Computer Aided Diagnosis System based on Machine Learning Techniques for Lung Cancer

Hamada R. H. Al-Absi[1], Brahim Belhaouari Samir[2], Khaled Bashir Shaban[3], and Suziah Sulaiman [1]
[1]Department of Computer & Information Sciences,
[2]Department of Fundamental and Applied Sciences,
Faculty of Science and Information Technology
Universiti Teknologi PETRONAS,
31750 Tronoh, Perak, Malaysia
[3]Computer Science and Engineering Department,
College of Engineering,
Qater University, P.O. Box:
2713, Doha, Qatar

hamada.it@gmail.com, brahim_belhaouari@petronas.com.my,
khaled.shaban@qu.edu.qa, suziah@petronas.com.my

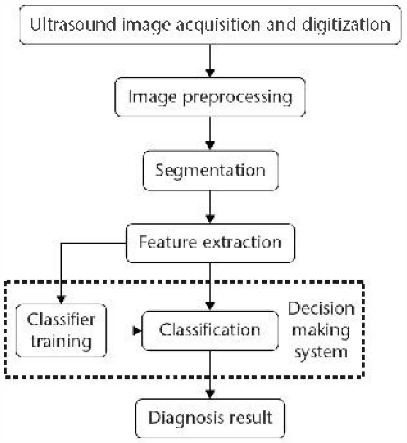
Abstract - **Cancer is a leading cause of death worldwide. Lung cancer is a type of cancer that is considered as one of the most leading causes of death globally. In Malaysia, it is the 3rd common cancer type and the 2nd type of cancer among men. In this paper, machine learning techniques have been utilized to develop a computer-aided diagnosis system for lung cancer. The system consists of feature extraction phase, feature selection phase and classification phase. For feature extraction/selection, different wavelets functions have been applied in order to find the one that produced the highest accuracy. Clustering-K-nearest-neighbor algorithm has been developed/utilized for classification. Japanese Society of Radiological Technology's standard dataset of lung cancer has been used to test the system. The data set has 154 nodule regions (abnormal) and 92 non-nodule regions (normal). Accuracy levels of over 96% for classification have been achieved which demonstrate the merits of the proposed approach.**

I. Introduction

Cancer is a leading cause of death worldwide. According to the World Health Organization (WHO) [1], cancer accounts for 7.4 million deaths (13% of all deaths) in 2008. More than 70% of all cancer deaths take place in low and middle income countries. It is projected that deaths caused by cancer will be rising to reach 13.1 million by the year 2030 [1]. In Malaysia, according to the National Cancer Registry (NCR) [2], the top five common types of cancer that affect the population regardless of sex are breast, colorectal, lung, cervix and nasopharynx. Other types that are also affecting Malaysians include liver cancer and prostate cancer. 21,773 cases were registered in NCR in the year 2006 that have been affected by one type of cancer in Malaysia. Since the cause of cancer is still unknown, early detection and treatment of cancer is considered the most promising approaches to reduce the number of deaths. In order to diagnose diseases, medical imaging modalities such as Mammography, Computed Tomography and Magnetic Resonance Imaging, etc. have been developed to produce images of organs that help to identify abnormalities. Radiologists refer to these images to diagnose diseases; however, radiologists may miss subtle regions or lesions that are small in size and difficult to perceive [3]. Several Computer Aided Diagnosis (CAD) systems have been developed to assist radiologists to detect subtle regions and identify cancerous cells [4] some of these systems will be discussed in section 2.

A. Lung Cancer

Lung cancer is one of the major causes of death in many parts around the world [5]. GLOBOCAN – which is a project under the International Agency for Research in Cancer (part of WHO) that aims "*to provide contemporary estimates of the incidence of, mortality and prevalence from major type of cancers, at national level, for 184 countries of the world*" - reported that lung cancer is the first type of cancer among men and the third for both sexes globally [5]. In Malaysia, it is the 2nd common cancer among men and the third in general with 2100 Malaysians being affected by this disease every year [2]. Specialized CAD systems have been developed to aid radiologists in detecting pulmonary nodules that indicate the existence of lung cancer [8-11]. CAD systems in general consist of three stages; preprocessing, feature extraction and classification. Figure 1 shows a block diagram of the main processing components of a computer aided diagnosis system.



