ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Non Invasive Technique for Classification of Pulmonary Cancer Based on Computerized Tomography: Design and Implementation

[1]**Mohammed O. Osman,** MSc, [2]**Alnazier O. Hamza,** PhD, [3]**Zeinab A. M. Mohamed,** PhD, [4]**Mohamed O. Khider,** MSc & [5]**Ali S. A. Altib,** MSc

ABSTRACT

Lung cancer is one of the leading causes of death in USA and Europe. Most cancers that start in the lung, known as primary lung cancers, are carcinomas that derive from epithelial cells. The most common symptoms are coughing (including coughing up blood), weight loss, shortness of breath, and chest pains.Different diagnostic procedure have been followed in attempt to differentiate between the benign and malignant tumor such as Computed Tomography (CT) or Magnetic Resonance Imaging (MRI) images, angiogram, chest X-ray and biopsy which is the most accurate procedure that determines the pulmonary tumor type.Studies indicate that radiologists do not detect all abnormalities on images that are visible on retrospective review, and they do not always correctly characterized abnormalities that are found. In the clinical interpretation of medical images, limitations in the human eye-brain visual system, reader fatigue, distraction, the presence of overlapping structures that camouflage disease in images, and the vast number of abnormal cases seen in screening programs provide cause for detection and interpretation errors.A simple yet effective system for lung nodule detection is developed, the system implementation of feature extraction, normalization, selection, and classification stages were described. The system was verified on real data from a reference dataset and compared to previous systems. The results obtained show a better performance in both the sensitivity and false positive rate. This indicates its potential for clinical use as a first reviewer to boost the diagnostic process.There are several future directions which might further improve the CAD systems for human CT lung images.

Keywords: Lung cancer, malignant lung tumor, computed tomography, Non invasive techniques.

INTRODUCTION

Lung cancer is the leading cause of cancer death among both men and women in the United States[1] and Europe. Lung cancer, also known as carcinoma of the lung or pulmonary carcinoma, is a malignant lung tumor characterized by uncontrolled cell growth in tissues of the lung. If left untreated, this growth can spread beyond the lung by process of metastasis into nearby tissue or other parts of the body. Most cancers that start in the lung, known as primary lung cancers, are carcinomas that derive from epithelial cells.

Cite this article as: Mohammed O. Osman, Alnazier O. Hamza, Zeinab A. M. Mohamed, Mohamed O. Khider & Ali S. A. Altib, " Non Invasive Technique for Classification of Pulmonary Cancer Based on Computerized Tomography: Design and Implementation", International Journal & Magazine of Engineering, Technology, Management and Research (IJMETMR), ISSN 2348-4845, Volume 10 Issue 7, July 2023, Page 1-10.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

The main primary types are small-cell lung carcinoma (SCLC) and non-small-cell lung carcinoma (NSCLC). The most common symptoms are coughing (including coughing up blood), weight loss, shortness of breath, and chest pains[2].

Tobacco smoking is the most important cause of lung cancers with 80%–90% arising in cigarette smokers[3]. About 10–15% of cases occur in people who have never smoked[4]. These cases are often caused by a combination of genetic factors[5] and exposure to radon gas[4], asbestos,or other forms of air pollution[6], including second-hand smoke[7][8]. Lung cancer may be seen on chest radiographs and CT scans. The diagnosis is confirmed by biopsy [9] which is usually performed by bronchoscope or CT-guidance.

Treatment and long-term outcomes depend on the type of cancer, the stage (degree of spread), and the person's overall health, measured by performance status. Common treatments include surgery, chemotherapy, and radiotherapy. NSCLC is sometimes treated with surgery, whereas SCLC usually responds better to chemotherapy and radiotherapy[10]. Overall, 16.8% of people in the United States diagnosed with lung cancer survive five years after the diagnosis[11], while outcomes on average are worse in the developing world. Worldwide, lung cancer is the most common cause of cancer-related death in men and women, and was responsible for 1.6 million deaths annually, as of 2012[12].

Computer Aided Diagnosis (CAD) has been defined as a diagnosis made by a radiologist with the benefit of information generated by computerized image analysis [13]. Although some investigators distinguish between the concepts of computer-aided detection a CAD, others have interpreted CAD broadly as encompassing both the detection task and the classification task [13].

The history of CAD for chest radiography has a long history [14], and was faced badly by the main challenge of extracting the nodules from a chest background, which are formed from a superimposed three-dimensional structure that is projected into a two-dimensional image.

