

A Computer Based Feature Extraction of Lung Nodule in Chest X-Ray Image

Nitin S. Lingayat and Manoj R. Tarambale

Abstract—Feature extraction is one of the most important step in CAD (Computer Assisted Diagnosis) system. It helps CAD system to take correct decision and increase its accuracy by providing distinguish feature of malignant and benign tumor. Computer based system is proposed in this paper for feature extraction of lung nodule from the X-ray image. In recent years, the image processing mechanisms are widely used in several medical areas for early detection and in deciding treatment stages, where the time and cost factor is very important to discover the disease in the patient. Among the cancer, lung cancer is one of the most common causes of death worldwide. Therefore, early detection using diagnostic tests promises to reduce mortality from lung cancer. Present paper deals with the problem of developing a computer based system for the extraction of maximum features from the segmented suspicious area from the lung X-ray image. Further, these properties can be used to classify lung tumor as benign or malignant from the X-ray image directly. Calculated features will help the CAD system to take correct decision.

Index Terms—Benign, feature extraction, irregularity index, lung nodule, malignant, neoplasia, variance.

I. INTRODUCTION

In medical field, neoplasia term is used for cancer. The term neoplasia means new growth, the new growth produced is called neoplasm or tumor (except new growth of tissues and cells those exist in the process of embryogenesis, regeneration, hormonal stimulations etc.). Satisfactory definition of neoplasm or tumor is a mass of tissues formed as a result of abnormal, excessive, uncoordinated, autonomous and purposeless proliferation of cells. The branch of science dealing with the study of neoplasms or tumors is called Oncology. Neoplasm may be benign when they are slow growing and localized without causing much difficulty to the host or malignant when they proliferate rapidly, spread throughout the body and may eventually cause death of the host. The common term used for all malignant tumors is cancer and malignant tumor of lung is referred as lung nodule [1].

Today, cancer is one of the most formidable health problem faced by mankind. Worldwide survey indicates cancers in all forms are causing almost 17% of death and in the developed countries; cancer is second cause of death next to

cardiovascular diseases. According to WHO (World Health Organization), estimation during year 2008, more than 12 million death are because of cancer. More people die from cancer than from AIDs, malaria, and tuberculosis combined. In India during year 2008, about 950000 new cancer cases being detected and nearly half of these cases die each year. Changes in life style, changes in environment and increase in life expectancy are the main causes of cancer. World-wide survey indicates that among the cancer, LUNG CANCER is the first most common cause of death in both men and women together [2].

In medical imaging, nodule detection is one of the most challenging tasks. Nodules are difficult to detect in digital images (X-ray) because of low contrast, large variation in density, varying size and location of the lung nodule within an area of a complicated anatomy (such as the hilum and ribs). At present, increased work load on interpretation of digital images (X-Ray and CTs) by radiologist can be a potential source of error due to fatigue in detecting subtle lesion are one of the problem in analyzing X-ray image. The main objective of the proposed work is to extracted maximum distinguish features from the lung nodule, which will helps the CAD system in classifying the lung tumor as malignant or benign [3]. Feature extraction properties are feed as an input to the CAD system.

Fig. 1 shows a CAD system. In first step, hard copy of images is converted into the softcopy with the help of high resolution scanner. These scanned images are stored in the storage drive for further analysis and processing. These scanned images are imported into the CAD system. CAD system performs two main tasks namely detection task and classification task. Detection task further classified as image acquisition, enhancement and segmentation. In image acquisition all the scanned images are resized to the standard size. As the CAD system algorithm are developed for specific size image. Enhancement process of digital image is required before applying any segmentation and feature extraction algorithm. During the scanning process and storing process of X-ray image, a lot of irrelevant information (noise) is added. For removing irrelevant information and for enhancing the X-ray image, low pass filter, high pass filter and point processing techniques are employed. After enhancing the image, segmentation is done to extracted interested area from the rest of the image.

Second task of CAD system is classification task. In this task feature extraction and classification is done. In feature extraction mathematical, textural and geometrical properties are calculated from the segmented region. Further these properties are used for classification purpose. With the help of intelligent system and feature extracted properties, CAD system decide whether the given image contain some

Manuscript received April 20, 2013; revised July 16, 2013.

Nitin S. Lingayat is with the Electrical Department, DR. Babasaheb Ambedkar Technological University's Institute of Petrochemical Engineering, Lonere, Dist: -Raigard, Pin-402103 India (e-mail: nslingayat@yahoo.com).

