Lung Nodule Detection in CT Scan Image Based on GLCM and RLM Features Using Support Vector Machine (SVM) Method

Zaimah Permatasari
Department of Electrical Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
zaimah.permatasari@gmail.com

Mauridhi Hery Purnomo
Department of Electrical Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
hery@ee.its.ac.id

I Ketut Eddy Purnama
Department of Electrical Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
ketut@te.its.ac.id

***Abstract* — Lung cancer is the most common cause of cancer death globally. Early detection of lung cancer will greatly beneficial to save the patient. This study focused on the detection of lung cancer using classification with the Support Vector Machine (SVM) method based on the features of Gray Level Co-occurrence Matrices (GLCM) and Run Length Matrix (RLM). The lung data used were obtained from the Cancer imaging archive Database, consisting of 500 CT images. CT images were grouped into 2 clusters, including normal and lung cancer. The research steps include: image processing, region of interest segmentation, and feature extraction. The results indicate that the system can detect the CT-image of SVM classification where the default parameter only provides an accuracy of 85.63%. It is expected that the results will be useful to help medical personnel and researchers to detect the status of lung cancer. These results provide information that detection of lung nodules based on GLCM and RLM features that can be detected is better. Furthermore, selecting parameters C and γ on SVM.**

Keywords—cancer, nodule, support vector machine (SVM).

I. INTRODUCTION

Cancer is an abnormal and uncontrolled growth of cells that can harm surrounding tissue and spread far from its origin. It causes death and can grow in any type of cell in the human body [1]. Lung cancer is the most common cause of cancer death, which is not only the leading cause of death in the male population but also in the female population of 13.6% and death from lung cancer was recorded at 11.1%. Malignancy causes death and can grow from any type of cell in the human body [2]. The mortality rate can decline and if cancer can be detected and treated early, the chances of cure are higher. There are two primary types of lung cancer, including Small Cell Lung Cancer (SCLC) and Non-Small Cell Lung Cancer (NSCLC).) [3], The cells in these subtypes differ in size, shape, and chemical make-up when looked at under a microscope. But they are grouped together because the approach to treatment and prognosis (outlook) are often very similar [4].

Early detection of lung cancer will accelerated the patient's recovery. The instrument used to detect lung cancer is a CT Scan (Computed Tomography). CT scan images will provide different results between normal and abnormal lung and lung cancer stages. The inspection process using CT Scan definitely requires quite complex, expensive equipment and requires experts in its implementation. Even experts can make mistakes in distinguishing between normal and abnormal lungs. Therefore, many researchers have provided various alternative solutions to help doctors by utilizing varied image processing techniques [5].

Lung cancer nodules that are observed by these techniques were observed with image processing. The nodule was called pulmonary nodule. A pulmonary nodule is a small round or oval-shaped growth in the lung. It is sometimes also called a spot on the lung or a coin lesion. Pulmonary nodules are generally smaller than 3 centimetres in diameter. If the growth is larger than that, it is known as a pulmonary mass. A mass is more likely to represent cancer than is a nodule [6]. A pulmonary nodule appears in the lung as a spherically shaped mass which can be distorted by surrounding anatomical structure and there are no limitations on size or distribution in lung tissue. The pulmonary nodules are classified into certain categories; nodule connected to the pleural surface, other connections to neighbouring vessels by thin structure. Pre-diagnosis approaches help to locate the risk of lung cancer disease in the very early stage [7].

Several studies have been done in determining lung cancer by using several methods. Research has been conducted by Singh et al (2010) on the development of the Cellular Neural Network (CNN) algorithm to detect the limits and areas of lung cancer from X-ray images [8]. Lung cancer detection using artificial neural network and fuzzy clustering methods investigated by another study [9]. Research conducted by Almas Pathan and Bairu Saptalkar detects lung cancer using Neural Network on X-ray images [10]. Other research has been done by Zagreb and Croastia regarding the system for the classification of asthma and chronic obstructive pulmonary disease (COPD) based on fuzzy rules and the trained neural network [11]. also proposed a classification method using SVM. The first stage, extraction and reconstruction of the lung parenchyma is performed and then scaled up to highlight its structure. Extraction is done with Gaussian Filter and Median Filter [12].

