. In order to obtain the histogram-based statistic properties, the texture of an image can be calculated using the following features (8) - (13) [13].

1) Mean

$$m = \sum_{i=0}^{L-1} i \,.\, p(i) \tag{8}$$

The formula resulted mean of object brightness. In this case, *m* refers to number of mean value, *i* refers to the grey level in the image and *p(i)* represents the probability of emergence of *i* and *L* presenting the highest grey level.

2) Standard Deviation

$$\sigma = \sqrt{\sum_{i=1}^{L-1}(i-m)^2\, p(i)} \tag{9}$$

Standard deviation (σ) refers to the level of statistic spread measuring the pattern of data spread and provides the level of contrast.

3) Skewness

$$Skewness = \sum_{i=1}^{L-1}(i-m)^3\, p(i) \tag{10}$$

Skewness refers to the level of asymmetrical towards the mean. It will be negative (-) if the histogram curve tends to be on the left side of from the value means and it will be positive (+) if otherwise.

4) Energy

$$Energy = \sum_{i=0}^{L-1}[p(i)]^2 \tag{11}$$

Energy refers to the stating level of pixel intensity distribution towards the extent of grey level which commonly known as uniformity.

5) Entropy

$$Entropy = -\sum_{i=0}^{L-1} p(i)\log_2(p\,(i)) \tag{12}$$

Entropies present a level complexity of an image. The higher of the value represent the high complexity of the image. Entropy also indicates the quantity of information contained in the data spread.

6) Smoothness

$$Smoothness = 1 - \frac{1}{1+\sigma^2} \tag{13}$$

The level of smoothness of an image could be measured by the smoothness value. The low smoothness value shows that the image has the rough intensity.

2.4.2 Gray Level Co-occurrence Matrices (GLCM)

The GLCM method was firstly published by Haralick in 1973 with 28 values of features. GLCM uses the texture measurement in second order by considering the relationship between the pair of two pixels of original image [13].

In example, *f(x, y)* refers to the image with the size of *Nx* and *Ny* that has a pixel with probability. Thus, the *L* level and

$\vec{r}$ are the spatial direction vectors. $GLCM\vec{r}(i,j)$ is defined as the number of pixels with $j \in 1,..,L$ occurred in the offset $\vec{r}$ towards the pixel with the values of $i \in 1,..,L$that can be stated in the formula [16] :

$$GLCM_{\vec{r}}(i,j) = \#\{(x_1,y_1),(x_2,y_2) \in (N_x,N_y) \times (N_x,N_y) | f(x_1,y_1) = j^{\vec{r}} = (\overrightarrow{x_2-x_1,y_2-y_1})\} \quad (14)$$

In this case, offset $\vec{r}$ refers to angle and distance of pixel. For example, Fig. 5 shows four directions for GLCM.

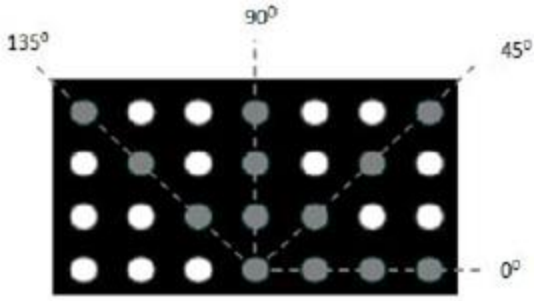


Fig. 5. The direction of GLCM (0°, 45°, 90°, and 135°)

Five features of GLCM used include Angular Second Moment (ASM), contrast, Inverse Different Moment (IDM), entropy, and correlation as explained as follows (15) – (22) [17]:

1) ASM

$$ASM = \sum_{i=1}^{L}\sum_{j=1}^{L} GLCM\ (i,j)^2 \quad (15)$$

Angular Second Moment (ASM) measures the uniformity or the size of the properties of homogeneity of image.

