Feature Extraction for Lesion Margin Characteristic Classification from CT Scan Lungs Image

Yosefina Finsensia Riti[1], Hanung Adi Nugroho[2], Sunu Wibirama[3]
Department of Electronical Engineering and Information Technology Engineering Faculty, Universitas Gadjah Mada
Yogyakarta,Indonesia
[1]finsensia.mti13@mail.ugm.ac.id, [2]adinugroho@ugm.ac.id, [3]sunu@ugm.ac.id

Budi Windarta[4], Lina Choridah[5]
Department of Radiology, Medical Faculty
Universitas Gadjah Mada
Yogyakarta, Indonesia
[4]budiwinarta@ugm.ac.id, [5]linachoridah@ugm.ac.id

Abstract— **Lung cancer is one of the common cancer which occurred in both male and female. Revealed by WHO data, in 2012, this disease become one of the major cause of death in worldwide with the mortality rate about 1.59 million. An early detection of lung cancer by using Computed Tomography (CT) Scan can provide more opportunity to survive. However, the diagnosis of lung cancer by reading the CT scan image which performed by radiologists may lead to an error. A computer-based digital image processing is a solution to improve the accuracy and consistency in reading the CT Scan image result. This study aim is to identify the morphological characteristic of regular and irregular margins by using feature extraction method. In this research, image processing divided into several stages refer to the segmentation process with Otsu method, feature extraction with number of features such as convexity, solidity, circularity, and compactness, and the last is classification by using Multi Layer Perceptron (MLP). The classification process of features convexity, solidity, circularity, and compactness, resulted in the accuracy value of 85%, sensitivity of 85%, and specificity of 85%.**

Keywords—Lung Cancer; CT Scan; Lesion Margin, Feature Extraction and Classification

I. Introduction

Lung cancer is a cancer which commonly occurred in any gender. In 2012, *World Health Organization* (WHO) statistic data showed that lung cancer was the first cause of death in worldwide compare with other five types of cancers such as liver cancer, gastric cancer, colorectal cancer, breast cancer and esophagus cancer. The number of deaths recorded caused by lung cancer was at approximately 1.59 million in 2012 [1]. In the year of 2014, lung cancer became the major cause of the deaths for men in Indonesia, while for women are breast cancer and cervical cancer which it claimed infected 22,479 men and 8,390 women [2]. The main cause of lung cancer is the addiction of smoking cigarettes, carcinogenic environment such as radioactive gas and also air pollution. In addition, genetic factors also have a contribution to cause lung cancer [3] [4].

Lung cancer is basically a tumor which growing rapidly and spreads to the other organs. The occurrence of cancer is characterized by an abnormal growth cell that could damage the cells of normal tissue [5]. Based on its histopatology, lung cancer is classified into two types: Small Cell Lung Cancer (SCLC), a cancer of the lung with a small cell, and Non-Small Cell Lung Cancer (NSCLC), lung cancer with small cells consisting of Squamous Cell Cancer (ACC), adenocarcinoma (ADC), and large cells [4] [5]. From these type of cancer, NSCLC caused 80-90% of deaths in the world [4] [5]. From these type of cancer, NSCLC caused 80-90% of deaths in the world [3] [6] [7].

The lung cancer detection can be done by taking a screening using Computed Tomography (CT) Scan. The CT Scan result then observed on morphological guide of lung cancer as the diagnostic criteria such as the size of the tumor, enhancement, irregular spiculated margin [8] [9], lobulated, water bronchograms, ground glass opacity, and heterogeneous density [10]. The lung cancer diagnosis by using CT Scan image which conducted by a radiologist may lead to an error influenced by the blurring of anatomical structures surrounding the lung area, the small size of lesions, and also the different experiences of the radiologist generate a different interpretation [11]. To avoid the errors and to improve the accuracy and consistency, a computer-based digital image processing is necessary as the second opinion to read the CT Scan image result.

This study aims to identify the lung cancer morphological characteristics of regular and irregular margin which can be used by radiologists as one of the parameters to determine the malignancy degree of the lesion.