This study focuses on the detection of lung cancer on CT scan images using the Support Vector Machine (SVM) method based on the features of Gray Level Co-occurrence Matrices (GLCM) and Run Length Matrix (RLM). It is hoped that it will be useful to assist medical personnel and researchers to detect the status of lung cancer.

II. METHODS

A. Material

The lung data used originates from the cancer imaging archive database consisting of SO CT-images. CT image is grouped into 2 clusters, normal and lung cancer. Normal CT image consists of 50 images and 60 lung cancer tomography images. Figure 1 (a) shows the example of a normal lung CT image while Figure (b) shows a lung cancer image.

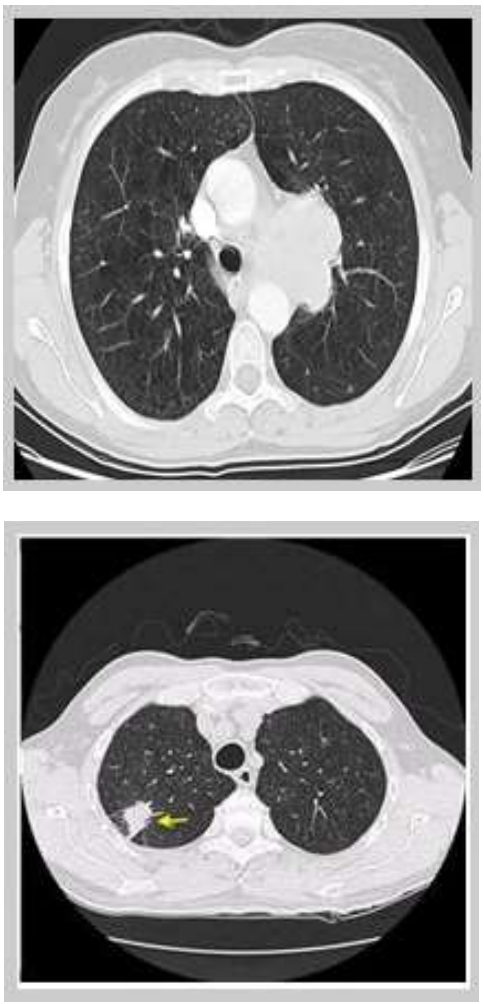

Fig. 1. Image of CT Lungs (a) Normal (b) Lung cancer

B. Method

Before detecting the status, tomography image enhancement is done. The first step is preprocessing by doing grayscale and binary, then removing noise using a median filter. Adaptive histogram equalization is used to improve the contrast. Afterward, the segmentation process is applied to find a segment of the region of interest. Features are unique characteristics of an object, which are divided into two, including natural and artificial features. The natural features are an ever-present part of the image such as the brightness and edges of the object. Meanwhile, the artificial features are those obtained by a particular operation on the image, such as gray level histogram, etc. [13].

Textures are highly helpful in characterizing various images. They are commonly utilized by the human visual system for recognition and interpretation [14]. The intensity values corrected by the image are then corrected according to the GLCM matrix to extract their characteristics: homogeneity, energy, contrast, correlation, variance, and the RLM of the characteristics.: SRE1, LRE1, GLN1, RP1, RLN1, LGRE1, HGRE1.

SVM classification is performed using a polynomial kernel with data samples obtained from GLCM and RLM feature extraction.