The rest of the paper is organized as follows: Section II presents some related work, Section III explains the machine learning system, section IV reports the results

978-1-4673-1938-6/12/$31.00 ©2012 IEEE

obtained and section V concludes the finding and some future work to be carried out.

II. Related Work

Many techniques have been developed in order to increase the detection accuracy rates of lung cancer in CAD systems. For instance, Hardie et al. [8] presented a CAD system for pulmonary nodule detection in chest radiography. The proposed system was tested using a data set that consists of 167 chest radiography that contains 181 lung nodules; the system utilized adaptive distance-based threshold algorithm for nodule segmentation, after that, features were computed for each nodule using geometric features, intensity features and gradient features. Lastly, a Fisher linear discriminant classifier was used to classify the computed features. The system could detect 78.1% of those nodules. Lee et al. [9] proposed a lung nodule detection using an ensemble classifier aided by clustering. Lung scans of 32 patients who included 5721 images were used to test the method. The system obtained sensitivity of 98.33% and specificity of 97.11%. Another system was proposed by Dehmeshki et al. [10] to detect lung nodules using shape-based genetic algorithm template matching (GATM). In this system, a preprocessing step for enhancement was performed using spherical-oriented convolution-based filtering scheme, and a 3D geometric shape feature was calculated to obtain the fitness function. The system was evaluated using a data set of 70 CT images containing 178 nodules. The system could detect 160 nodules with a rate of 90%. An automatic method for lung cancer detection was also introduced by Sousa et al. [11]. The system has six stages, namely, thorax extraction, lung extraction, lung reconstruction, structure's extraction; tubular structures elimination, and false-positive reduction. Each stage performs a specific task that resulted in detecting lung nodules. The sensitivity achieved was 84.84%, and specificity achieved was 96.15%.

These were examples of systems that have been developed to detect lung cancer. General issues of existing systems are the high number of false positives and false negatives. Therefore, there is a continuous need and importance to develop computer-aided diagnosis to help in lung cancer detection and prediction. This work aims at developing a system with advanced and effective capabilities in classifying of lung nodules and reducing the number of false positives and false negatives. In particular, the proposed system presents two stages of feature selections which results in selecting only relevant features that result in better performance. In addition, we apply the Cluster k-nearest neighbor algorithm which combines two algorithms K means clustering and K-nearest neighbor, this classifier has been reported to produce high accuracy in [13].

III. Lung Cancer Diagnosis System with Cluster K-Nearest Neighbor Classifier

This section presents a computer aided system for lung cancer diagnosis. The system begins by training the classifier using an image dataset and then a separate testing dataset is used to validate the performance. The system consists of feature extraction, feature selection and classification stages. The following sub-sections explain each stage.

A. *Feature Extraction based on Wavelet Transform*

Feature extraction is an important step in CAD systems. The extraction of features that represent medical images of specific organs is an important issue. Multi-resolution methods such as Wavelet Transform (WT) have been widely used in feature extraction due to its quality compared with other feature extraction methods.

The transform of a signal is considered as a different way of representing the signal. And it save the information content present in the signal. Wavelets [12] make use of different sets of basic functions to permit the decomposition of continuous and discrete signals. Wavelet Transform offers a time-frequency representation of the signal.

