Computer Aided Detection of Large Lung Nodules using Chest Computer Tomography Images

Mai Mabrouk
Faculty of Engineering, MUST University, 6th of October, Egypt

Ayat Karrar
Faculty of Engineering, Cairo University, Giza, Egypt

Amr Sharawy
Faculty of Engineering, Cairo University, Giza, Egypt

ABSTRACT

Lung cancer is the most common cancer which leads to death for both women and men, so the early detection of lung cancer increases the therapy success. Different techniques are used to provide the early detection such as Computer Aided Detection (CAD) system. In this paper, we present an automatic Computer Aided Detection (CAD) system to detect a large lung nodule from lateral Chest Radiographs of computed tomography (CT) images to reduce false positive rates. Basic image processing techniques such as Bit-Plane Slicing, Erosion, Median Filter, Dilation, Outlining, radon transform and edge detection are applied to the CT scan images in order to detect the lung region. A total of 22 image features were extracted from the enhanced image based on statistical features such as standard deviation, average and mean. A fisher score ranking method is used as a feature selection method to select best ten features (standard deviation, variance, range, maximum grey level, seven invariant moments except the second, sixth and seventh invariant moments and 5th percentile, 9th percentile). Thus optimal screening modalities have both high sensitivity and specificity based on artificial neural network (ANN) significantly more accurate than using K-Nearest Neighborhood (KNN) classifier with accuracy 98% and 96% respectively in detecting large lung nodule with equivalent diameter ranging from 22.65 mm to 41.62 mm.

Keywords

Computer Aided Diagnosis (CAD), Computed tomography (CT), Radon transform, Artificial Neural Network (ANN), K-Nearest Neighborhood (KNN).

1. INTRODUCTION

Lung cancer is a disease that consists of uncontrolled cell growth in tissues of the lung which may lead to metastasis that is the infestation of adjacent tissue and infiltration beyond the lungs. Carcinomas are the vast majority of primary lung cancers which derived from epithelial cells. Lung cancer, the most common cause of cancer-related death in men and women, is responsible for 1.3 million deaths worldwide annually, as of 2004 [1].

Lung cancer can be seen on traditional x-ray and computed tomography (CT scan). The diagnosis is confirmed with a biopsy. Bronchoscopy or CT-guided biopsies are usually used in sample extraction. Treatment and prognosis depend on the histological type of cancer, the stage (degree of spread), and the patient's performance status, but overall only 14% of people diagnosed with lung cancer survive five years after the diagnosis[2].

Recently, Computed Tomography (CT) is 10-20 times more sensitive than standard x-ray techniques, so that it is the most effective for early detection of lung cancer. Computer Aided Diagnosis (CAD) system is one of the most effective applications used in detection of lung nodules. With CAD, radiologists use the computer output as a "second opinion" and make the final decisions. Methodological research currently focuses on segmentation and feature extraction, a complete structure of a CAD system is shown in Fig 1. The CAD software technology meets three main objectives:

- Improve the quality of diagnosis (second opinion).
- Increase therapy success by early detection of cancer.
- Avoid unnecessary biopsies whereas, a lung biopsy is a dangerous procedure, with a 2% risk of serious complications (including death) [3].

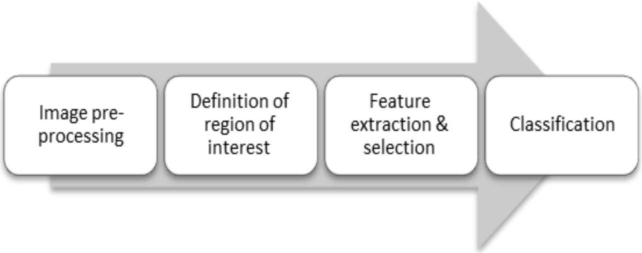


Image pre-processing
Definition of region of interest
Feature extraction & selection
Classification

Fig1. A complete structure of a CAD system.

In this paper, we also propose an algorithm approves that Artificial Neural Network (ANN) is significantly more accurate in early detection of large lung nodules than K-Nearest Neighborhood (KNN) with sensitivity 97% and 97% respectively, specificity 98% and 96% respectively and accuracy 98% and 96% respectively.