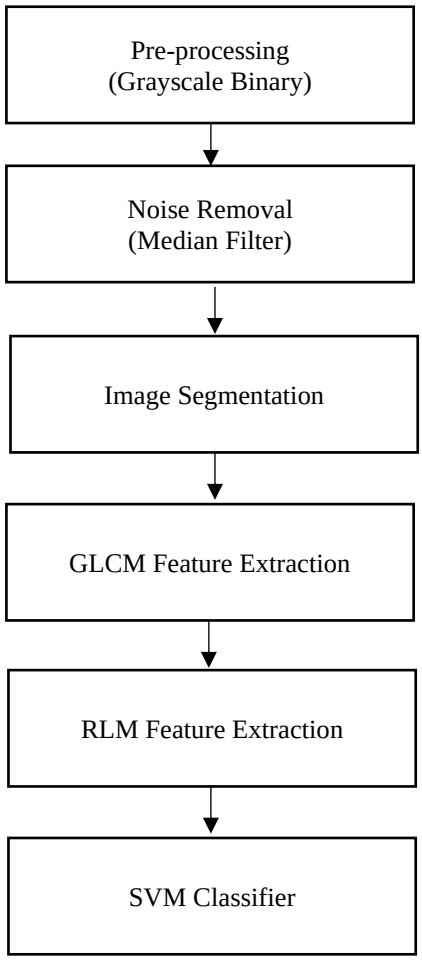

Fig. 2. Process of the research

III. RESULTS AND DISCUSSIONS

A. Image Processing

Image processing methods used for intensity improvement are grayscale, binary, median filters, and adaptive histogram equations. The CT images used have a standard dimension of 207x208 pixels. The results of pre-processing are shown in Figure 3. The result of the grayscale image is shown in Figure 3 (b). The binary image is shown by (c) while noise removal results using the median filter is shown by (d). Results present of histogram adaptive equalization process.

The grayscale process to convert images from RGB to grayscale images is the step to convert to the binary image. The median filter method is very useful for eliminating outliers/noise. Figure 3(c) shows that the image noise in lung cancer tomography has been reduced. The adaptive histogram equations aim to improve the image quality so that it is clearer and more intelligible. This can be seen from the image of the

result of the adaptive histogram equation, where the image results are ready for the image segmentation process.

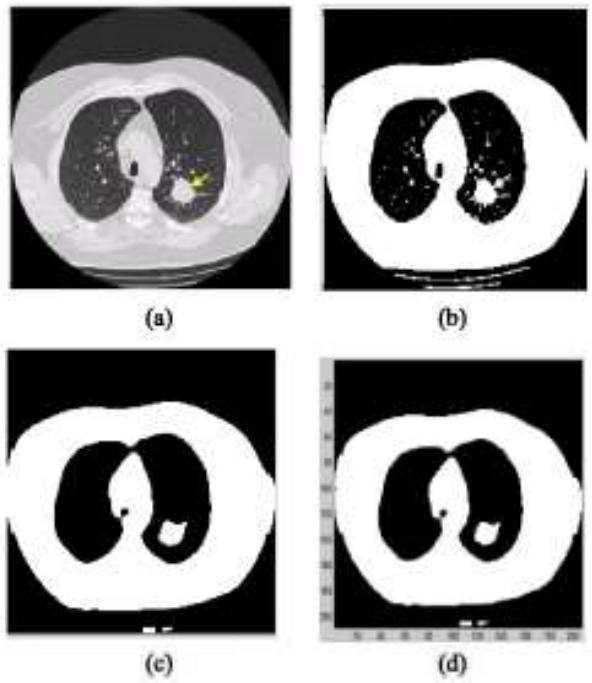

Fig. 3. (a) Grayscale image (b)Binary image (c) Noise removal reult using median filter (d) Results of Histogram adaptive equalization process

B. Image Segmentation

Segmentation method used to get the area of lung cancer is region props used to find the value of this centroid is represented in the form of coordinates X and Y which states the central axis coordinates. Once obtained the value of centroid then done the process of labelling. Labelling is used to recognize all objects that have the potential as a nodule. Objects that Eire candidates of the affected area will be normalized and then searched for its feature information. Figure 4 is an example result of image segmentation.