The Continuous Wavelet Transform (CWT) is provided where x (t) is the signal to be analyzed. Ψ (t) is the mother wavelet or the basis function. The mother wavelet is considered as source of all the wavelet functions used in the transformation during translation (shifting) and scaling (dilation or compression).

$$X_{WT}(\tau,s)=\frac{1}{\sqrt{|s|}}\int x(t).\Psi^{*}\left(\frac{t-\tau}{s}\right)dt \qquad (1)$$

In MATLAB, wavelet is implemented as a toolbox. This toolbox offers many wavelet functions that can be utilized. Wavelet families such as Daubechies, Haar, Symlet, Coiflet, Biorthogonal, Meyer, Battle-Lemarie, Mexican Hat and Morlet can easily be implemented with any MATLAB code. In this paper, Haar (haar) function, Daubechies (db1) function and Symlet (sym3) function were used to test the method.

B. *Feature Selection*

Once the feature extraction stage has been executed, a huge number of coefficients will be produced. It is important to reduce the coefficients by selecting those coefficients that contains the most important information that would contribute to high accuracy and ignoring the remaining. For this reason, we use two steps for feature selection through calculating the variance and then the energy.

We calculate the variance as follows:

Suppose m_1, m_2 and m_3 are the mean of class1, class2, and class3 respectively and m_T is mean of all classes.

Let $m_T=\frac{\sum_{i=1}^{n} m_i}{n}, n = number\ of\ classes$

so that $\sum(m_i-m_T)^2$ is not sufficient to quantify the classification contribution of the coefficients because it may give same values in the two cases. Therefore, there is a need to introduce another metric to quantify the coefficients contribution. We introduce another metric as follow:

$$Var\_mod - \frac{1}{n}\left[\sum_{i=1}^{n}\frac{(m_i-m_T)^2}{Var_i}\right], \qquad (2)$$

where Var_i is variance of the class i, m_i is mean of class i, m_i is mean of all classes and n is the number of classes. Var_i will be calculated using the following formula:

$$Var_i = \frac{\sum_{j-1}^{n_i} \left(x_j^i - m_i \right)^2}{n-1_i} \qquad i = 1, 2, 3, \ldots\ldots n_i \tag{3}$$

Where n_i is number of the features in class i.

The way to select the desired features coefficients will be as follow:

If variance modified of any feature is less than certain threshold $Var\_mod \le 1$, we will remove it, else we keep feature.

Energy is used as well to select the relevant coefficient, for that we apply the statistical energy to reduce the dimensions of the overall features by selecting only the features with high statistical energy values, since these features contains high amount of lung information more than the other features with low statistical energy value.

C. *Classification*

A classifier that is a combination between K-means modified algorithm and K-Near Neighbor (K-NN) is applied in this research. This classifier was developed by Samir Brahim [13]

- *Cluster-K-Nearest Neighbor Classifier (C-K-NN)*

Cluster-K-Nearest Neighbor is a classifier that combines two algorithms that are K-means modified algorithm and K-Nearest neighbor. K-means is used to cluster the data into classes and sub-classes with a centre point to represent each class and K-Nearest Neighbor is used to classify new data by calculating the Euclidean distance between the centre point of each class and the new data. With this combination the classification is more accurate in less time. The algorithm is explained further below:

Firstly, we will cluster each class C_i to number of sub-classes $C_{i,j}$, with means $\mu_{i,j}$, with $1 \le j \le m_i$ where m_i is number of sub-classes using K-means modified algorithm. This procedure will minimize the variance within each cluster and maximize the variance between the clusters. As traditional K-means algorithm suffering from determining the number of sub-classes and the initial K-vector, in this classifier, two algorithms were developed to choose the optimum initial K-vector for the minimum variance, namely near to near and near to mean. The next Figure 2 will show that each class C_i will be divided into number of sub-classes $C_{i,j}$ represented by the mean $\mu_{i,j}$ of the data.

