ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 2, Issue 3, September 2012

Lung Cancer Classification Using Image Processing

Dr. S.A.PATIL, M. B. Kuchanur

Abstract*: Lung cancer is one of the most common and lethal types of cancer. As occurs in almost all types of cancer, its cure depends in a critical way on it being detected in the initial stages, when the tumor is still small and localized. Thus an objectively standardized criteria is required for clinically and histological identification of the individuals suffering from lung cancer. Regular follow-up and professional screening is needed to intercept lung cancer during its curable stage. Death rates due to lung cancer are doubling every decade, no cure disseminated disease is available, so diagnosing and educating individuals at increased risk at an early stage is vital to increase survival rates. With this view, the present study will help the radiologist as well as physicians to identify quickly the deadliest form of lung cancer in the early stage. The texture feature estimation algorithms are applied to various lung cancer chest X-ray images such as small-cell (SC)and non-small-cell (NSC) type, as well as on tuberculosis (TB) images (49 images from each category). Initially, the identifying features are obtained from the X-ray images using image processing and analyzing methods. Then, these features are applied to an expert system to classify the lung cancers into malignant (SC, NSC) and benign (TB).

***Index terms*: Lung cancer, X-ray images, Image processing, ANN.**

I. INTRODUCTION

One of the most important and difficult tasks the radiologist has to carry out consists of the detection and diagnosis of cancerous lung nodules from chest radiographs. Some of these lesions may not be detected due to the fact that they may be camouflaged by the underlying anatomical structure, or the low-quality of the images or the subjective and variable decision criteria used by radiologists. Previous studies [1] showed that radiologists fail to diagnose small lung nodules in as many as 30% of positive cases. In recent research, digital image processing techniques have been used in developing CAD systems for locating suspected nodules [4], [8], but too many false-positive (FP) classifications/chest radiograph are made. These FP's include rib crossings, rib vessel crossings, vessel-vessel crossings and end-on vessels [9]. Thus, the problem to solve for early diagnosis of lung cancer is associated with the reduction of the number of FP classifications while maintaining a high degree of true-positive (TP) diagnoses, i.e., sensitivity. Several methods have been proposed to reduce the number of FP's while maintaining a high sensitivity [8]–[11]. Most of them operate in two phases, i.e., feature extraction and feature classification. Some morphology-based algorithms have been proposed to extract specific features, such as circularity, size, and contrast [8], [9] or local curvature [2], [11]. Other authors attempt to tackle this second phase with the implementation of artificial neural networks (ANN's), which act as classifiers [3], [10], [12]–[14]. The critical task is to define and specify a *good* feature space that means the type of features which will discriminate between nodules and non-nodules, malignant and benign etc. Interpreting a chest radiograph is extremely challenging. Superimposed anatomical structures make the image complicated. Even experienced radiologists have trouble distinguishing infiltrates from the normal pattern of branching blood vessels in the lung fields, or detecting subtle nodules that indicate lung cancer [9]. When radiologists rate the severity of abnormal findings, large inter-observer and even intra-observer differences occur [10], [11]. The clinical importance of chest radiographs, combined with their complicated nature, explains the interest to develop computer algorithms to assist radiologists in reading chest images [12]. These are problems that cannot be corrected with current methods of training and high levels of clinical skill and experience. These problems include the miss rate of detection of small pulmonary nodules, the detection of minimal interstitial lung disease and the detection of changes in preexisting intestinal lung disease. The present research work describes the computerized technique to identify the lung nodules by extracting various discriminating geometrical and textural features like area, perimeter, irregularity index, standard deviation, skew-ness, third moment, entropy etc. using image processing and analyzing algorithms. Then these features are applied as an input to the feed forward neural network for the classification of lung cancer. Thus the developed algorithms aid the physician to detect the cancer in a short time with more accuracy.