In this study, the initial stage of image processing is pre-processing to crop Region of Interest (RoI), segmentation, feature extraction and classification.

Several number of segmentation process on CT Scan lungs image have been applied by using watershed method [12], fuzzy possibilistic C-Means (FPCM), Fuzzy C Means [13], region growing [14], or Otsu threshodilng [15] [16]. In this study the segmentation process used Otsu thresholding method because it is able to select the most optimal threshold automatically and stably because it based on the image histogram [15]. After the image of the segmentation result was obtained, feature extraction is performed to obtain the characteristics value which used in the classification process. Extraction features are based on the geometric feature to identify the characteristic margin such as circularity [17] [18], area, perimeter [12] [13] [16] [18] [19], irregularity index (compactness) [13] [19] [20], equivalent diameter, convex area, solidity [19] [21], convexity [19], and eccentricity [12] [16] [21]. Furthermore, the features extraction is used to identify the margin morphological characteristics on the CT scans image including convexity, solidity, circularity, and compactness. The last stage is classification by using Multi Layer Perceptron (MLP) to distinguish classes between regular and irregular margins.

II. RESEARCH METHOD

Fifty four images of cancerous lesions are taken from DR.Sardjito Hospital Yogyakarta as image dataset are used in this study. The dataset was stored in RGB and bitmap (.bmp) types, and grouped into regular margin, consist of 27 images, and the irregular margin consist of 27 images). The steps of image processing of CT lung scans describe in Figure 1.

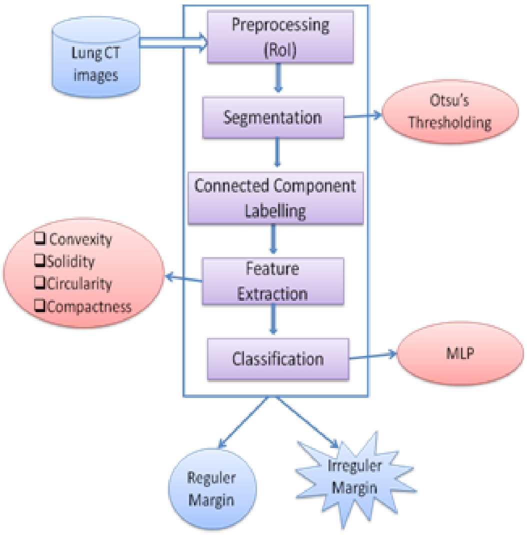


Figure1. Step of Image Processing of CT lung Scan

A. *Pre-Processing*

The initial stage of the whole process is pre-processing which gain the RoI area from the CT Scan image as shows in Figure 2.

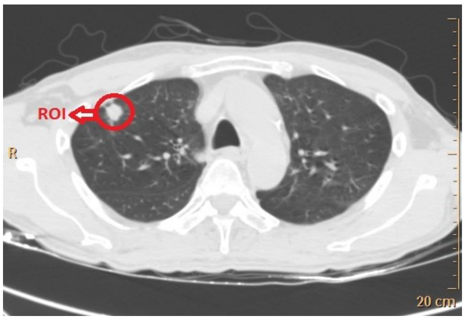


Figure 2. Original Lung Image of CT Scan and RoI

B. *Segmentation*

Segmentation process aims to divide the image into several areas, and separated between the object and background areas. RoI image which obtained in the previous stage then segmented by using thresholding Otsu.

The purpose of Otsu method is to divide the histogram of gray level image into two distinct areas automatically without requiring to input the threshold value. The approach taken by the Otsu method to perform a discriminant analysis is to determine the variables that can distinguish between two or more groups. The discriminant analysis could maximize these variables to separate the object and the background. The Otsu segmentation result are in the binary image form which only have the intensity value of 0 and 1. The value of 0 indicated the intensity of black (background), while the value of 1 indicated the intensity of the white (the object). Otsu thresholding method is computation the probalitiy of intensity value *i* in histogram [15] [16], then normalized and distributed.

$$\rho_i = \frac{n_i}{N}, \rho_i \geq 0, \sum_{i=1}^{L} \rho_i = 1 \qquad (1)$$

Where L: gray levels, n: number of pixels, and N: total pixels. After that, it is divided the histogram into classes and class background object ($\omega_0$dan ω_1), and calculate the mean of these two classes.