2) Contrast

$$Contrast = \sum_{n=1}^{L} n^2\{\sum_{|i-j|=n} GLCM\ (i,j)\} \quad (16)$$

Contrast measures the spatial frequency of image and the size of spread (inertia moment) of image matrix element. It also refers to the existence of the variation of grey level of image pixel.

3) IDM

$$IDM = \sum_{i=1}^{L}\sum_{j=1}^{L} \frac{GLCM\ (i,j)^2}{1+(i-j)^2} \quad (17)$$

Inverse Different Moment (IDM) is used to measure the homogeneity of image with the similar grey level. The homogeny image will have a large value of IDM.

4) Entropy

$$Entropy = -\sum_{i=1}^{L}\sum_{j=1}^{L}(GLCM(i,j))\log(GLCM\ (i,j)) \quad (18)$$

Entropy is the size of complexity of grey level of an image. The values will be low if the elements of GLCM are close to the value of 0 or 1 and the value will be high if the elements of GLCM have the relatively equal value.

5) Correlation

$$Correlation = \sum_{i=1}^{L}\sum_{j=1}^{L} \frac{(ij)(GLCM\ (i,j)) - \mu_i'\mu_j'}{\sigma_i'\sigma_j'} \quad (19)$$

Correlation is the size of the correlation of linearity from a number of pixel pair and provides the information regarding the linear structure in image.

Remark:

$$\mu_i' = \sum_{i=1}^{L}\sum_{j=1}^{L} i * GLCM(i,j) \quad (20)$$

$$\mu_j' = \sum_{i=1}^{L}\sum_{j=1}^{L} i * GLCM(i,j) \quad (21)$$

$$\sigma_j^2 = \sum_{i=1}^{L}\sum_{j=1}^{L} GLCM(i,j)\ (i-\mu_i')^2 \quad (22)$$

$$\sigma_i^2 = \sum_{i=1}^{L}\sum_{j=1}^{L} GLCM(i,j)\ (i-\mu_i')^2 \quad (23)$$

2.5. Classification

The artificial nerve tissue can be used for classification and to identify the pattern of object [18]. MLP is the formation of artificial nerve system that are mostly used in terms of education and application [19]. MLP has abilities to learn and give the better performance of classification are proven in a number of research [10] [20]. At the classification stage, this research used MLP method by using three layers, consisting of input layer, hidden layer, and output layer. The classification was conducted using Weka machine learning [21]. K-fold cross validation is chosen to evaluate the performance of training and testing feature from the dataset before being classified [22]. Technically, the architecture of MLP used in this research is illustrated in Fig. 6.

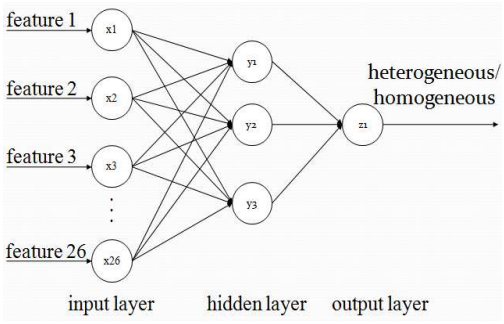


Fig. 6. Architecture of MLP for the classification of lesion density

The performance of classification is measured from the prediction of accuracy, sensitivity and specification aspects as expressed in (24) - (26). Where, TP, TN, FP and FN are true positive, true negative, false positive, and false negative, respectively.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (24)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (25)$$

$$Specificity = \frac{TN}{TN+FP} \quad (26)$$

3. Results and Discussion

Based on the experiment in this research, the texture feature extraction based on histogram and GLCM is taken from 50 images and the average value of feature shown in Table 1 for each features to difference heterogeneous and homogeneous lesion.