2. RELATED WORK

Rachid et al [4] uses pure basic image processing techniques lung regions extraction Hopfield artificial neural network (HANN) that shows a good Segmentation results in a short time. M. Gomathi et al [5] applied a Modified Fuzzy Possibilistic C-Mean algorithm to do the segmentation step. For classification, Support Vector Machine (SVM) and Extreme Learning Machine (ELM) are used. The experimental result shows that the accuracy of using ELM is better when compared to the usage of SVM for classification.

Volkan et al [6] developed two algorithms to classify all the samples in a batch jointly, one based on a probabilistic analysis and another based on a mathematical programming approach. Experiments on three real-life computer aided diagnosis (CAD) problems demonstrate that the proposed algorithms are significantly more accurate than a naive SVM which ignores the correlations among the samples.

Nancy et al [7] employs a classification algorithm for detecting Solid pulmonary nodules from CT thorax studies and describes some of the machine learning techniques. Korfiatis et al [8] identifies lung boundary by an automatic thresholding approach, false positive (FP) regions were subsequently removed using a Support Vector Machine (SVM) classifier employing morphological features extracted from corresponding nodule candidate regions of the enhanced and the original images.

Alessandro et al [9] introduces an approach based on: (1) a lung tissue segmentation pre-processing step, composed of histogram thresholding, seeded region growing and mathematical morphology; (2) a filtering step, whose aim is the preliminary detection of candidate nodules (via 3D fast radial filtering) (3) a false positive reduction (FPR) step, applies thresholds on region diameters, and a supervised FPR, which is based on support vector machines classification

Thangaraj et al [10] explains an approach for lung nodule detection in screening CT. A dot-enhancement filter is utilized effectively for the selection of nodule candidate. In addition a neural classifier is employed to reduce the false-positive rates. The experiments are conducted on real time collected data set (CT scan) to prove the efficiency of his approach.

3. MATERIAL AND METHODS

3.1 Dataset

The 12 digital chest radiographs used in this study consisting of 2911 2D CT images collected with approval from Cornell University [11] with equivalent diameters of lung nodules ranging from 22.65 mm to 41.62 mm. The in - slice (x, y) resolution is 0.79 × 0.79 mm and CT slice thickness is 1.25 mm.

3.2 Lung region extraction

A CT image of chest consists of different regions such as the background, lung, heart, liver and other organs' areas. The goal of lung region extraction step is to separate the lung regions, our regions of interest (ROIs), from the surrounding anatomy to avoid confusion. Our algorithm to extract lung region is shown in Fig 3. This algorithm begins with the bit – plane slicing [22] [4]. In terms of bit-plane extraction for a 8-bit image, it is seen that binary image for bit plane 7 is obtained by proceeding the input image with a thresholding gray-level transformation function that maps all levels between 0 and 127 to one level (e.g. 0) and maps all levels from 129 to 253 to another (e.g. 255). These binary images then are enhanced by using Erosion, Median filter and Dilation to eliminate irrelevant details that may causes some difficulties in lung region extraction step. The fast method for lung separation is used as in [21]; this process starts by estimating the mass center of each lung, as shown in Fig 2. It consisted on applying the Radon transform to the binary image for both horizontal and vertical directions.

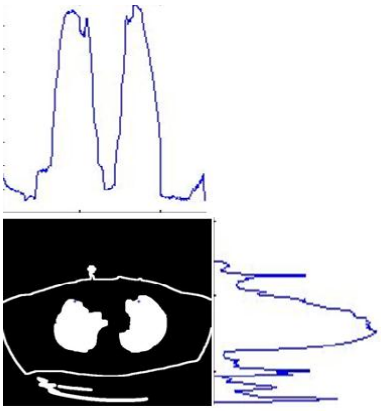


Fig2. Radon transform.

3.3 Feature extraction

Feature extraction is a step used to reduce the original dataset by measuring certain properties, or features, that distinguish one input pattern from another. The extracted features provide the characteristics of the input type to the classifier by considering the description of the properties that relate to the image into a feature space; Fig 5 shows applying the feature extraction and selection step and then input it to classifiers. Tumor is a heterogeneous tissue and the mean values of relaxation times are not at all sufficient to characterize the heterogeneity of the different tumor types [12].