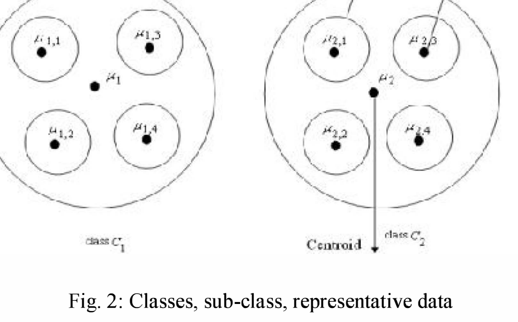


Fig. 2: Classes, sub-class, representative data

○ *Near to Near Algorithm*

This algorithm calculates the distance $d(x_{i,n}, x_{i,m})$ for all $x_i \in C_i$, and then starts to cluster each class to $N_i - 1$ sub-classes, where $card(C_i) = N_i$. Then each two closet data will put in the same sub-class $C_{i,1} = \{x_{i,n_0}, x_{i,m_0}\}$,

where: $\min\limits_{n \ne m} d(x_{i,n}, x_{i,m}) = (x_{i,n_0}, x_{i,m_0})$

The other data will be gathered at separated sub-class

$C_{i,j} = \{x_{i,j}\}, \forall j \in \{1\ldots\ldots N_i\} - \{n_0, m_0\}$

Now the index n_1 and m_1 be

$$\min_{\substack{n \ne m \\ (n,m) \ne (n_0,m_0)}} d(x_{i,n}, x_{i,m}) = (x_{i,n_1}, x_{i,m_1}) \tag{4}$$

The sub-class $C_{i,r}$ will split into two other sub-classes if the x_{i,n_1} and x_{i,m_1} belong to it.

But if they belong to different sub-classes, they will be put in that classes based on the classes *card*. The iteration will stopped after K-subclasses were obtained and the initial K-vector will become the means of each sub-classes.

○ *Near to Mean algorithm*

It's similar to near to near algorithm, however it deals with the mean of sub-class, therefore, the class C_i will split into two sub-classes:

$$C_{i,1} = \{x_{i,n_0}, x_{i,m_0}\} \tag{5}$$

and

$$C_{i,1} - \{x_{i,j} | j \notin \{n_0, m_0\}\} \tag{6}$$

where, $d(x_{i,no}, x_{i,mo}) = \min\limits_{n \ne m} (x_{i,n}, x_{i,m})$

Then the class will be updated C_i by replacing $x_{i,no}$ and $x_{i,mo}$ with their average,

$$\left| \begin{matrix} C_i^1 = \{\ldots\ldots x_{i,no-1}, s_0, x_{i,no+1}, \ldots\ldots x_{i,mo-1}, s_0, \\ x_{i,mo+1}, \ldots\} \end{matrix} \right| \tag{7}$$

where $so = (x_{i,no} + x_{i,mo}) / 2$.

then $x_{i,n1}$ and $x_{i,m1}$ will be

$$d(x_{i,n1}, x_{i,m1}) = \min d(x_{i,n}, x_{i,m}) \Big| d(x_{i,n}, x_{i,m}) \ne 0 \tag{8}$$

Then C_i^1 will replace all the data in that are equal to $x_{i,n1}$ or $x_{i,m1}$ by s_1, which is the mean of the union of the two subclasses where $x_{i,n1}$ and $x_{i,m1}$ belong to:

$$s_1 = \frac{C_{n1} x_{i,n1} + C_{m1} x_{i,m1}}{C_{n1} + C_{m1}} \tag{9}$$

Where, C_{n1} is the number of iteration of $x_{i,n}$ inside of

C_i^l and C_{ml} is the number of repetition of $x_{i,ml}$ inside of C_i^l. The algorithm will stop once the number of distinct vectors inside of C_i^r is equal to k. This classification algorithm does not need to keep all the data, but only the average of each subclass. This is the outstanding feature of this new clustering. To classify a new data or vector x, k-NN algorithm will be used, i.e., we assign x to the class $C_{\hat{i}}$ for which $\hat{i} = arg_i \min_{i,j} d(x, \mu_{i,j})$,

where $arg_i \; d(x, \mu_{io,jo}) = i_o$.

Further investigation about the k-NN algorithm is indeed needed to find the closest j- examples in the dataset and select the predominant class. The smallest closest examples could be found in the dataset and select the predominant class which have exactly k examples. Now, the mathematical derivation of the new classifier is clear. Figure 3 shows a diagram of the system.