II. MATERIAL AND METHOD

A. Image Samples for the Study

The presented work uses a set of digital images consisting of 49 small-cell types of lung cancer images, 49 non-small-cell types of lung cancer images, and 49 tuberculosis images, a total of 147 images (samples) each of 512 X 512 pixels in size. The images are obtained by scanning the X-ray images collected from private hospitals, and browsing the public database JSRT (Japanese Society of Radiological Technology) from internet. The digitized images are stored in the JPEG format with a resolution of 8 bits per plane. All images are stored as 512 X 512 X 256 JPEG raw data.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 2, Issue 3, September 2012

B. ***Pre-Processing Of Images***

Most of the pre-processing is done with the help of MATLAB software. Each image sample is scanned, and stored to a size of 512 X 512 pixels. Generally during the scanning, the quality of image is affected by different artifacts due to non uniform intensity, variations, motions, shift, and noise. Thus, the pre-processing of image aims at selectively removing the redundancy present in scanned images without affecting the details, that play a key role in the diagnostic and analysis process. Hence, image filtering becomes the important step in preprocessing. Therefore each image is median filtered to improve its quality (Fig. 1).

Fig. 1 Median Filtered X-Ray Image with Manually Segmented Lung Field Mask

C. Lung Field Segmentation

The first step in image analysis is the segmentation that separates the tumor from the background to produce the description of the tumor. To restrict the segmentation area as well as to remove the other bony structures (like shoulder bones), segmentation of lung fields is essential. In lung fields' segmentation, lung field masks have been prepared manually by deriving the peripheral pixel co-ordinates. Mask is a logical image denoting field area with logical '1' values and rest of the area with logical '0' values (Fig. 1). Lung fields are separated from the rest of the image portion by multiplying mask with the median filtered image.

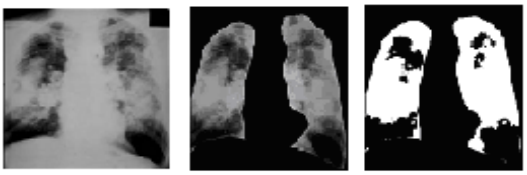
D. Lung nodule segmentation

Thresholding technique is applied on the separated lung fields' images. The valley point value between the two peaks of the histogram is used as threshold value for segmentation of nodule (Fig. 2). In most of the cases it is seen that with this threshold value, the other lung field area apart from the nodule having similar gray (pixel) values is also segmented. Further the region labeling (in case of small-cell type lung cancer images (Fig. 3)) and region growing (for the non-small-cell type of lung cancer images) algorithms are applied for separating the nodules from the background. Morphological operations like dilation and erosion are used to remove the artifacts from the image while segmentation.

(a) (b) (c)

Fig. 2 Segmented NSCLC Type Lung Fields after Multiplication, (B) Image after Thresholding, (C) Separated Nodule

Region-growing is one of the conceptually simplest approaches to image segmentation. In this algorithm, neighboring pixels of similar amplitude are grouped together to form segmented region. In the first stage of the process, pairs of quantized pixels are combined together in groups called *atomic regions* if they are of the same amplitude and are four connected. In the second stage, merging of the weak and common boundaries between the regions is carried out.



(a) (b) (c)

Fig. 3 (A) SCLC Original Image, (B) Separated Lung Fields, (C) Separated Cancerous Portion

Region-labeling replaces each pixel by a negative number representing the label of the region to which the pixel belongs. The algorithm uses a list to keep track of pixels that are yet to be labeled. Insertion and removing of the pixel are the two operations carried out with reference to the list. Pixel is inserted at the end of the list and, pixel is removed from the front of the list.