Fig. 4. Centroid location

C. Feature Extraction

In texture feature extraction, the difference is the texture that is the determinant characteristic of the image. The statistical technique for texture feature extensions is GLCM & RLM.

1) GLCM (Gray Level Co-occurrence Matrix)

GLCM is utilized to measure the spatial dependence of grey levels in an image. In GLCM, the number of rows and columns exactly equals the number of grey levels in the image. The co-occurrence matrix is constructed in four spatial orientations (0º, 45º, 90º, and 135º). The value of each element is updated to coincide at the same pixel. The texture features measured using GLCM include contrast, correlation, dissimilarity, energy, entropy, homogeneity, mean, variance, and standard deviation.

$$Contrast = \sum i \sum j \, (i-j)^2 C(i,j) \quad (1)$$

$$Energy = \sum i \sum j \, C^2(i,j) \quad (2)$$

$$Entropy = \sum i \sum j \, C(i,j) \log C(i,j) \quad (3)$$

$$Homogeneity = \sum i \sum j \frac{C(i,j)}{i+|i+j|} \quad (4)$$

$$Correlation = \mathrm{u} \sum_i^{L_i} = 1 \sum_j^{L_j} = 1 \quad (5)$$

2) RLM (Run Length Matrix)

In RLM, the statistic concerned is the number of value pairs at the grey level [15]. GLRLM feature extraction has 7 attributes, namely SRE, LRE, GLN, RP, RLN, LGRE, HGRE.

$$SRE = \frac{1}{n_r} \sum_{\mathrm{i}=0}^{\mathrm{M}} \sum_{\mathrm{j}=0}^{\mathrm{N}} \frac{\mathrm{p(i,j)}}{\mathrm{j}^2} \quad (6)$$

$$LRE = \frac{1}{n_r} \sum_{i=0}^{M} \sum_{j=0}^{N} \mathrm{p(i,j)} * \mathrm{j}^2 \quad (7)$$

$$GLN = \frac{1}{n_r} \sum_{i=0}^{M} \left(\sum_{j=0}^{N} p(i,j) \right)^2 \quad (8)$$

$$RP = \frac{n_r}{\sum_{i=0}^{M} \sum_{j=0}^{N} p(i,j) * j} \quad (9)$$

$$RLN = \frac{1}{n_r} \sum_{j=0}^{N} \left(\sum_{i=0}^{M} p(i,j) \right)^2 \quad (10)$$

$$LGRE = \frac{1}{n_r} \sum_{\mathrm{i}=0}^{\mathrm{M}} \sum_{\mathrm{j}=0}^{\mathrm{N}} \frac{\mathrm{p(i,j)}}{\mathrm{i}^2} \quad (11)$$

$$HGRE = \frac{1}{n_r} \sum_{i=0}^{M} \sum_{j=0}^{N} \mathrm{p(i,j)} * \mathrm{i}^2 \quad (12)$$

The GLRLM attribute is used to measure the distribution of short runs, to measure the distribution of long runs, to measure the similarity of gray level values throughout the image, to measure the homogeneity and distribution of runs from an image, to measure the long similarity of runs across the entire image, to measure the distribution of low gray level values. (low gray level values), and measures the distribution of high gray level values (Galloway).