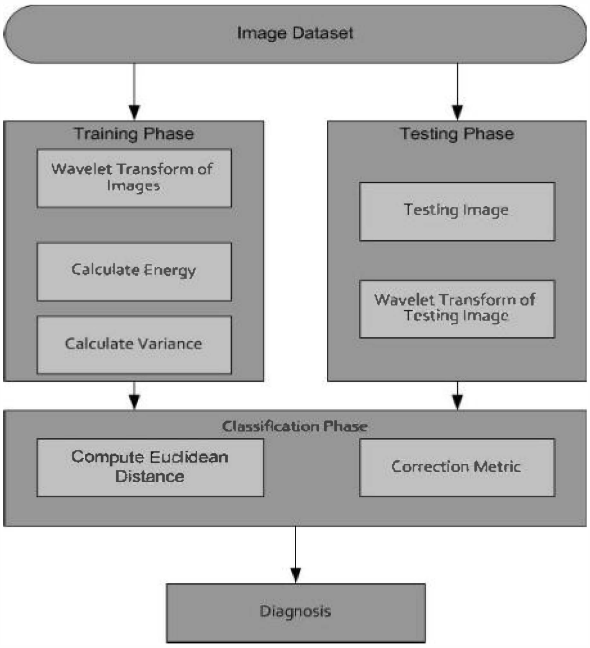


Fig. 3: The proposed system for lung diagnosis

IV. RESULTS & DISCUSSION

A standard dataset has been used to carry out experiments to examine the efficacy of the proposed CAD system. The collected results show the merits of the approach and encourage further investigation in this direction.

D. *Dataset*

JSRT (Japanese Society of Radiological Technology) standard dataset of chest radiographs [14-15] is used to test the methods introduced. The dataset contains 247 chest radiographs, where 154 images contain nodules (Abnormal) and 93 images do not contain any nodules (Normal). Regions of 60x60 were selected from the original 2048x2048 images. Figure 4 shows an example of one chest radiography.

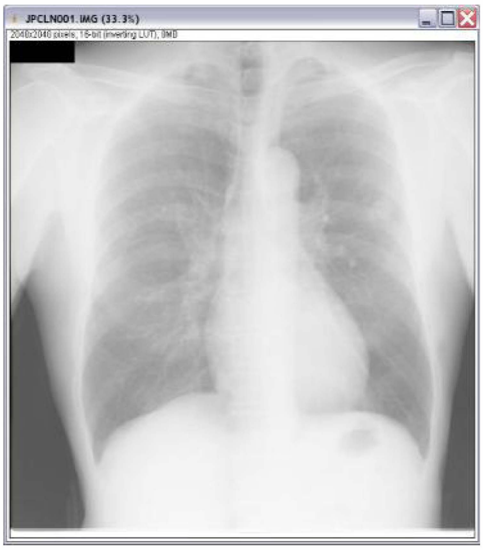


Fig. 4: Example of the JSRT (JPCLN001.IMG) [13]

The abnormal regions were selected based on the coordinates that are provided with the dataset, and the normal regions were extracted randomly from the 93 non nodule images. Figure 5 shows example selected regions used for training and testing of the system