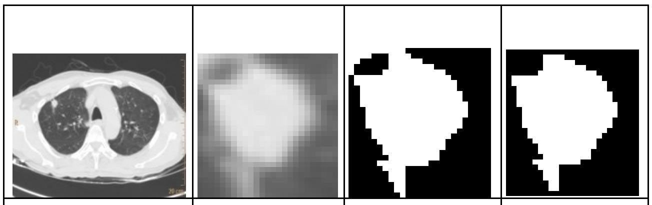
$$\omega_0 = \sum_{i=1}^{k} \rho_i = \omega(k) \text{ and } \omega_1 = \sum_{i=k+1}^{k} \rho_i = 1 - \omega(k) \qquad (2)$$

$$\mu_0 = \sum_{i=1}^{k} i\rho_i / \omega_0 \text{ and } \mu_1 = \sum_{i=k+1}^{L} i\rho_i / \omega_1 \qquad (3)$$

The threshold values obtained by calculate the optimal threshold between class variance (BCV).

$$\sigma^2 (k) = \omega_0 \omega_1 (\mu_0 - \mu_1)^2 \qquad (4)$$

Otsu segmentation result obtained the lesions form, hence there are still some false areas around the lesion which also detected it removed by using connected component labeling. The results of Otsu segmentation and connected component labeling (CCL) as shows in Figure 3.

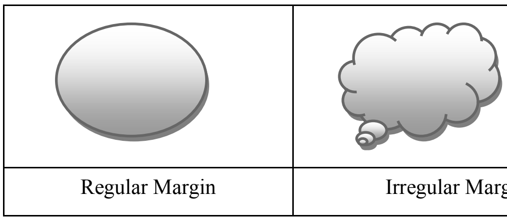


Original Image | RoI | Otsu Thresholding | CCL Result

Figure 3. Phase of Segmentation Processing

C. Feature Extraction

Feature extraction is a further step to obtain lung area characteristic after segmentation process. All the calculated feature from the image carries some informations about lung nodules. These contain important informations to detect the malignant or non-malignant lung nodules [13]. In this study, the features used to identify the margins of lesions characteristics included convexity, solidity, circularity, and compactness. Figure 4 shows the example of a model that represents the image of the lesion of regular and irregular margin.



Regular Margin | Irregular Margin

Figure 4.The Representation of The Margin of CT Scan Lung Image

a. Convexity

Convexity is a measure of comparison convex image to its hull convex. The values of convexity are obtained from the perimeter length ratio (circum ference/margin) between hull convex surrounding the object with the perimeter object length. For irregular margin, the convexity values are closer to zero [19]. Mathematically, convexity was formulated in equation (5).

$$convexity = \frac{Convex\ perimeter}{ObjectPerimeter} \tag{5}$$

b. Solidity

Solidity is a size of comparison of object area compare to its hull convex by utilizing the pixels which construct the hull convex. For irregular margin, the solidity value is close to zero [19]. Mathematically, solidity is formulated in equation (6).

$$solidity = \frac{ObjectArea}{ConvexArea} \tag{6}$$

c. Circularity

Circularity is a measure of average ratio of euclidean distance from the center to the margin of the object and the standard deviation of the distance from the center point to the margin of the area. The circularity value of irregular value is less than 1 and the circularity value of regular margin is more than 1 [17]. Mathematically, circularity is defined in equation (7) with N refers to the number of pixels and (y,x) refers to the pixel coordinates.