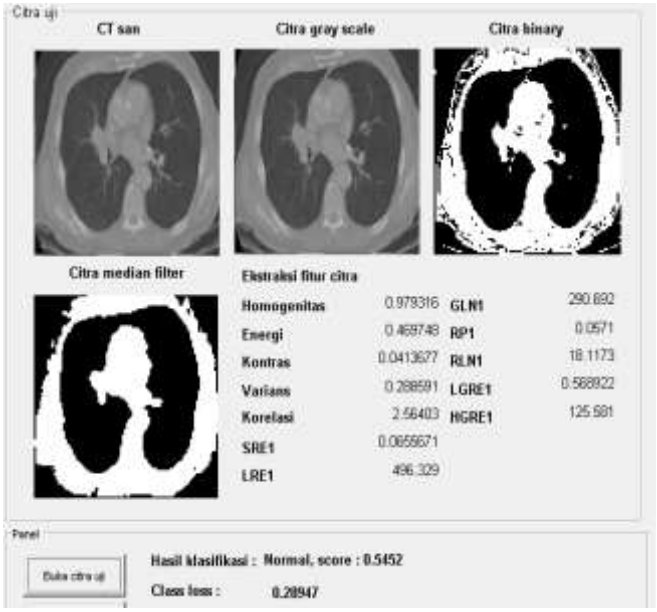

Fig. 5. The results of processing and feature extraction of tests

The GUI presents the results of processing and feature extraction of test figure 5 which is shown by the parameter values of homogeneity, energy, contrast, variance, correlation, SRE1, LRE1, GLN1, RP1, RLPN1, LGRE1, and HGRE1. The results of the process of the image classification test are shown in Figure 5, in which image 1 indicates class 1, which is defined as a normal class with a classification result of 0.5452 and a class loss of 0.28947.

TABLE I. FEATURE EXTRACTION OF GLCM DEFINED AS NORMAL CLAS

Homogeneity	Energy	Contrast	Correlation	Variants
0.9793	0.4697	0.0414	25,640	0.2886
0.9805	0.4630	0.0390	25,097	0.2948
0.9790	0.5060	0.0420	26,622	0.2633
0.9838	0.5258	0.0325	26,823	0.2487
0.9745	0.5493	0.0510	27,544	0.2367
0.9873	0.6094	0.0255	28,420	0.1991

TABLE II. FEATURE EXTRACTION OF RLM DEFINED AS NORMAL CLASS

SRE	LRE	GLN	RP	RLP
0.0243	34,363	5,634,676	0.0254	215,096
0.0366	47,944	4,685,083	0.0211	177,078
0.0288	30,365	5,584,095	0.0254	220,768
0.0370	69,038	3,964,822	0.0176	184,848
0.0434	40,677	4,998,205	0.0226	180,995
0.0071	46,604	4,920,618	0.0218	241,429

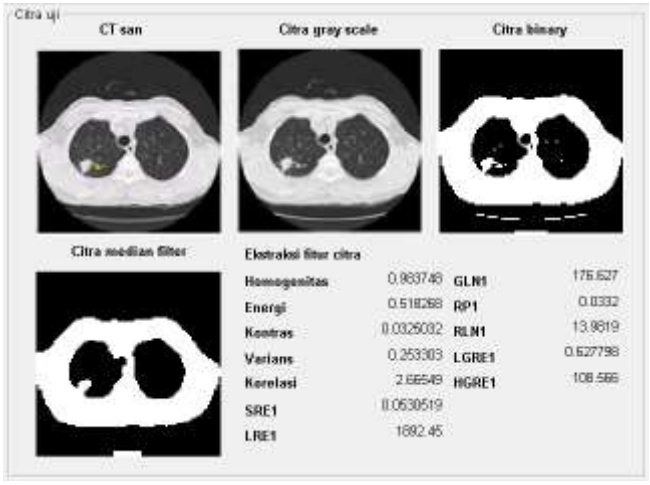

Fig. 6. The results of processing and feature extraction of tests.

GUI shows the results of processing and feature extraction of test Figure 6 which is shown by the parameter values of homogeneity, energy, contrast, variance, correlation, SRE1, LRE1, GLN1, RP1, RLPN1, LGRE1, and HGRE1. The results of the image classification test process are shown in Figure 4.32, in which image 6 indicates class 2 which is defined as a cancer class with a classification result of 0.83556 and a class loss of 0.26315.