Table 1. Average result of the histogram and GLCM-based texture feature extraction for each features

No	Statistic feature of histogram and GLCM	Average value of heterogeneous	Average value of homogenous
1	Mean	92,50337	152,49886
2	Standard Deviation	99,42738	99,76297
3	Skewness	3,01751	-11,26714
4	Energy	0,34104	0,14217
5	Entropy	2,36069	2,7899
6	Smoothness	0,13258	0,13326
7	ASM 0^0	0,28815	0,07602
8	ASM 45^0	0,27486	0,06830
9	ASM 90^0	0,28558	0,07292
10	ASM 135^0	0,27288	0,06902
11	Contrast 0^0	1708,03977	1463,88758
12	Contrast 45^0	2617,44874	2336,75677
13	Contrast 90^0	1765,57904	1619,49413
14	Contrast 135^0	2478,66052	2251,99059
15	IDM 0^0	0,63694	0,53278
16	IDM 45^0	0,59110	0,44343
17	IDM 90^0	0,61557	0,47510
18	IDM 135^0	0,59197	0,45022
19	Entropy 0^0	3,53509	4,34803
20	Entropy 45^0	3,68478	4,59926
21	Entropy 90^0	3,59275	4,49295
22	Entropy 135^0	3,69508	4,58720
23	Correlation 0^0	0,00009	0,0001
24	Correlation 45^0	0,00009	0,0001
25	Correlation 90^0	0,00009	0,0001
26	Correlation 135^0	0,00009	0,0001

From the conducted experiment result, the ranked features difference influential are contrast, skewness, mean, entropy, ASM, energy, IDM, standard deviation, smoothness, and correlation. The value of all texture feature extraction method based on the histogram and GLCM required in the classification process. There are 50 images input of data which contains 25 heterogeneous image and 25 homogeneous image. The total combined features are 26 features of each image. Table 2 shows the accuracy, sensitivity, and specificity classifying rate.

Table 2. Result of the classification of 26 texture feature

Texture feature extraction	Classifier	Accuracy	Sensitivity	Specificity
Histogram	MLP	80%	88%	72%
GLCM	MLP	96%	96%	96%
Histogram + GLCM	MLP	98%	96%	96%

Classification based on MLP is used in this research could facilitate the process of classification of heterogeneous and homogeneous lesion with the highest accuracy. Fig. 7 shows the confusion matrix of proposed method describes that TP = 24; the number of image with characteristics heterogeneous lesion recognizable as heterogeneous from 25 cases, TN = 25; all number of image with characteristics homogeneous lesion recognizable as homogeneous, FN = 1; only one image with characteristics heterogeneous lesion density recognizable as homogeneous lesion, and FP = 0; there is no image with characteristics homogeneous lesion recognizable as heterogeneous lesion.

```
=== Confusion Matrix ===

  a  b   <-- classified as
 24  1 |  a = Heterogeneous
  0 25 |  b = Homogeneous
```

Fig. 7. The confusion matrix of proposed method

4. Conclusion

This research proposes a method to identify the characteristics and classification of lesion density of primary lung cancer by using the histogram and GLCM-based texture feature extraction. The combination of histogram and GLCM-based texture feature extraction obtained achieved the accuracy of 98%, sensitivity of 96%, and specification of 96%. The obtained results are able to show quantitatively that the two methods are able to identify the characteristics of the difference of lesion density between the heterogeneous and homogeneous lung cancer. Thus, it can help the radiologists in interpreting the image. Furthermore, the proposed method in this research can be recommended as one part of CAD development to diagnose the lung cancer. In future work, it is suggested to propose other segmentation technique and feature extraction for density lesion detection.

Acknowledgements

The authors thank Department of Radiology of RSUP Dr. Sardjito Hospital, Yogyakarta that has provided the database for this research.

References

1 W. H. Organization, *Cancer Media Centre*, 2015. [Online]. Available: http://who.int/mediacentre/factsheets/fs297/en/. [Accessed: 09-Apr-2016].

2 Kementrian Kesehatan RI, *Stop Kanker*, Indonesia: Pusat Data and Informasi Kesehatan, 2015.

3 N. C. Institute, *Lung Cancer—Patient Version*. [Online]. Available: http://www.cancer.gov/types/lung. [Accessed: 13-Apr-2016].