In this study, 22 statistical features are used. These features are: mean [24], standard deviation, variance, skewness, kurtosis [23] [13], entropy [14] [15], nine percentile features were used ranging from the first percentile up to the ninth percentile [16] [13], seven invariant moments [13], maximum of gray level [13], and range of gray level [13]. Ten of these features only are the most significant features to be considered in the classification step. These features are (Standard Deviation, Variance, Maximum grey level, Range of gray level , all seven invariant moments except the second, sixth and seventh invariant moments and 5th percentile, 9th percentile) and they are calculated as follows:

1. Standard Deviation

It is a statistical measure of spread or variability. The standard deviation is the root mean square (RMS) deviation of the values from their mean.

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (1)$$

Where n is the number of pixels in the image and $\bar{x}$ is the mean of x

2. Variance

It is the Square of the standard deviation.

$$v = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad (2)$$

3. Maximum grey level.

4. Range

It is the difference between the highest and lowest values in the sample.

$$R = g_{\max} - g_{\min} \qquad (3)$$

Where g is the grey level of the pixels.

5. Seven invariant moments

It is a shape description technique; Moments can provide characteristics of an object which singularly represent its shape. Classification in the multidimensional moment invariant feature space performed invariant shape recognition. From the second and third order values of the normalized central moments a set of seven invariant moments can be computed which are independent of rotation. These features are calculated as:

$$I_1 = \eta_{20} + \eta_{02} \qquad (4)$$

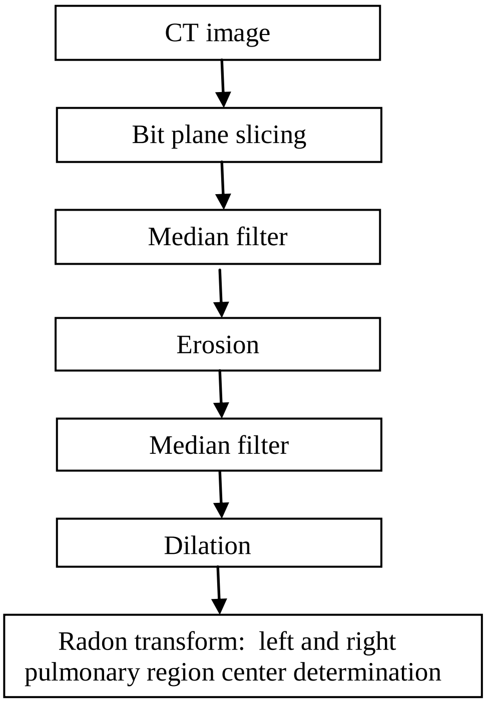
$$I_2 = (\eta_{20} - \eta_{02})^2 + (2\eta_{11})^2 \qquad (5)$$

$$I_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \qquad (6)$$

$$I_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \qquad (7)$$

$$\begin{aligned} I_5 = {} & (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})\left[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2\right] \\ & + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})\left[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right] \end{aligned} \qquad (8)$$

$$\begin{aligned} I_6 = {} & (\eta_{20} - \eta_{02})\left[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right] \\ & + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \end{aligned} \qquad (9)$$


CT image
Bit plane slicing
Median filter
Erosion
Median filter
Dilation
Radon transform: left and right pulmonary region center determination

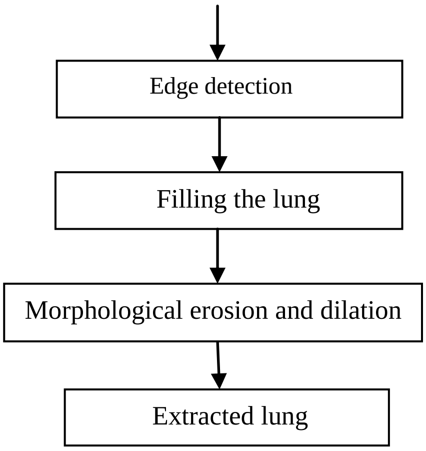


Fig 4. Extraction lung region.

$$\begin{aligned} I_7 = {} & (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})\left[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2\right] - \\ & (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})\left[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right] \end{aligned} \qquad (10)$$

Where $\eta_{ij} = \dfrac{\mu_{ij}}{\mu_{00}^{(1+\frac{i+j}{2})}}$ (11)

And $\mu_{pq} = \sum_x \sum_y (x - \bar{x})^p (y - \bar{y})^q f(x, y)$ (12)

Where $\bar{x} = \dfrac{m_{10}}{m_{00}}$ (13)

And $\bar{y} = \dfrac{m_{01}}{m_{00}}$ (14)

Where $m_{ij} = \sum_x \sum_y x^p y^q I(x, y)$ (15)

Where *I*(*x*, *y*) are the pixel intensities

6. Percentile

The pth percentile of a list is the number such that p percent of the elements in the list are less than that number

$$P_{n=\frac{100}{N}}\left(n - \frac{1}{2}\right) \qquad (17)$$

Where N is the number of elements in the sample, n is the rank of the percentile. Nine percentile features were used ranking from (10, 20 ….90). In this work we reduce the number of features significantly by using feature Selection technique, while increasing the detection accuracy [17].