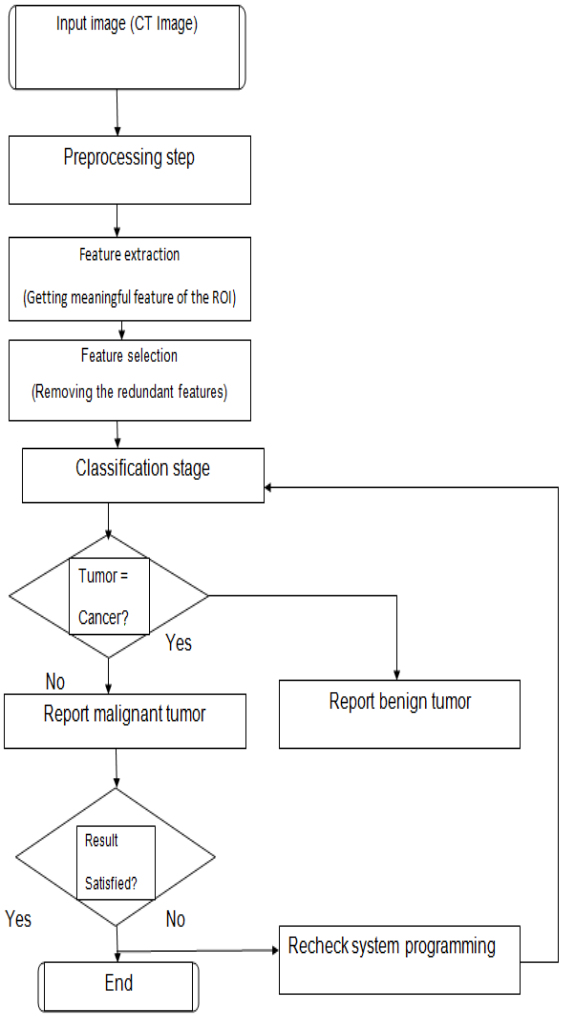
Many researchers have been conducted in the last 6-10 years for developing a CAD system. The approaches followed could be grouped into two main categories, either model based or density based. In model based, nodules were regarded as spherical object that has varying diameters within the known range of 3mm-30mm, and the goal of the research was to find a technique to search for objects having similar shape. The main two strategies followed were by either doing template matching, or by applying a filter that can strengthen the objects that have similar geometric shape as the model the filter was build on. Recently, nodules are regarded more generally as ellipsoid objects rather than spherical. The second approach is the density based, which relies on the fact that nodules tend to have higher density value than the surrounding environment they are in. In this approach, techniques such as multiple thresholding and region growing are usually employed. The most outstanding work

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

done in the last few years was the development of the LIDC database which has led to waves of researches in developing CAD systems for detecting pulmonary nodules. 8 a comprehensive CADe system for classification lung tissues in chest CT scans by using different combinations of 2-D features based on statistical texture features. Min–max normalization method (Rescaling) utilized. Sequential forward selection (SFS) with logistic regression model as evaluation function was used to select the significant feature set. Finally, k-nearest neighbor (KNN) and support vector machine (SVM) classifiers used for classification. The performance of the developed system was verified on LIDC reference database with quantitative measures such as the area under curve (AUC) of the receiver operating characteristics curves (ROC), sensitivity, specificity and accuracy to assess the performance.

Methodology

This section presents the methodology used in this research and the new CAD system for detection and characterization of pulmonary nodules on CT images.The CAD system involve following steps as shown in the Computer Aided Diagnosis system Block Diagram (Figure 1 below).The methods of lung cancer diagnostic involve The first stage of the entire system concerns loading the image, In the second stage, enhanced CT images that contain abnormal masses and normal masses were passed into a segmentation stage where a variable window size for the region of interests (ROI) were taken from the CT images in order to select a dataset of features for the next process which is feature extraction. The resulting features were then used to train a classifier known as K-Nearest Neighbour(KNN) and Support Vector Machine (SVM), in order to discriminate the pattern recognition of the masses into two classes; normal and abnormal. The output can be used for cancer classification.