E. Features extraction

Geometrical features like area, diameter, perimeter, and irregularity index have been estimated from the separated lung nodules. The number of pixels having the values '1'in the image array gives the area of the segmented tumor image. The algorithm is developed which estimates the area using *bit quads*, 2-by-2 pixel patterns. The number of boundary pixels in the tumor image is estimated as the perimeter of the tumor image. According to the morphology of tumor, the shape of the tumor is circular in nature. To find out the irregularity in the circular shape, the circularity index is measured by using the equation like, $I = 4\pi A / p^2$: where, P is the perimeter of the tumor and A is area of the tumor in pixels. Geometrical features for the figure 2 (c) are included in table I. Texture or the contrast features are important features used in the classification of the lung cancers. Contrast features are again classified under two categories, first order statistic and second order statistic.

Table I Geometrical Features

Sr. No.	Features	Value
1	Area	2815
2	Perimeter	226.85
3	Diameter	59.686
4	Irregularity Index (I)	0.69

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 2, Issue 3, September 2012

Texture related features like average gray level, standard deviation, smoothness, third moment; uniformity and entropy are estimated using Gray Level Co-Occurrence Matrix technique (GLCM). First order statistic features with reference to figure 3 (b) are included in table II.

Table II Texture Features

Sr. No.	First-order statistic features	Values
1	Avg. gray level	48.57
2	Std. deviation	61.36
3	Smoothness	0.06
4	Third moment	2.50
5	Uniformity	0.28
6	Entropy	4.37

A co-occurrence matrix, also referred to as a co-occurrence distribution, is defined over an image to be the distribution of co-occurring values at a given offset. Mathematically, a co-occurrence matrix C is defined over an n x m image I, parameterized by an offset (Δx, Δy), as:

$$C(i,j) = \sum_{p=1}^{n} \sum_{q=1}^{m} 1, \text{ if } I(p,q) = i \text{ and } I(p+\Delta x, q+\Delta y) = j$$

$$C(i,j) = \sum_{p=1}^{n} \sum_{q=1}^{m} 0, \text{ otherwise}$$

The 'value' of the image originally referred to the grayscale value of the specified pixel. The value could be anything, from a binary on/off value to 32-bit color and beyond. Note that 32-bit color will yield a 2^{32} X 2^{32} co-occurrence matrix. It is also possible to define the matrix across two different images. Really any matrix or pair of matrices can be used to generate a co-occurrence matrix, though their main applicability has been in the measuring of texture in images, so the typical definition, as above, assumes that the matrix is in fact an image. Note that the (Δx, Δy) parameterization makes the co-occurrence matrix sensitive to rotation. We choose one offset vector, so a rotation of the image not equal to 180 degrees will result in a different co-occurrence distribution for the same (rotated) image. This is rarely desirable in the applications co-occurrence matrices are used in, so the co-occurrence matrix is often formed using a set of offsets sweeping through 180 degrees (i.e. 0, 45, 90, and 135 degrees) at the same distance to achieve a degree of rotational invariance. Similarly GLCM technique is also used in estimating contrast features like, contrast, correlation, energy, and homogeneity. While calculating these feature values the spatial relationship between the two adjacent pixels in various directions (represented by angles 0, 45, 90, and 135 degrees) is considered (Fig. 4). In the presented work, all the features required in the classification of lung cancers are extracted from 147 samples employing the various algorithms as briefed above.

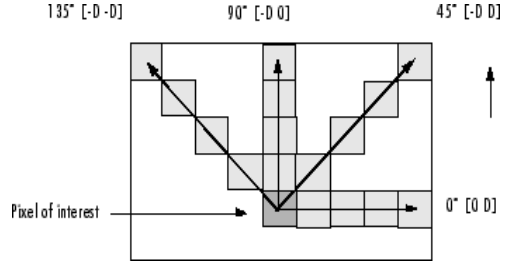


Fig. 4 Offsets and Their Directions

Table III Texture features

Second-order statistic features	For offset [0 1] 0^0	For offset [-1 1] 45^0	For offset [-1 0] 90^0	For offset [-1-1] 135^0	Avg. value
Contrast	0.15	0.19	0.11	0.19	0.16
Correlation	0.97	0.97	0.98	0.97	0.97
Energy	0.36	0.35	0.36	0.35	0.35
Homogen-eity	0.98	0.97	0.98	0.97	0.97

Table III Illustrates the Second Order Statistic Feature Values for the Figure 3 (B).