TABLE III. FEATURE EXTRACTION OF GLCM DEFINED AS CANCER CLASS

Homogeneity	Energy	Contrast	Correlation	Variants
0.9796	0.5148	0.0408	26,779	0.2573
0.9816	0.4719	0.0368	25,487	0.2857
0.9781	0.4959	0.0437	26,447	0.2703
0.9835	0.5216	0.0331	26,746	0.2514
0.9809	0.4645	0.0382	25,157	0.2929
0.9803	0.5271	0.0395	27,000	0.2493

TABLE IV. FEATURE EXTRACTION OF RLM DEFINED AS CANCER CLASS

SRE	LRE	GLN	RP	RLP
0.0542	17,196	686,375	0.0313	236,405
0.0184	47,754	503,486	0.0223	260,811
0.0022	56,481	434,509	0.0190	250,855
0.0036	23,827	595,745	0.0268	287,443
0.0208	19,038	408,307	0.0122	284,848
0.0068	40,677	504,365	0.0247	180,995

D. Cancer Lung Detection

In the learning process using SVM, the extracted image data were divided into training data and test data. In this study, data distribution was done using k-fold = 5. Therefore, the distribution of training data was 96 consisting of 48 normal data and 48 tumor data. The 24 tests data consists of 12 normal data and 12 tumor data. In the classification process, the data are binary separated to normal class and tumor class with linear kernel approach, RBF kernel, polynomial kernel, and sigmoid kernel. The sample data for calculating the SVM classification process using a polynomial kernel with 20 data samples from GLCM and RLM feature extraction is explained as follows. Classification data is presented in the table along with the label values as follows:

TABLE V. OF TEST RESULTS OF CT SCAN IMAGE

No	Text Image	Accuracy	Loss Class	Image Class	Prediction	Results
1	Image 1	0.5452	0.2895	1	1	Normal
2	Image 2	0.1985	0.2895	1	1	Normal
3	Image 3	0.7454	0.3158	1	1	Normal
4	Image 4	0.4700	0.2895	1	1	Normal
5	Image 5	0.5843	0.2895	1	1	Normal
6	Image 6	0.2894	0.2334	1	1	Normal
7	Image 7	0.3098	0.1343	1	1	Normal
8	Image 8	0.1984	0.3422	1	1	Normal
9	Image 9	0.2912	0.2734	1	1	Normal
10	Image 10	0.4291	0.1433	1	1	Normal
11	Image 11	0.8356	0.2632	2	2	Cancer
12	Image 12	0.7523	0.2895	2	2	Cancer
13	Image 13	0.8131	0.3158	2	2	Cancer

TABLE VI. IMAGE CLASSIFICATION RESULTS BASED ON KERNEL SVM

Data	Kernel	Acu	Spec	Sens
Normal Image	Linear	69.90%	48.89%	100.00%
	RBF	60.89%	50.45%	92.43%
	Sigmoid	61.43%	43.63%	88.73%
	Polynomial	75.34%	53.76%	95.27%
Cancer Image	Linear	74.86%	57.92%	85.23%
	RBF	82.63%	60.23%	75.32%
	Sigmoid	70.24%	51.82%	83.65%
	Polynomial	85.63%	66.35%	100.00%

The learning process carried out is based on GLCM and RLM data on SVM which consists of linear kernel, RBF, sigmoid, and polynomial. In the linear kernel, the default parameter used is parameter C = 1. In the RBF kernel, the default parameter used is parameter C = 1 and = 1.8. In the sigmoid kernel, the default parameters used are C = 1, = 1.8, and r = 0.0. In the polynomial kernel, the default parameters used are C = 1, = 1.8, r = 0.0, and d = 3. The best learning result based on the classification process using SVM is at an accuracy of 85.63% with a specificity of 66.35%, and sensitivity of 100.00%. The results obtained from the SVM classification using the default parameter only provide an accuracy of 85.63%.