$$circularity = \frac{\frac{1}{N}\sum_{i=1}^{N}|(y_i,x_i)-(\bar{y}_c,\bar{x}_c)|}{\frac{1}{N}\left[|(y_i,x_i)-(\bar{y}_c,\bar{x}_c)|-\left(\frac{1}{N}\sum_{i=1}^{N}|(y_i,x_i)-(\bar{y}_c,\bar{x}_c)|\right)\right]^2} \tag{7}$$

d. Compactness

Compactnes or irregularity index is a form of roundness from the object margin. Irregularity index can be obtained by measuring the ratio between the object and the square of the object perimeter. The compactness value of irregular margin is less than 1 [13]. Mathematically, compactness is formulated on equation (8).

$$I = 4\pi\, x \frac{Area}{(perimeter)^2} \tag{8}$$

D. Classification

Classification is the final step in image processing to obtain the required information. In this study classification is used to facilitate the detection of lesion morphological characteristics of on lungs CT scan image related with the existing classes, including the class with the regular and irregular margins. In this research Multi Layer Perception (MLP), one model of Artificial Neural Network (ANN) is used as the classification method, which consist of several layers of perceptron: input layer, hidden layer and output layer, and in each layer there are several number of neurons [22]. MLP can produce a proper classification with the direct learning process, and the determination of the optimal weights by using back propagation training process [21] [23]. The classification process by using the MLP describes in Figure 5.

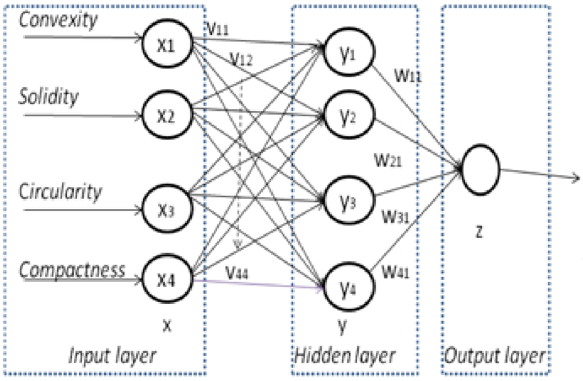


Figure 5.Architecture of MLP for the classification of the Margin of CT Scan Lung Image

Figure 5, shows that x1,...x4 are the quantitative values of the extracted features, y1,... y4 are the output from the input layer, z is the output of the hidden layer, vxy and wyz are the weight value changing by ongoing training process.

The classification results then measured by using confusion matrix, consisting of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) [22]. The confusion matrix can be calculated for the level of accuracy, sensitivity, and specificity, which can be defined by following equation.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{9}$$

Accuracy is a ratio between the number of correct identified data by the total of all data.

$$Sensitivity = \frac{TP}{TP+FN} \tag{10}$$

Sensitivity is measured the classifier ability to identify the data considered to be correct according to the TPR (*True Positive Rate).*

$$Spesificity = \frac{TN}{TN+FP} \tag{11}$$

Specificity is a measure of the classifier ability to identify the data considered wrong according to TNR. *(True Negative Rate*).

III. RESULTS AND DISCUSSION

In this study, the data consist of 54 images grouped in lesions with a regular margin (27 images), and the irregular margin (27 images). The extraction results from each feature can be seen in Table 1.

Table 1. Sample of The Results of Feature Extraction of CT Scan Lung Image

No.	Convexity	Solidity	Circularity	Compactness	Margin
1	0,96	0,99	0,92	0,81	Regular
2	0,89	0,92	0,7	0,67	Regular
3	0,85	0,81	0,45	0,45	Regular
4	0,97	0,97	0,66	0,62	Regular
5	0,89	0,86	0,67	0,62	Regular
6	0,57	0,69	0,21	0,19	Regular
7	0,94	0,98	0,81	0,77	Regular
8	0,93	0,96	0,74	0,7	Regular
9	0,79	0,82	0,44	0,41	Regular
10	0,93	0,92	0,81	0,74	Regular
11	0,66	0,76	0,27	0,24	Irregular
12	0,81	0,8	0,58	0,46	Irregular
13	0,83	0,82	0,54	0,51	Irregular
14	0,97	1,01	0,97	0,83	Irregular
15	0,75	0,79	0,38	0,35	Irregular
16	0,74	0,8	0,42	0,38	Irregular
17	0,67	0,8	0,28	0,27	Irregular
18	0,9	0,93	0,71	0,67	Irregular
19	0,96	0,99	0,92	0,81	Irregular
20	0,85	0,86	0,57	0,54	Irregular