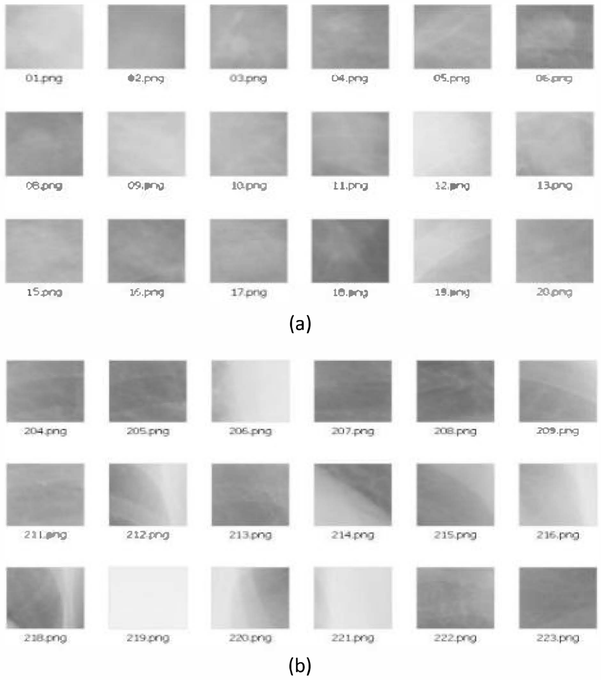


Fig. 5: a) Abnormal Images affected by nodules, b) normal images

E. *Results*

The system which was developed in MATLAB has been tested with many wavelet functions to determine the function and the level (scale level) that produces the optimum result. The table below (Table 1) reports the results of the method when three wavelets functions (haar, db1 and sym3) were tested.

TABLE I
RESULTS WITH THREE WAVELET FUNCTIONS AT DIFFERENT SCALES

Function Level	Performance	Function		
		haar	db1	sym3
1	Accuracy	0.9060	0.8889	0.9231
	False Positives	0.0465	0.0465	0
	False Negatives	0.0541	0.0676	0.0270
2	Accuracy	0.9487	0.9573	0.9316
	False Positives	0.0233	0.0233	0.0233
	False Negatives	0.0135	0.0135	0.0405
3	Accuracy	0.9573	**0.9658**	0.9487
	False Positives	0.0233	0	0.0233
	False Negatives	0.0135	0.0135	0.0405
4	Accuracy	0.9573	0.9573	0.9573
	False Positives	0.0233	0.0233	0
	False Negatives	0.0135	0.0135	0.0405
5	Accuracy	0.9573	0.9573	0.9573
	False Positives	0	0.0233	0
	False Negatives	0	0.0135	0.0270
6	Accuracy	0.9487	0.9573	0.9573
	False Positives	0.0233	0.0465	0.0233
	False Negatives	0.0135	0.0270	0.0405

As shown in the table above, function db1 produces the highest accuracy with 0.9658 at level 3. In comparison, the performance of the other two functions ranges between 0.9060 and 0.9573. Furthermore, haar function at level 5 produced zero for false positive and false negative with an accuracy of 0.9573.

Figure 6 demonstrates the effect of the number of coefficients to the performance of the classifier. In this figure (where db1 wavelet function was used with level 3), it is shown that:

- When the number of coefficients is large, low accuracy is produced with high false positives and false negatives, and that is due to the huge number of features, which could be redundant.
- When the number of coefficients is low, the accuracy is low and the false positives and false negatives are high. That is because not enough information that represents the images is available for the classifier.
- However, the best number of coefficients that supplies the classifier with enough information produces the highest accuracy and low false positive and false negatives.

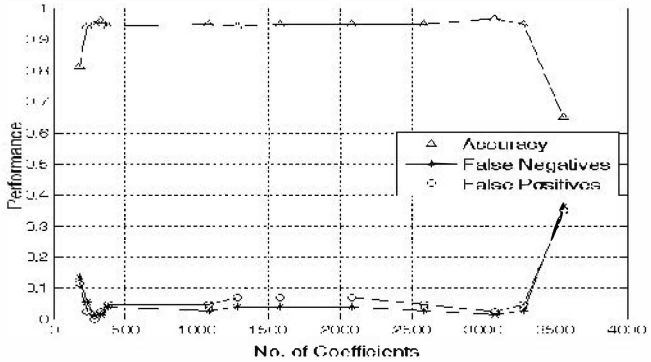


Fig. 6 The affect of number of Coefficients to the performance of the classifier.

This clearly demonstrates the importance of the feature selection method in increasing the accuracy rate of a classifier.