IV. CONCLUSIONS

The conclusions obtained can be mentioned, namely:

1. Because the texture feature extraction method tends to get the texture characteristics of the CT Scan image However, using a combination of extraction features of GLCM and RLM can improve classification accuracy. This is explained by increasing the number of features that can be extracted by using two methods of feature extraction.
2. The appearance of noise in an MRI image can be denoised using a median filter with thresholding set to obtain a cancer object which is said to be an optimal metric value [4 4]. Three holding values that are too high result in the loss of cancer objects from the MRI image, failing the classification process.
3. The results of image feature extraction with the acquisition of SVM classification with default parameters only provide 85.63% accuracy. According to the assessor, this value is included in the sufficient category. Therefore, in this study, optimization of SVM parameters was carried out to get the best hyperplane. The best and optimal hyperplane will provide a high system accuracy value. If the accuracy is high, then the system can classify the data according to its class properly.

REFERENCES

[1] American Cancer Society. (2013). What are the risk factors for non-smallcell lung cancer?. 19 Desember 2013. < http : // w w w . cancer. org>.

[2] Ancuceanu, R. V., and Victoria, I, 2004, Pharmacologically Active Natural Compounds for Lung Cancer, Altem. Med. Rev., 9 ,4 ,4 0 2 -4 1 9

[3] M.S.Tarawneh, "Lung Cancer Detection Using Image Processing Techniques", Leonardo Electronic Journal o f Practices and Technologies, Issue 20, pp. 147-158, January-June 2012.

[4] http://www.cancer.org/cancer/lungcancer-nonsmallcell/detailedguide/non-small-cell-lung-cancer-what-is-non-small-cell-lung-cancer

[5] Kusumadewi, S.. Membangun Jaringan Syaraf Tinian dengan M atlab dan Exel Link. Yogyakarta: Graha Emu.

[6] Cleveland Clinic Medical Professional (2021). Pulmonary Nodules. 28 May 2021. < https://my.clevelandclinic.org/health/diseases_conditions /hic_Pulmonary_Nodules >

[7] Vinod kumar, Kanwal Garg," Neural network based approach for detection of abnormal regions of lung cancer in x ray", International journal of engineering research & Technology (IJERT), ISSN:2278-0181, vol 1 Issue 5, July-2012 p 1-7

[8] S. Singh, R. Vijay, Y. Singh, "Artificial Neural Network and Cancer Detection", IOSR Journal o f Computer Engineering, pp. 20-24.

[9] F. Taher, N. Werhi, H. Al-Ahmad, R. Sammouda, "Lung Cancer Detection by Using Artificial Neural Network and Fuzzy Clustering Methods",American Journal o f Biomedical Engineering 2 012,2(3), pp. 136-142.

[10] Almas Pathan, Bairu.K.saptalkar, "Detection and Classification o f Lung Cancer Using Artificial Neural Network", International Journal on Advanced Computer Engineering and Communication Technology Vol-1 Issue:1

[11] A. Badnjevic, M. Cifiek, D. Koraga, D. Osmankovic, "Neuro-fiizzy classification o f asthma and chronic obstructive pulmonary disease", BMC Med Inform Decis Mak. 2015; 15(Suppl 3): SI. Published online 2015 Sep 11.

[12] A. O. De Carvalho Filho, W. B. De Sampaio, A. C. Silva, A. C. de Paiva, R. A. Nunes, and M. Gattass, "Automatic detection of solitary lung nodules using quality threshold clustering, genetic algorithm and diversity index," Artif. Intell. Med., vol. 60, no. 3, pp. 165–177, 2014.

[13] R.M.Haralick, K. Shanmugam, R.M.Haralick, K. Shanmugam,, 'Textural features for image classification," IEEE Trans. System Man. Cybernetics, vol. SMC-3, pp. 610-621,1973

[14] D. Tian, "A Review on Image Feature Extraction and Representation Techniques", International Journal o f Multimedia and Ubiquitous Engineering Vol. 8, No. 4, July, 2013, pp. 385-396.

[15] Nurtanio, Inggrid, Astuti, E.R.,Purnama, I.K.E., Hariadi, M., Purnomo, M. H., Classifying Cyst and Tumor Lesion Using Support Vector Machine Based on Dental Panoramic Images Texture Feature, IAENG International Journal of Computer Science, Vol. 40, No.1, Feb 2013.