The results of the features extraction in Table 1 were classified according to the existing classes by using classification method of MLP. The classification results can be seen in Table 2, where in the measurement column it showed the number of TP = 23, is the number of images with characteristics of regular lesions margins recognized by regular, and TN = 23, is the number of images with characteristics irrreguler margins of lesions recognized irrregular. While FP = 4, is the number of images with regular lesion characteristics identified irregular, and FP = 4, is the number of images with the regular lesion characteristics identified by irregular. From the results of the confusion matrix, the accuracy, sensitivity, and specificity value are calculated, and the results of the classification of features shows convexity, solidity, sircularity, and compactness provided an accuracy value of 85%, sensitivity of 85% and specificity of 85%.

Table 2. Results of The Classification of CT Scan Lung Image

Measurement	**Features (Convexity,Solidity, Circularity, Compactness)**
TP	23
TN	23
FP	4
FN	4
Accuracy	85%
Specificity	85%
Sensitivity	85%

IV. CONCLUSION

Based on the morphological classification of margin CT scan image result by using the Otsu thresholding segmentation method, with several features extraction such as convexity, solidity, circularity, and compactness, and classification with MLP, from 54 image data, it can be concluded that the feature classification with MLP capable to distinguish the characteristics of the regular and irregular lesion margin with an accuracy value of 85%, sensitivity 85% and specificity of 85%.

Thus, it capable to assist the radiologists to interpret the images and can be used as consideration in the diagnosis of lung cancer. Nevertheless, further research is needed to develop methods that can improve the classification accuracy results.

V. Acknowlegment

Researcher is grateful to the Department of Radiology, Faculty of Medicine, University of Gadjah Mada, Yogyakarta, Indonesia, which has been willing to give the image dataset of lung cancer for research purposes, and for any participation in providing information and commentary for this research.

References

[1] "Cancer Media Centre," 2015. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs297/en/. [Accessed: 05-May-2016].

[2] World Health Organization, "Country Cancer Profilles." [Online]. Available: http://www.who.int/cancer/country-profiles/idn_en.pdf?ua=1. [Accessed: 09-Apr-2016].

[3] F. Hosseinzadeh, M. Ebrahimi, B. Goliaei, and N. Shamabadi, "Classification of Lung Cancer Tumors Based on Structural and Physicochemical Properties of Proteins by Bioinformatics Models," *PLoS One*, vol. 7, no. 7, 2012.

[4] J. Cabrera, A. Dionisio, and G. Solano, "Lung Cancer Classification Tool Using Microarray Data and Support Vector Machines," *Information, Intell. Syst. Appl. (IISA), 2015 6th Int. Conf.*, pp. 1 – 6, 2015.

[5] N. Hadavi, J. Nordin, and A. Shojaeipour, "Lung Cancer Diagnosis Using CT-Scan Images Based on Cellular Learning Automata," *Comput. Inf. Sci. (ICCOINS), 2014 Int. Conf.*, pp. 1–5, 2014.

[6] G. Nuemi, F. Afonso, a. Roussot, L. Billard, J. Cottenet, E. Combier, E. Diday, and C. Quantin, "Classification of hospital pathways in the management of cancer: Application to lung cancer in the region of burgundy," *Cancer Epidemiol.*, vol. 37, no. 5, pp. 688–696, 2013.

[7] M. V. Dass, M. A. Rasheed, and M. M. Ali, "Classification of Lung cancer subtypes by Data Mining technique," *Control. Instrumentation, Energy Commun. (CIEC), 2014 Int. Conf.*, pp. 558–562, 2014.