V. Conclusion and Future Work

The paper presented a method for lung cancer diagnosis based on cluster k nearest neighbor algorithm. The results shown in the previous section demonstrates the potential of the method in cancer classification. 96.58 % acuracy has been accomplished so far. Further experiments with more wavelet functions will be carried out to increase the accuracy of the method. In addition, to further improve the results, curvelet transform will be utilized to in the system in future experiments and compared with wavelets.

Acknowledgment

The authors would like to thank the Japanese Society for Radiological Technology for providing the chest radiography images (JSRT Standard Dataset). The authors would also like to thank Univeristi Teknologi PETRONAS for supporting this work.

References

[1] World Health Organization (WHO), “Cancer facts sheet, 2010” http://www.who.int/mediacentre/factsheets/fs297/en/ (Accessed February 2012).

[2] National Cancer Registry, Malaysia, “Malaysian Cancer Statistics – Data and Figure Peninsular Malaysia 2006”. http://www.makna.org.my/PDF/MalaysiaCancerStatistics.pdf (Accessed June 2011).

[3] Jin Mo Goo. A Computer-Aided Diagnosis for Evaluating Lung Nodules on Chest CT: the Current Status and Perspective. Korean J Radiol. 2011 Mar-Apr; 12(2): 145–155.

[4] J. Tang, R. Ranjayyan, I. El Naqa, Y. Yang, Computer aided detection and diagnosis of breast cancer with mammogram: recent advances. IEEE Transaction Information Technology in Biomedicine, 13 (2), 2009, pp. 236-251.

[5] Qiang Li, Recent progress in computer-aided diagnosis of lung nodules on thin-section CT. Computerized Medical Imaging and Graphics 31 (2007); pp. 248–257.

[6] GLOBOCAN, Cancer Fact Sheet 2008, http://globocan.iarc.fr/factsheets/populations/factsheet.asp?uno=900 (Accessed February 2012).

[7] Huang, Y.-L., Computer-aided Diagnosis Using Neural Networks and Support Vector Machines for Breast Ultrasonography. J Med Ultrasound, 2009. 17(1): p. 17–24.

[8] Russell C. Hardie, Steven K. Rogers, Terry Wilson, Adam Rogers; Performance analysis of a new computer aided detection system for identifying lung nodules on chest radiographs. Medical Image Analysis 12 (2008); pp. 240–258.

[9] S.L.A. Lee, A.Z. Kouzani, E.J. Hu, Random forest based lung nodule classification aided by clustering. Computerized Medical Imaging and Graphics 34 (2010); pp. 535–542

[10] Jamshid Dehmeshki, Xujiong Ye, XinYu Lin, Manlio Valdivieso, Hamdan Amin; Automated detection of lung nodules in CT images using shape-based genetic algorithm. Computerized Medical Imaging and Graphics 31 (2007); pp.408–417

[11] João Rodrigo Ferreira da Silva Sousa, Aristófanes Corr˘ea Silva, Anselmo Cardoso de Paiva, Rodolfo Acatauassú Nunes, Methodology for automatic detection of lung nodules in computerized tomography images. Computer Methods and Programs in Biomedicine 98 (2010); pp. 1–14.

[12] Cunjian Chen and Jiashu Zhang, "Wavelet energy entropy as a new feature extractor for face recognition," in Image and Graphics, 2007. ICIG 2007. Fourth International Conference on, 2007, pp. 616-619.

[13] Brahim Belhaouari, samir (2009) Fast and Accuracy Control Chart Pattern Recognition using a New cluster-k-Nearest Neighbor. Journals of Word Academy of Science, Engineering and Technology.

[14] Shiraishi J, Katsuragawa S, Ikezoe J, Matsumoto T, Kobayashi T, Komatsu K, Matsui M, Fujita H, Kodera Y, and Doi K.: Development

of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. AJR 174; 71-74, 2000

[15] Japanese Society of Radiological Technology Lung Dataset. http://www.jsrt.or.jp/web_data/english01.html?category=3 (Accessed: October 2011)