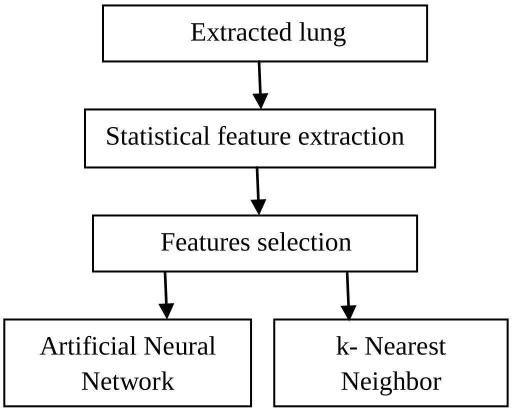


Fig.5 Feature extraction and classification

3.4 Feature selection

Feature selection is a problem that has to be addressed in many areas, especially in artificial intelligence. The main issues in developing feature selection techniques are choosing a small feature set in order to reduce the cost and running time of a given system, as well as achieving an acceptably high recognition rate. This has led to the development of a variety of techniques for selecting an optimal subset of features from a larger set of possible features; Fig 5 shows applying the feature extraction and selection step and then input it to classifiers. The Fisher score ranking technique is used to select the best ten features that give best results by calculating the difference [17], which is described in terms of mean and standard deviation, between the positive (abnormal) and negative (normal) examples relative to a certain feature. Equation (17) defines the Fisher score, in which R_i is the rank of feature *i,* describing the proportion of the substitution of the mean of the feature *i* values in the positive examples (*p*) and the negative examples (*n*), and the sum of the standard deviation. The bigger the R_i, the bigger the difference between the values of positive and negative examples relative to feature *i* ,so we can differentiate between the normal and abnormal tissues ; thus, this feature is more important for separating the positive and negative examples.

$$\boldsymbol{R_i} = \frac{\mu_{i,p} - \mu_{i,n}}{\sigma_{i,p} + \sigma_{i,n}} \qquad (17)$$

3.5 Classification

The classification step is the final step in our model where both the features of the training images and the test images are the input of the classifier, while the output is the image type. In our model we used two classifiers; Artificial Neural Network (ANN) and K- Nearest Neighborhood (KNN).

1. K-Nearest Neighbor (k-NN) Classifier.

A simple classifier is a Nearest Neighbor (NN) classifier [26], where each pixel, is classified in the same class as the training data with the closest intensity. There is a possibility of an NN classifier yielding an erroneous decision if it is obtained single neighbor is an outlier of some other class. To avoid this and improve the robustness of the approach, the k-NN classifier works with k patterns in the neighborhood of the test pattern. The k-NN classifier is considered a non-parametric classifier since it makes no underlying assumption about the statistical structure of the data [12].

2. Artificial Neural Network (ANN).

The Multilayer Perception is a feed forward network, capable of generating nonlinear boundaries. It has been successfully applied here to solve some difficult and diverse problems [18]. It consists of ten input nodes, two hidden layers one with three nodes and one with four nodes, and one output node after exploring the performance of various configurations[19] [25].

3.6 Performance measures

All classification result could have an error rate and on occasion will either fail to identify an abnormality, or identify an abnormality which is not present. It is common to describe this error rate by the terms true and false positive and true and false negative as follows: *True Positive (TP):* the classification result is positive in the presence of the clinical abnormality. *True Negative (TN):* the classification result is negative in the absence of the clinical abnormality. *False Positive (FP)*: the classification result is positive in the absence of the clinical abnormality [12]. *False Negative (FN):* the classification result is negative in the presence of the clinical abnormality. Sensitivity, specificity and accuracy are used to measure the classifiers performance [27] [12]. To provide clinically pertinent definitions for *sensitivity* and *specificity,* we follow the definition in [20] and point out how these performance measures should be interpreted for cancer screening. *Sensitivity* and *specificity* measure the number of false positives and false negatives and are useful in measuring the effectiveness of screening methods. Sensitivity and specificity are defined as follows; the sensitivity of a screening test is its ability to detect those individuals with cancer. It is computed by taking the number of true positives (TPs) and dividing it by the total number of cancer cases (TP FN). The specificity of a test is its ability to identify those individuals who actually do not have cancer. It is computed by dividing the true negative (TN) by the sum of the TN and FP cases as shown in next equations:

$$\text{Sensitivity} = TP/\ (TP+FN)\ *100\% \qquad (18)$$

$$\text{Specificity} = TN/\ (TN+FP)\ *100\% \qquad (19)$$

$$\text{Accuracy} = (TP+TN)/\ (TP+TN+FP+FN)*100\ \% \qquad (20)$$

4. RESULTS

In this paper, we have described a CAD scheme for the automated detection of lung nodules on thin slice CT. Fig. 4 shows the results of applying step by step the proposed lung regions extraction method to a given CT image. The system is based on an automated detection approach utilizing a classifier rather than a rule based scheme. Automated delineation of lung boundaries is considered another advantage of the proposed system. The results of performance measures are shown in table 1. Table 2 shows sensitivity, specificity and accuracy of the classifiers that are used to evaluate the performance of the classifier. Results provided that the accuracy of ANN classifier is better than KNN classifier.

Table1. Results of performance measures.

Error rate	Artificial neural network	k- nearest neighborhood
TP	100	150
TN	190	276
FP	2	12
FN	4	6

Table2. Results of classification accuracy.

Classifier	Sensitivity	Specificity	Accuracy
Artificial neural network	97%	98%	98%
k- nearest neighborhood	97%	96%	96%

5. CONCLUSIONS

In this paper, the computer based technique for automatic classification of CT slices as normal or abnormal with various CT image features using two classifiers is proposed. In the first phase of the proposed technique, the lung region is extracted from the chest tomography image. The different basic image processing techniques are used for this purpose. The inputs of classifiers were CT images after extraction and selection of 22 statistical features, ten features of the twenty two achieved best results were selected by using feature selection method as fisher score ranking method which input to the Artificial Neural Network (ANN) and k- Nearest Neighbor (KNN) classifiers. The performances of the classifiers in terms of statistical measures such as sensitivity, specificity and classification accuracy are analyzed. The results indicated that the ANN approach yielded the better performance when compared to the KNN classifier.