Input Images:

The database used in this research consisted of 401 cases contains (904 images), randomly obtained from LIDC-IDRI (Lung Image Database Consortium). The LIDC-IDRI database

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

is publicly available in the Cancer Imaging Archive (TCIA), and currently contains 1010 CT scans of the chest, collected in different equipment and different configuration parameters (for example, slice thickness, pixel size and total number of slices). The nodule size ranges from 3 mm to 30 mm and can be primary lung cancer, metastatic disease, benign nodule or indeterminate nature. All nodules were evaluated by four experienced radiologists that through software tools extracted regions of the nodules and described their likelihood of malignancy. The probability was divided into five degree: highly unlikely, moderately unlikely, indeterminate, moderately suspicious and highly suspicious.The nodules obtained in our study were diagnosed by consensus by at least four radiologists,904 nodules were used (including 452 abnormal and 452 normal).

Feature Extraction

Feature extraction in pattern recognition is based on finding mathematical methods for reducing dimensionality of pattern representation.

Extraction procedure consists of three steps:

• Region of interest (ROI) selection where selection of normal and abnormal ROIs in LIDC-IDRI dataset is done manually according to the interpretation of the database and its annotation table.

• Extraction of the selected ROIs with size 32x32.

• Calculation of features from the selected ROIs. First order statistical or histogram-based features, which depend only on pixel values, with seventeen features calculated from extracted ROIs (mean, standard deviation, third-order moment, smoothness, skewness, kurtosis, harmonic mean and variance, percentiles). Other utilized features include the Haralick's texture features, which are are: Energy (EG), Correlation (CO), Inertia (IN),Entropy (EN), Inverse Difference Moment (IDM), Sum Average (SA), SumVariance (SV), Sum Entropy (SE), Difference Average (DA), Difference Variance (DV), Difference Entropy (DE), Information measure of correlation-1(ICO-1) and Information measure of correlation-2(ICO-2).

Feature selection

Feature selection is an important part of any classification scheme. Only a few features may be relevant and hence useful while many may contain irrelevant or redundant information that may result in degradation of classification performance. The success of a classification scheme largely depends on the selected features and the extent of their role in the model.

If the feature changes consistently with pathology, then the hypothesis of a statistically significant, difference between the set of values for normal and abnormal cases will be true, Inferential Statistical tests like T-test should detect a difference.

Classification

To differentiate between nodule and non-nodule regions, a classifier is trained to discriminate between the classes in the feature space based only on the significant selected features. In this work, we used k-nearest neighbor (KNN) and the support vector machine (SVM) classifiers. Both can be used to perform a hard classification, where the output is a class label 0 label for non-nodule and 1 label for nodule.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

• KNN classifier

The KNN classifier is one of the simplest and widely used non-parametric machine learning algorithms. An object is classified based on computing the Euclidean distances from its k-neighbors. The unknown test cases assigned to the closest cluster based on a majority vote among its labeled k-neighbors. In this work, to minimize the errors due to the dimensionality of the feature space and sparsely of samples within that space, we have used nearest neighbor classifier (K=1), (K=3), (K=5), (K=7)

• SVM classifier

A Support Vector Machine is a supervised machine learning algorithm that can be employed for both classification and regression purposes. SVMs are more commonly used in classification problems and as such, this is what we will focus on in this post.SVMs are based on the idea of finding a hyper plane that best divides a dataset into two classes. Support vectors are the data points nearest to the hyperplane, the points of a data set that, if removed, would alter the position of the dividing hyperplane. Because of this, they can be considered the critical elements of a data set.

In SVM the original input space is mapped into a high dimensional dot product space called feature space. In this space the optimal classification hyperplane is determined to maximize the generalization ability of the classifier [8].

Cross Validation

Cross validation, sometimes called rotation estimation, is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. In a prediction problem, a model is usually given a dataset of known data on which training is run (training dataset), and a dataset of unknown data (or first seen data) against which the model is tested (testing dataset)[4].The goal of cross validation is to define a dataset to "test" the model in the training phase (i.e., the validation dataset), in order to limit problems like over fitting give an insight on how the model will generalize to an independent dataset (i.e., an unknown dataset, for instance from a real problem), etc. One round of crossvalidation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (called the training set), and validating the analysis on the other subset (called the validation set or testing set). To reduce variability, multiple rounds of cross-validation are performed using different partitions and the validation results are combined (e.g. averaged) over the rounds to estimate a final predictive model [10].

Performance measures

The performance measures include four tests:

• Sensitivity (SE): The sensitivity of a test is the ability of the test to identify correctly the affected individuals Proportion of persons testing positive among affected individuals.

• Specificity (SP): The specificity of a test is the ability of the test to identify correctly non-

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

affected individuals Proportion of persons testing negative among non-affected individuals.

• Positive Predictive Value (PPV): Is the proportion of abnormal cases correctly identified.