4 A. Icksan, R.M . Faisal, Elisna, P. Astowo, H. Hidayat and J. Prihartono, *Kriteria Diagnosis Kanker Paru Primer berdasarkan Gambaran Morfologi pada CT Scan Toraks Dibandingkan dengan*

Sitologi, Indonesian Journal of Cancer (2008).

5 S. Uyun, *In term of : Model Komputasi penentuan faktor Resiko Kanker Payudara berdasarkan Pola dan Persentase Densitas Mamografi*, Ph.D. Thesis, Universitas Gadjah Mada, Indonesia, 2014.

6 L. Devan, R. Santosham and R. Hariharan, *ANOVA of Texture based Feature Set for Lung Tissue Characterization using Low-Dose CT Images*, 7(1) (2014) 974-1925.

7 S. A. Patil and M. B. Kuchanur, *Lung Cancer Classification Using Image Processing*, 2(3) (2012) 2277-3754.

8 K. M. M. Tun and A. S. Khaing, *Feature Extraction and Classification of Lung Cancer Nodule using Image Processing Techniques*, 3(3) (2014) 2278–0181.

9 S. K. V. Anand, *Segmentation coupled Textural Feature Classification for Lung Tumor Prediction*, 10th Int. Conference Communication and Computing Technologies., Nagercoil, Tamil Nadu, India, 2010, pp. 518–524.

10 M. Y. Ahmad, A. Mohamed, Y. A. M. Yusof, and S. S. M. Ali., *Colorectal Cancer Image Classification Using Image Pre-Processing and Multilayer Perceptron*, Int. Conference on Computer and Information Science., Kuala Lumpur, Malaysia, 2012, pp. 275–280.

11 D. Mitrea, S. Nedevschi, M. Abrudean, and R. Badea, *Colorectal Cancer Recognition from Ultrasound Images , Using Complex Textural Microstructure Cooccurrence Matrices , Based on Laws ' Features*, Int. Conference Telecomunnications Signal Processing., Prague, Czech Republic, 2015, pp. 458–462.

12 P. Valarmathi and S. Robinson, *Efficacy of Feature Selection Techniques for Multilayer Perceptron Neural Network to Classify Mammogram*, 6th Int. Conference on Advanced Computing., Chennai, India, 2014, pp. 26–31.

13 A. Kadir and A. Susanto, *Pengolahan Citra Teori dan Aplikasi*, 1st ed. Yogyakarta, Indonesia: ANDI, 2012.

14 F. Y. Shih, *Image processing and pattern recognition: fundamentals and techniques*. Johm Wiley & Sons, 2010.

15 D. Putra, *Pengolahan Citra Digita*l, 1st ed. Yogyakarta: ANDI, 2010.

16 S. D. Newsam and C. Kamath, *Comparing Shape and Texture Features for Pattern Recognition in Simulation Data*, IS&T/SPIE's Annual Symposium on Electronic Imaging., San Jose, CA, United States, 2005, pp. 106–117.

17 R. M. Haralick, K. Shanmugam, and I. Dinstein, *Textural Features for Image Classification*, IEEE Transactions on Systems, Man, and Cybernetics., 3(6) (1973) 610–621.

18 R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. Wiley, 2001.

19 Suyanto, *Artificial Intelligence*. Bandung, Indonesia: Informatika, 2007.

20 I. S. Isa, Z. Saad, S. Omar, M. K. Osman, K. A. Ahmad, and H. A. M. Sakim, *Suitable MLP network activation functions for breast cancer and thyroid disease detection*, 2nd International Conference on Computational Intelligence, Modelling and Simulation, CIMSim, Bali, Indonesia, 2010, pp. 39–44.

21 M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, *The WEKA data mining software*, SIGKDD Explor., vol. 11, no. 1, 2009, p. 10.

22 P. Refaeilzadeh, L. Tang and H. Liu, *Cross-Validation*, Encycl. Database Syst. (2009) 532–538.