F. TB Image Analysis

A main problem in the texture analysis of chest radiographs is the complex background of superimposed normal anatomical structures to which the analysis must be somehow insensitive. One way to solve this problem would be to restrict texture analysis to regions of interest. An alternative could be to pre-process the images so as to remove normal background structures. Here the approach is to divide the separated lung fields in parts (4 here) and analyze each part separately, with texture features extracted solely from these parts. In this way, the classifier should capture knowledge regarding the normal variation within that particular part. The avg. gray level, standard deviation (second moment), skew (third moment), uniformity, entropy etc., of each filtered image are computed as texture features. According to the morphology of TB, the TB infection mostly affects the posterior segment of upper lobes. Therefore for the analysis, the separated lung field image is divided in to 4 sections like upper right (UR), upper left (UL), lower right (LR), and lower left (LL) (shown in Fig. 5). Then using GLCM technique the 1st and 2nd order statistic features are estimated for each section. The 1st order statistic feature values for the image shown in Fig. 5 are included in the Table IV. Tables V depict the 2nd

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 2, Issue 3, September 2012

order statistic feature values (avg. values considering all offsets) for 4 sections for the image included in Fig. 5.

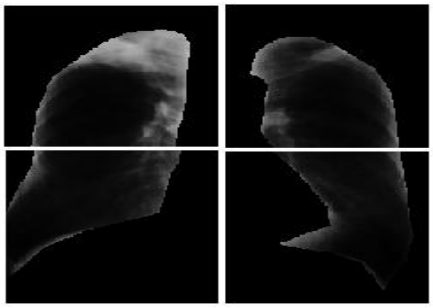


Fig. 5 TB Image Divided into 4 Sections

Table IV 1st Order Statistic Feature Values For TB Image Shown In Fig. 5

	UR	UL	LR	LL
Avg. Gray Level	96.231	81.723	73.587	25.284
Std. Deviation	84.279	89.796	78.815	46.871
Smoothness	0.0985	0.1103	0.0872	0.0327
Third Moment	0.3062	3.1721	3.1139	2.674
Uniformity	0.1596	0.2843	0.2423	0.517
Entropy	5.343	4.2395	4.7598	2.931

Table V 2nd Order Statistic Feature Values For TB Image Shown In Fig. 5

Sr. No.	Second Order Statistic Features	Avg. Value (UR)	Avg. Value (UL)	Avg. Value (LR)	Avg. Value (LL)
1	**Contrast**	0.379	0.353	0.164	0.292
2	**Correlation**	0.965	0.971	0.931	0.969
3	**Energy**	0.220	0.329	0.537	0.280
4	**Homogeneity**	0.964	0.968	0.971	0.968