• Negative Predictive Value (NPV): Is the proportion of benign cases correctly identified.

Results

As stated earlier the total number of ROIs used from the LIDC-IDRI database was 904 (including 452 Non-nodules and 452 nodules). The training and testing of KNN and SVM classifiers was based on independent half-and-half training and testing.The first stage of the entire system concerns loading the image, In the second stage, CT images were passed into a segmentation stage where a window for the region of interest (ROI) was taken from the CT images in order to build a dataset for the process of feature extraction (figure 2 shows an example of the ROI). The resulting features were then used to classifier known as of KNN and SVM in order to discriminate the pattern recognition of the masses into two classes; normal and abnormal. With the lung cancer system fully developed using MATLAB, the means to evaluate the efficacy of the methods and algorithms used, is crucial. A total of 904 ROIs extracted from CT lung images have been compiled for this project.

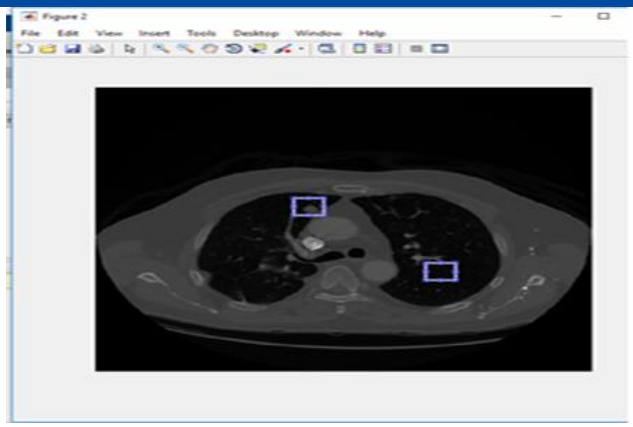


Figure 2:Region of Interest (ROI)

Feature extraction

Based on adjusted and segmented images produced frompreviousstages, a set of features are extracted from each image. We have then applied our reduction algorithm on these features. Before the classification model can be built, meaningful features of the ROIs delineated during the process of segmentation, need to be extracted and used as input in the classification process. TheHaralick features were used in our developed technique for feature extraction.

Feature Selection

Normalization and feature selection boost the performance of the classification process. Normalization will avoid numerical instabilities in the process of training the classifiers and allow variations in different features to be well represented equally with no dominating features that happen to have wider numeric ranges. Feature selection reduces the time of classification and will choose the optimal set that yield the best performance.Feature selection chooses optimal set of features, Number of the selected features depends on the datasets used. The training and testing method (half-and-half or

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

cross validation) is chosen according to the available number of samples. If we have large number of samples, it is preferred to use half-and-half.

Classification

The final procedure of the proposed system is to confirm the suspicious region and determine if it is a true nodule utilizing features obtained from previous stages.In this work, we used-nearest neighbour (KNN) and the support vector machine (SVM) classifiers. Both can be used to perform ahard classification.

The KNN classifier is one of the simplest and widely used non-parametric machine learning algorithms. An object is classified based on computing the Euclidean distances from its neighbours. The unknown test cases assigned to the closest cluster based on a majority vote among is labelled neighbours. In this work, to minimize the errors dueto the dimensionality of the feature space and sparsity of samples within that space, we have used nearest neighbour classifier (K=1, 3, 5, 7)

The SVM classifier is a supervised learning model. It has been applied to classification and regression problems with exceptionally good performance on a range of binary classification tasks. In SVM the original input space is mapped into a high dimensional dot product space called feature space. In this space, theoptimal classification hyper-plane is determined to maximize he generalization ability of the classifier.

Table 1 below presents the set of parameters associated with the used classifiers. Half-and-Half method used for training and testing of KNN and SVMclassifiers.

Table 1: parameters of used classifiers

Classifier	**Parameters**
KNN	K=(1,3,5,7) Euclidean distance
SVM	Linear kernel function

Classification results

Table 2 here under demonstrates the results of the training and testing phases utilizing KNN and SVM classifiers.Figures 3 and 4 shows the Region of Convergence (ROC) curves for the KNN classifiers (K=5 and K=7 respectively).