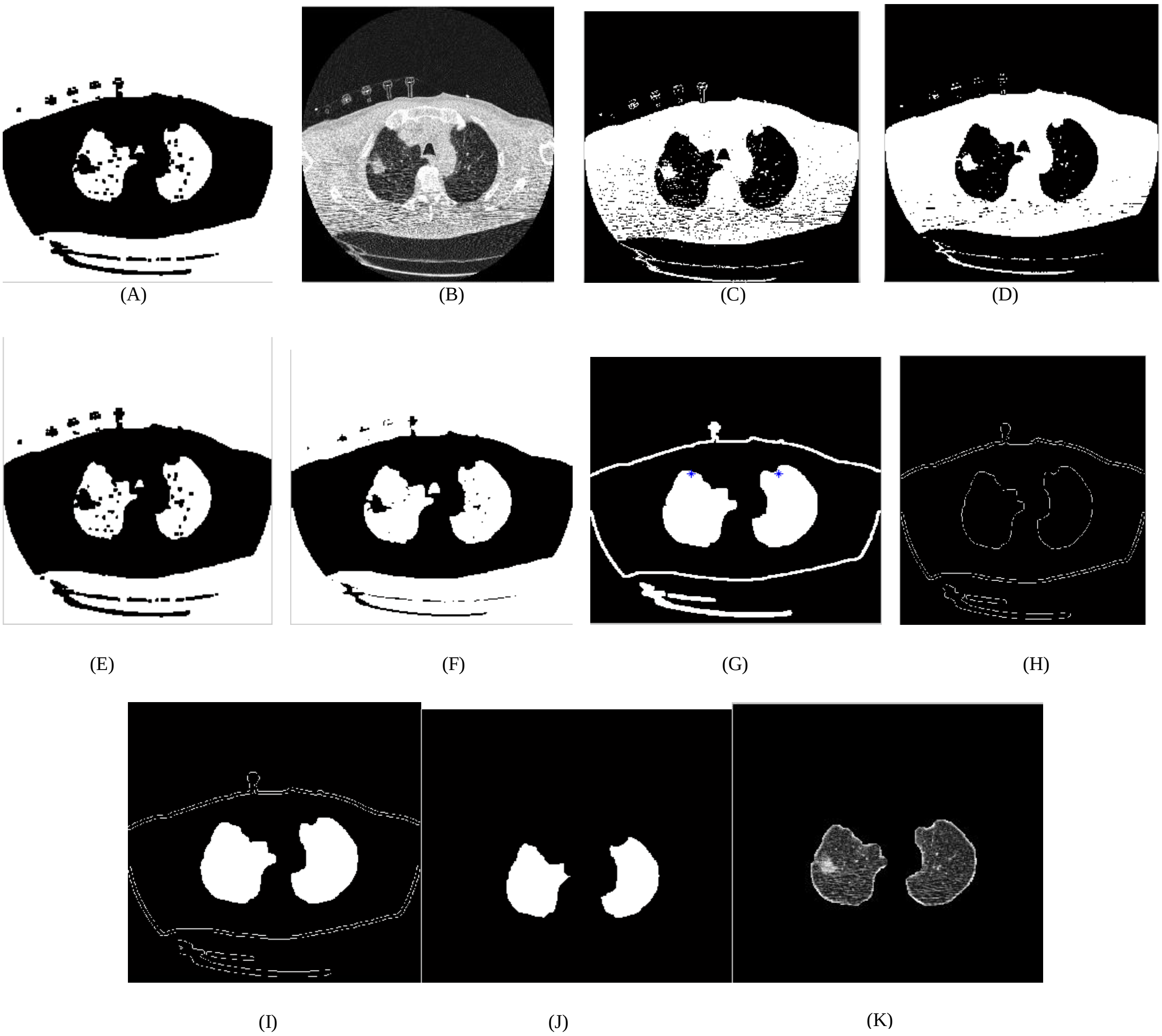


Fig4: Lung regions extraction algorithm: A. original CT image, B. bit-plane- slicing, C. median filter , D.erosion, E. median filter, F. dilation, G. Radon transform., H. Edge detection , I. Filling the lung, J.morphological operators and K. Extracted lung.

6. REFERENCES

[1] World Health Organization (WHO), February 2006.

[2] Minna. JD and Schiller JH," Harrison's Principles of Internal Medicine (17th ed.)", McGraw-Hill, pp. 551–562, 2008.

[3] Http//www3.cancer.gov/bip/lidc_comm.htm.

[4] Rachid Sammouda1. Jamal Abu Hassan1,.Mohamed Sammouda2, Abdulridha Al-Zuhairy 3 and Hatem abou ElAbbas," Computer Aided Diagnosis System for Early Detection of Lung Cancer Using Chest Computer Tomography Images", GVIP 05 Conference, CICC, Cairo, Egypt, 19-21 December 2005.

[5] M. Gomathi and P. Thangaraj, "A Computer Aided Diagnosis System for Lung Cancer Detection using Machine Learning Technique", European Journal of Scientific Research, Vol.51 No.2, pp.260-275, 2011.

[6] Volkan Vural1. Glenn Fung2. Balaji Krishnapuram2. Jennifer Dy1 and Bharat Rao2," Batch Classification with Applications in Computer Aided Diagnosis", European Conference on Machine Learning (ECML), vol. 4212, p. 449-460, 2006.

[7] R Bharat Rao. Jinbo Bi. Glenn Fung, Marcos Salganicoff. Nancy Obuchowski and David Naidich, "LungCAD: A Clinically Approved, Machine Learning System for Lung Cancer Detection ", Knowledge Discovery and data mining conference (KDD), p. 1033-1037, August 12-15, 2007.

[8] P. Korfiatis. C. Kalogeropoulou. L. Costaridou," Computer Aided Detection of Lung Nodules in Multislice Computed Tomography", the International Special Topic Conference on Information Technology in Biomedicine IEEEITAB, 2006.

[9] Alessandro Riccardi. Todor Sergueev Petkova. Gianluca Ferri. Matteo Masotti and Renato Campanini," Computer-Aided Detection of Lung Nodules via 3D Fast Radial Transform, Scale Space Representation, and Zernike MIP Classification ",the international journal of medical physics research & practice,vol.38, 1962 ,2010.