G. ANN based classification

Further these estimated features are applied as input pattern to an expert system, which is designed to test the effectiveness of the input features so as to discriminate the lung cancer images. Artificial Neural Network (ANN) theory and practice suggest that, in a diagnostic application, the network should be trained with a balanced mixture of inputs from each diagnostic class. With this approach, a set of image samples consisting of 50% of small-cell-type, 50% of non-small-cell-type, and 50% of tuberculosis category are used for training and testing the network. In the presented work, a three layered, feed-forward artificial neural network is used. It has the first layer as the input layer consisting of eight nodes, one hidden layer with 8 hidden nodes (h1 to h8), and three output nodes (o1, o2 and o3) in the output layer. The input layer uses 8 inputs as follows --- Avg. gray level (AGL), Std. deviation (STD), Smoothness (SMT), Third moment (THM), Uniformity (UNF), Entropy (ENT), Contrast (CNT), and Energy (ENG). Three outputs will indicate whether the input sample is from SCLC (100), NSCLC (010) or TB (001) category. The network is trained by using feed forward Back-Propagation algorithm. Back-propagation algorithm is used for training the network that works on a Levenberg-Marquardt method. For each training pattern presented to the input layer of the network, error at the nodes in the output layer of the network is estimated. Back-propagation algorithm refers to the propagation of error of the nodes from the output layer to the nodes in the hidden layers. These errors are used to update the weights of the network. The amount of weights to be added or subtracted to the previous weight is governed by delta rule. Experimentation started with approximately 50% of samples (25 samples) from each category used to train the network with 6 numbers of hidden nodes in hidden layer. Five unknown samples from each category are used for to test the network. The transfer function used in between input and hidden layer is of tansigmoid type. Pure linear type of transfer function is used in between hidden and output layer. Levenberg – Marquardt (LM) training algorithm is used to train the network with learning rate of 1 and an error tolerance of 0.005. Poor (33.33%) classification performance has been found during this experimentation. Improved classification performance has been observed when the number of training samples has been increased up to 39 (for each category). Classification performance up to 83% has been observed for a network of 8:8:3 type with 39 training samples (from each category) and 10 testing samples. The number of epochs required to meet the goal is 1892. Initially, the momentum and weights are randomized to zero.

III. RESULTS AND DISCUSSION

Being scattered nature of the infection area in case of SCLC type of images only *area* of the infected portion is estimated. It is seen that, *irregularity index* in SCLC type of images is always less than 0.1. Geometrical features like *area*, *perimeter*, and *irregularity index* have been estimated meaningfully for the segmented nodules of NSCLC type of images. First and second order statistic features are calculated for all types of images. According

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 2, Issue 3, September 2012

to the TB database, TB infection appears mostly in the upper lobes of the lungs. Therefore texture feature estimation is carried out after dividing the TB image into 4 sections (to separate upper and lower lobes). Significant difference is seen in the 1st and 2nd order statistic feature values of the upper and lower lobes in case of TB database. Both feature values are marginally larger in upper lobes than the lower one. Results include only upper lobe values. The feature extraction techniques discussed in the above section have been applied on 147 images (49 from each category).

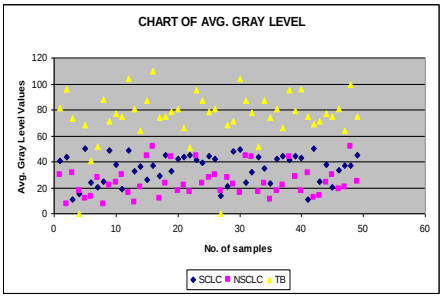


Fig. 6 Clustering Of Avg. Gray Level Values

Graphical representation of few sampled texture features is included in figures 6, 7, and 8.

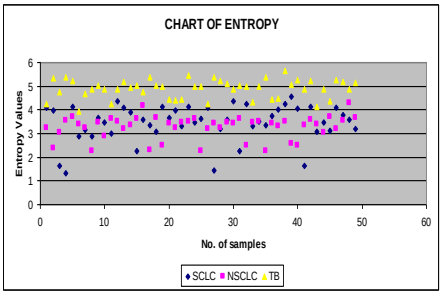


Fig. 7 Clustering Of Entropy Values

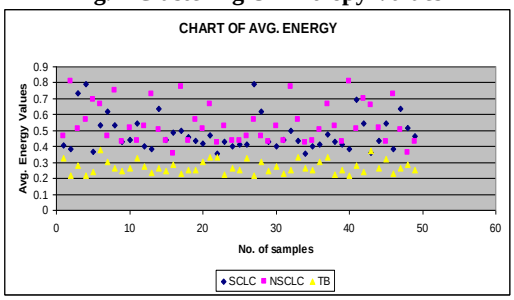


Fig. 8 Clustering Of Avg. Energy Values

IV. CONCLUSION

A fast and effective method to detect the lung nodules, and separate the cancer images from other lung diseases like TB is becoming increasingly needed due to the fact that the incidence of lung cancer has risen dramatically in recent years and an early detection can save thousands of lives each year. In this research work, an attempt is made to detect the lung tumors from the cancer images and supportive tool is developed to differentiate cancer images from the TB images. For this purpose, a method of feature extraction is derived and implemented. As an outcome, the following conclusions are confirmed in this research work.