Table 2: Classification Results (Sensitivity and Specificity)

	Train			Test		
Classifier	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
KNN Classifier (K=1)	100%	100%	100%	84%	78%	81%
KNN Classifier (K=3)	100%	100%	100%	83%	91%	87%
KNN Classifier (K=5)	100%	100%	100%	83%	91%	87%
KNN Classifier (K=7)	100%	100%	100%	**84%**	**92%**	**88%**
KNN Classifier (K=9)	100%	100%	100%	82%	90%	86%
SVM Classifier (Linear)	87%	79%	83%	86%	87%	86.5%
SVM Classifier (Gaussian)	95%	96%	95.5%	85%	82%	83.5%

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

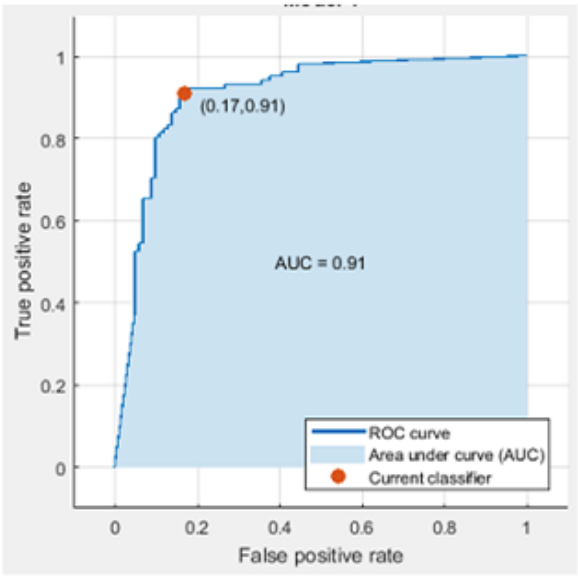


Figure 3: ROC curve for KNN classifier (K=5)

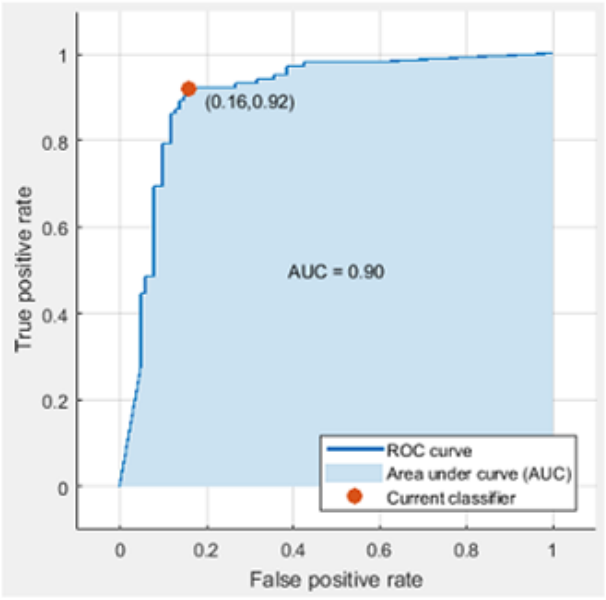


Figure 4:ROC curve for KNN classifier (K=7)

Conclusion

A simple yet effective system for lung nodule detection is developed, The system implementation of feature extraction, normalization, selection, and classification stages were described.The system was verified on real data from a reference dataset and compared to previous systems. The results obtained show a better performance in both the sensitivity and false positive rate.This indicates its potential for clinical use as a first reviewer to boost the diagnostic process.

There are several future directions which might further improve the CAD systems for human CT lung images, some of them are:

- The acquisition of large databases from different institutions with various image qualities for clinical evaluation and improvement in the CAD systems
- Improve the classification accuracy by extracting more efficient features and increasing the training data set.
- Further experiments and evaluation are therefore desirable to establish whether the proposed approaches have generic applications.

References

[1] Cao W, Chen HD, Yu YW, Li N, Chen WQ. Changing profiles of cancer burden worldwide and in China: a secondary analysis of the global cancer statistics 2020. Chinese Medical Journal. 2021 Apr 5; 134(07):783-91.

[2] Jameson J, Fauci AS, Kasper DL, Hauser SL, Longo DL, Loscalzo J. Harrison's Principles of Internal Medicine. 18th ed., McGraw Hill Professional; 2012. Chapter 89. ISBN: 0071748903, 9780071748902

[3] Minna JD, Roth JA, Gazdar AF. Focus on lung cancer. Cancer cell. 2002 Feb 1; 1(1):49-52.