[10] M. Gomathi and Dr. P. Thangaraj," Lung Nodule Detection using a Neural Classifier", IACSIT International Journal of Engineering and Technology, Vol.2, No.3, June 2010.

[11] (2002) Cornell university website. [Online]. Available: http://www.via.cornell.edu/databases.

[12] H. Selvaraj1, S. Thamarai Selvi2, D. Selvathi3 and L. Gewali1,"Brain MRI Slices Classification Using Least Squares Support Vector Machine", Vol. 1, No. 1, Issue 1, P. 21 - 33, 2007.

[13] Mohamed A. Alolfe, Abo-Bakr M. Youssef, Yasser M. Kadah, and A. S. Mohamed "Development of a Computer-Aided Diagnostic System for Cancer Detection from Digital Mammograms", 25th National Radio Science Conference, March 18-20, 2008.

[14] Yasser M. Kadah, Aly A farag, Ahmed M. badawy and Abou-Baker M. Youssef, "Classification algorithm for quantitative tissue characterization of diffuse liver disease from ultrasound," IEEE Trans. Med. Imag., vol. 15, pp.466-478, August 1996.

[15] Rafael Gonzalez and Richard woods," Digital Images Processing". Prentice Hall, 2002.

[16] CHAP T. LE, Introductory Biostatistics. A John Wiley & Sons Publication, April 2003.

[17] Dima Stopel, Zvi Boger, Robert Moskovitch, Yuval Shahar, and Yuval Elovici" Improving Worm Detection with Artificial Neural Networks through Feature Selection and Temporal Analysis ", International Journal of Applied Mathematics and Computer Sciences, Vol.1, Number1, August 2006.

[18] J. A. Freeman and D. M. Skapura, “Neural Networks, Algorithms, Applications and Programming Techniques”, Addison- Wesley Publishing Company, (2002).

[19]S. Haykin, “Neural networks: A comprehensive Foundation”, 2nd ed. Englewood Cliffs, NJ: Prentice Hall, 1999.

[20] D. G. Altman and J. M. Bland, “Diagnostic tests 1: Sensitivity and specificity,” Br. Med. J., vol. 308, pp. 1552–1552, 1994.

[21] José Silvestre Silva, Augusto Silva and Beatriz Sousa Santos," Lung Segmentation Methods in X-ray CT Images", 5th Iberoamerican Symposium on Pattern Recognition, vol.15,pp. 583-598, 2000.

[22] M. Gomathi and P. Thangaraj," A Computer Aided Diagnosis System for Detection of Lung Cancer Nodules Using Extreme Learning Machine", International Journal of Engineering Science and Technology, Vol. 2(10), 2010.

[23] Ted W. Way, Berkman Sahiner, Heang-Ping Chan, Lubomir Hadjiiski, Philip N. Cascade, Aamer Chughtai, Naama Bogot, and Ella Kazerooni, "Computer-aided diagnosis of pulmonary nodules on CT scans: Improvement of classification performance with nodule surface features", the international journal of medical physics research & practice, vol. 36, 2009 Jul;36(7):3086-98.

[24] Guo Xiuhua, Sun Tao, Wu Haifeng, He Wen, Liang Zhigang,Zhang Mengxia,Guo Aimin1 and Wang Wei1," Support Vector Machine Prediction Model of Early-stage Lung Cancer Based on Curvelet. Transform to Extract Texture Features of CT Image", World Academy of Science, Engineering and Technology 71, vol 17, 2010

[25] Hui Chen, Wenfang Wu, Hong Xia, Jing Du, Miao Yang and Binrong Ma, "Classification on pulmonary nodules using neural network ensemble", lecture notes in computer science, vol. 6677, 2011.

[26] Ayman El-Baz, Matthew Nitzken, Fahmi Khalifa, Ahmed Elnakib, Georgy Gimel’farb, Robert Falk and Mohammed Abo El-Ghar, " 3D Shape Analysis for Early Diagnosis of Malignant Lung Nodules ", Lecture Notes in Computer Science, Vol. 6801, 2011.

[27] R.Nithya, B.Santhi, “Mammogram Classification Using Maximum Difference Feature Selection Method", Journal of Theoretical and Applied Information Technology, Vol. 33, pp 197 - 204, 2011.