A. Area

Before estimating the various geometrical or texture features, it is essential to separate the lung field area from the rest of the image portion. Manually segmented lung field masks are used to separate the lung fields from the background. It is observed that, due to the scattered nature of the infected portion in case of SCLC type of images, the area values are sufficiently larger than the NSCLC type images. In case of NSCLC type of images the position of the lung nodules is seen anywhere in the lung field portion. Variations in the area values are also seen in case of lung nodules. Hence it is concluded that area of the lung tumor do not provide any specific information about the classification.

B. Diameter and Perimeter

Being spreaded nature of the affected portion in small-cell type of lung cancer images the developed algorithm does not provide any meaningful information about diameter and perimeter values. In case of non-small-cell type of lung cancer images it is observed that, the circularity of most of the lung nodules is very poor. Therefore developed algorithm is unable to estimate diameter values in case of NSCLC type images. This concludes that, a diameter and perimeter value does not provide any substantial information about the classification.

C. Irregularity Index

It is observed from the segmentation in case of small-cell type of lung cancer images that, in almost all the cases the segmented portion is touching the irregular lung field boundaries. Therefore irregularity index values are smaller than 0.1 in almost all the cases. Sufficient variations in irregularity index values are seen in case of non-small-cell type of lung cancer images. Hence this feature alone is not sufficient to discriminate the cancer image from TB image.

D. Texture and Contrast Features

After separating the lung field area from rest of the image portion, texture and contrast feature values are estimated for all the SCLC, NSCLC, and TB images. In case of TB images, infection occurs in the upper areas of the lungs. Therefore segmented TB images are further divided in to the 4 sections and texture and contrast features are estimated only for the upper lobes. By using statistical approach, six texture features are calculated using gray level co-occurrence matrix technique. It is observed that, almost all the texture features contribute for discrimination in case of cancer and TB images. Spatial relationship between the pixel values in various directions is studied using multiple GLCMs. Multiple GLCMs are created to estimate contrast features like, contrast, energy, correlation and homogeneity. It is observed that, not all these features

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 2, Issue 3, September 2012

contribute for classification of cancer and TB images. Only contrast and energy are the main discriminating features, whereas features, correlation and homogeneity have overlapping range of values for both the classes.

E. ANN and Classification Accuracy

The diagnostic results obtained are found to be very promising. As high as 83% accuracy in classification is achieved using training data sets of reasonable size. Classification accuracy is improved as the numbers of training samples are increased. The present study also concludes that, back- propagation algorithm of ANN is a good choice for classification of cancer and TB images. This supervised training algorithm produce results faster than the other traditional classifier. Graphical representation of the classification accuracy is shown in Fig. 9. This method enables the researchers or the clinicians to formulate new strategies and techniques in immunophenotyping research, wherein human decision making is partly substituted for by the expert system.

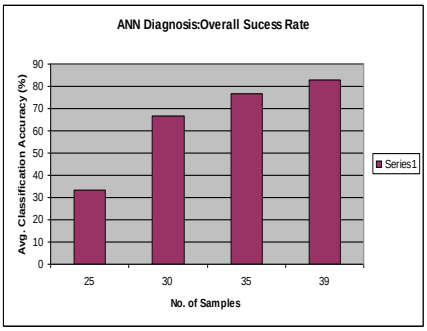


Fig. 9 Classification Accuracy

REFERENCES

[1] R. S. Fontana, D. R. Sanderson, L. B. Woolner, W. F. Taylor, W. E. Miller, and J. R. Muhm, "Lung cancer screening: The Mayo program," J. Occupat. Med., vol. 28, pp. 746 – 750, 1986.