[4] Thun MJ, Hannan LM, Adams-Campbell LL, Boffetta P, Buring JE, Feskanich D, Flanders WD, Jee SH, Katanoda K, Kolonel LN, Lee IM, Marugame T, Palmer JR, Riboli E, Sobue T,

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Avila-Tang E, Wilkens LR, Samet JM. Lung cancer occurrence in never-smokers: an analysis of 13 cohorts and 22 cancer registry studies. PLoS Med. 2008 Sep 30; 5(9):e185. doi: 10.1371/journal.pmed.0050185. Epub 2008 Sep 9. PMID: 18788891; PMCID: PMC2531137

[5] Alberg AJ, Samet JM. Murray &Nadel's Textbook of Respiratory Medicine. 5th ed. Saunders Elsevier; 2010. Chapter 46. ISBN 978-1-4160-4710-0.JM

[6] O'Reilly KM, Mclaughlin AM, Beckett WS and Sime PJ.Asbestos-related lung disease. American Family Physician.March 2007; 75 (5): 683–688. PMID 17375514.

[7] Carmona.The Health Consequences of Involuntary Exposure to Tobacco Smoke: A Report of the Surgeon General. U.S. Department of Health and Human Services.2006.

[8] Ong EK, Glantz SA. Tobacco industry efforts subverting International Agency for Research on Cancer's second-hand smoke study. The Lancet. 2000; 355(9211):1253-1259.

[9] Lu C, Onn A, Vaporciyan AA et al. Holland-Frei Cancer Medicine. 8th ed. People's Medical Publishing House.2010. Chapter 78. ISBN 978-1-60795-014-1

[10] Chapman, S; Robinson G; Straddling J; West S. Oxford Handbook of Respiratory Medicine. 2nd Ed. Oxford University Press. 2009. Chapter 31. ISBN 978-0-19-954516-2.

[11] Mustafa M, Azizi AJ, IIIzam E, Nazirah A, Sharifa S, Abbas S. Lung cancer: risk factors, management, and prognosis. IOSR Journal of Dental and Medical Sciences. 2016; 15(10):94-101.

[12] Pinsky PF. Lung cancer screening with low-dose CT: a world-wide view. Translational lung cancer research. 2018; 7(3):234.

[13] G. Samuel, III. Armato, M. Geoffrey, F. Michael, R. Charles, Y. David,Lung image database consortium—developing a resource for the medical imaging research community, Radiology 2004; 24: 739–748.

[14] Manfred Tillich, Detection of Pulmonary Nodules with Helical CT: Comparison of Cine and Film-Based Viewing. AJR 2015; 169:1611-1614.

Author Details:

Mohamed Omer Osman, MSc, is a Part time lecturer at the University of Medical Sciences and Technology, Faculty of Engineering, Department of Biomedical Engineering, Khartoum, Sudan. He has a Second-class,Division One Bachelor of Science and M.Sc. degrees in biomedical engineering from the Sudan University of Science and Technology, Khartoum, Sudan. Currently he is a Ph.D. Student of Biomedical Engineering at Sudan Academy for Science, Khartoum, Sudan.
Email: mohammedomer121@gmail.com

Alnazier Osman Hamza, PhD,is a professor of biomedical engineering at the University of Medical Sciences & Technology, he received his BS degree in Medical Imaging from Sudan University of Science and Technology, his MS in Medical Physics and Biomedical Engineering from Surry University, UK, and PhD in Medical Physics and Biomedical Engineering from University of Natal, South Africa.Email: alnazier_67@hotmail.com.

Zeinab Adam Mustafa Mohamed, PhD, is a professor of biomedical engineering at the Department of Biomedical Engineering, College of Engineering, Sudan University of Sciences and Technology. Email: zenab42000@yahoo.com

Mohamed Omer Khider, MSc, is a Part time lecturer at the University of Medical Sciences and Technology, Faculty of Engineering, Department of Biomedical Engineering, Khartoum, Sudan. He has a first-class Bachelor of Science and M.Sc. degrees in

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

biomedical engineering from the University of Medical Sciences and Technology, Khartoum, Sudan. Currently he is a Ph.D. Student of Biomedical Engineering at the University of Medical Sciences and Technology, Khartoum, Sudan. Email: birreema25@hotmail.com.

Ali Salah Aldin Ali, MSc, is a Part time lecturer at the Sudan International University, Faculty of Engineering, Department of Biomedical Engineering, Khartoum, Sudan. He has a second -class Bachelor of Science from Almughtrabeen University and M.Sc. degree in Biomedical Engineering from the University of Medical Sciences and Technology, Khartoum, Sudan. Email: alisalahaltib@gmail.com