[8] F. Li, S. Sone, H. Abe, H. Macmahon, and K. Doi, "Malignant versus benign nodules at CT screening for lung cancer: comparison of thin-section CT findings.," *Radiology*, vol. 233, no. 3, pp. 793–798, 2004.

[9] R. B. Kuravatti, B. Sasidhar, and R. B. D. R, "A Novel Method for Classification of Lung Nodules as Benign and Malignant using Artificial Neural Network," *Int. J. Eng. Comput. Sci.*, vol. 3, no. 7641, pp. 7641–7645, 2014.

[10] A. Icksan, R. . Faisal, and E. All, "Kriteria Diagnosis Kanker Paru Primer Berdasarkan Gambaran Morfologi Pada CTScan Thoraks Dibandingkan dengan Sitologi." 2008.

[11] Q. Li, F. Li, and K. Doi, "Computerized Detection of Lung Nodules in Thin-Section CT Images by Use of Selective Enhancement Filters and an Automated Rule-Based Classifier," *Acad. Radiol.*, vol. 15, no. 2, pp. 165–175, 2008.

[12] A. Chaudhary and S. S. Singh, "Lung Cancer Detection on CT Images by Using Image Processing," *2012 Int. Conf. Comput. Sci.*, pp. 142–146, 2012.

[13] N. S. Lingayat and M. R. Tarambale, "A Computer Based Feature Extraction of Lung Nodule in Chest X-Ray Image," *Int. J. Biosci. Biochem. Bioinforma.*, vol. 3, no. 6, pp. 624–629, 2013.

[14] S. S. Parveen and C. Kavitha, "Detection of lung cancer nodules using automatic region growing method," *2013 4th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2013*, pp. 4–9, 2013.

[15] Z. Fu and Y. Han, "A Circle Detection Algorithm Based on Mathematical Morphology and Chain Code," *2012 Int. Conf. Comput. Meas. Control Sens. Netw.*, pp. 253–256, 2012.

[16] K. Mya, M. Tun, and A. S. Khaing, "Feature Extraction and Classification of Lung Cancer Nodule using Image Processing Techniques," *Int. J. Eng. Res. Technol.*, vol. 3, no. 3, pp. 2204–2210, 2014.

[17] F. Taher, N. Werghi, and H. Al-ahmad, "Computer Aided Diagnosis System for Early Lung Cancer Detection," *Syst. Signals Image Process. (IWSSIP), 2015 Int. Conf.*, pp. 5–8, 2015.

[18] K. Varalakshmi, "Classification of Lung Cancer Nodules using a Hybrid Approach Percentage %," *J. Emerg. Trends Comput. Inf. Sci.*, vol. 4, no. 1, pp. 63–68, 2013.

[19] M. R. Tarambale and N. S. Lingayat, "Soft Tool Developement for Characterization of Lung Nodule From Chest X-Ray Image," *Int. J. Image Process. Vis. Sci.*, no. 1, pp. 7–12, 2012.

[20] G. Vijaya, A. Suhasini, and P. R, "Automatic Detection of Lung Cancer in CT Images," *IJRET Int. J. Res. Res. Eng. Technol.*, vol. 3, no. 7, pp. 166–172, 2014.

[21] S. K. V. Anand, "Segmentation coupled Textural Feature Classification for Lung Tumor Prediction," *Commun. Control Comput. Technol. (ICCCCT), IEEE Int. Conf.*, pp. 518–524, 2010.

[22] F. V. Farahani, A. Ahmadi, and M. H. F. Zarandi, "Lung Nodule Diagnosis from CT Images Based on Ensemble Learning," *Comput. Intell. Bioinforma. Comput. Biol. (CIBCB), 2015 IEEE Conf.*, pp. 1–7, 2015.

[23] K. Roy, C. Chaudhuri, M. Kundu, M. Nasipuri, and D. K. Basu, "Comparison of the multi layer perceptron and the nearest neighbor classifier for handwritten numeral recognition," *J. Inf. Sci. Eng.*, vol. 21, no. 6, pp. 1247–1259, 2005.