[2] R. N. Strickland, "Tumor detection in non stationary backgrounds," IEEE Trans. Med. Imag., vol. 13, pp. 491–499, June 1994.

[3] S. B. Lo, S. L. Lou, J. S. Lin, M. T. Freedman, and S. K. Mun, "Artificial convolution neural network techniques and applications for lung nodule detection," IEEE Trans. Med. Imag., vol. 14, pp. 711– 718, Aug. 1995.

[4] M. L. Giger, K. Doi, H. Mac Mahon, C. E. Metz, and F. F. Yin, "Image feature analysis and computer-aided diagnosis in digital radiography.

[5] Automated detection of nodules in peripheral lung fields," Med. Phys., vol. 15, no. 2, pp. 158 – 166, 1988.

[6] M. L. Giger, N. Ahn, K. Doi, H. Mac Mahon, and C. E. Metz, "Computerized detection of pulmonary nodules in digital chest images: Use of morphological filters in reducing false-positive detections," Med. Phys., vol. 17, no. 5, pp. 861 – 865, 1990.

[7] T. Matsumoto, H. Yoshimura, K. Doi, M. L. Giger, A. Kano, H. Mac Mahon, K. Abe, and S. M. Montner, "Image feature analysis of false-positive diagnoses produced by automated detection of lung nodules," Invest. Radiol., vol. 27, no. 8, pp. 587 – 597, 1992.

[8] J. S. Lin, P. A. Ligomenides, Y. M. F. Lure, M. T. Freedman, and S. K. Mun, "Application of neural networks for improvement of lung nodule detection in radiographic images," in Proc. Symp. Compute. Assist. Radiol (S/CAR'92), pp. 108 – 115, 1992. M. J. Carreira, M. G. Penedo, D. Cabello, and J. M.Pardo, "Computer-aided lung nodule detection in chest radiography," in Lecture Notes in Computer Science: Image Analysis Applications and Computer Graphics, vol. 1024. Berlin, Germany: Springer-Verlag, pp. 331 – 338, 1996.

[9] J. M. Boone, V. Sigillito, and G. S. Shaber, "Neural networks in radiology: An introduction and evaluation in signal detection task," Med. Phys., vol. 17, no. 2, pp. 234 – 241, 1990.

[9] Y. S. P. Chiou, Y. M. F. Lure, and P. A. Ligomenides, "Neural networks image analysis and classification in hybrid lung nodule detection (HLND) system," in Proc. IEEE Workshop on Neural Networks for Signal Processing, pp. 517 – 526, 1993.

[10] J. S. Lin, S. B. Lo, A. Hasegawa, M. T. Freed man, and S. K. Mun, "Reduction of false positives in lung nodule detection using a two- level neural classification," IEEE Trans. Med. Imag., vol. 15, pp. 206 – 216, Apr. 1996.

AUTHOR'S PROFILE

First Author Completed B.E. in Electronics from Shivaji University in 1988. Joined I.I.T., Bombay for M.Tech in Bio-Medical Engineering during 1995 and completed in 1997. Currently working as a Professor and Head in Electronics & Telecommunication Engg. dept. of DKTE's, Textile & Engg. Institute, Ichalkaranji, Maharashtra State, India. Total publications & presentations on credit are about 35. Received PhD (Electronics Engg.) in May 2011. Research topic was, "A novel neural network approach to detect and classify lung cancer using chest radiographs". Having life membership of ISOI, BMESI, ISTE, IETE etc. Total teaching experience is of 22 years, and one year is of Industrial.

Second Author Completed B.E in Mechanical Engg. From PDA Engineering college, Gulbarga and M.E (Machine design) from BMS Engineering College, Bangalore. Currently working as Associate Professor at BLDEA Engg. College, Bijapur, and having 27 years of teaching experience. He is having 06 publications on his credit.