Manoj R. Tarambale is with Marathwada Mitra Mandal's college of Engineering, University of Pune, Pune, India (tel.: +91 9822117427; e-mail: manoj_tarambale@yahoo.com).

suspicious part or not. Present paper deals with the feature extraction properties, how they are calculated and how they are useful for the detection of lung nodule.

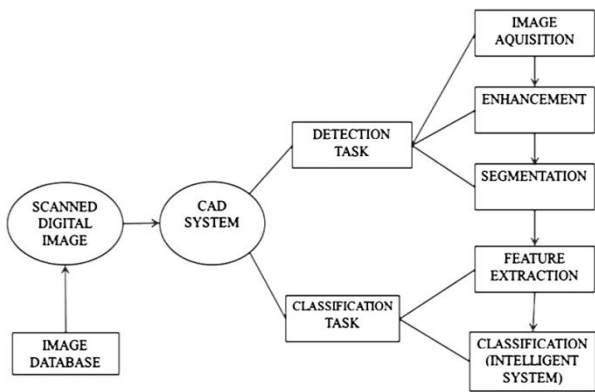


Fig. 1. CAD system.

II. BLOCK DIAGRAM

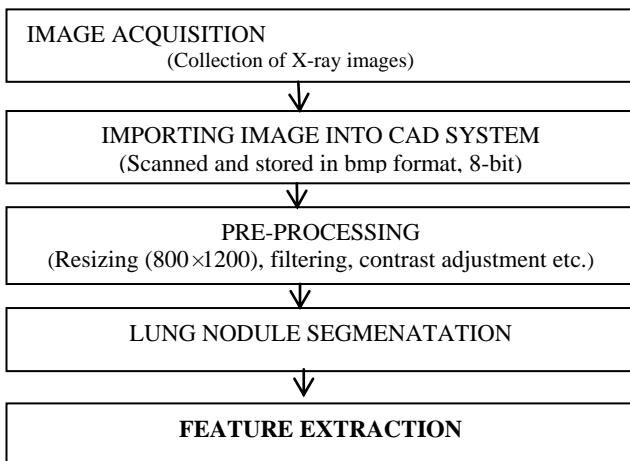


Fig. 2. Stages of propose algorithm.

Fig. 2 shows a sequence of steps which are to be followed before extracting features from the chest X-ray image. These steps are required for efficient and accurate calculation of features from the X-ray image. The present paper deals with the feature extraction step.

III. METHOD

JSRT public database and images which are collected from nearby hospitals are digitized and imported into the CAD system with the help of high resolution scanner (Scanjet 2400, HP India) and stored in 256 gray levels (8 bit). Imported images are resized to 800×1200 pixels and stored in bmp format. Preprocessing algorithm is applied to remove irrelevant information (noise). For removing noise (Gaussian and salt & pepper noise) low pass filters are used. To remove the blurring, which is introducing during the low pass filtering is removed by using high boost filter. Contrast, brightness and intensity problems are removed by using contrast stretching, histogram equalization, negativity and power law transformation etc. Using modified thresholding, labeling algorithm and edge detection, segmentation of the lung nodule X-ray image is carried out. Features such as geometric properties, textural properties and mathematical properties are

calculated from the segmented image by their respective formula. These calculated features are used as an input to the CAD system. CAD system used these features for classifying the lung nodule as malignant or benign. Present CAD system will not reduce the role of doctor/radiologist but it will provide a second opinion to the doctor. The result of this system along with doctor analysis will increase the accuracy of the diagnosis. Fig. 2 block diagram shows the overall process of the CAD system. Segmentation and feature extraction techniques are easily implemented and executed by using digital image processing toolbox of MATLAB software (see Fig. 3).

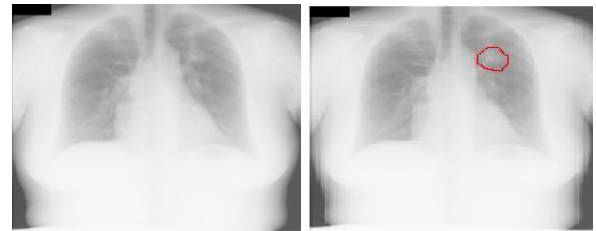


Fig. 3 a). Original image.

Fig. 3 b). Location of lung nodule marked with red circle.

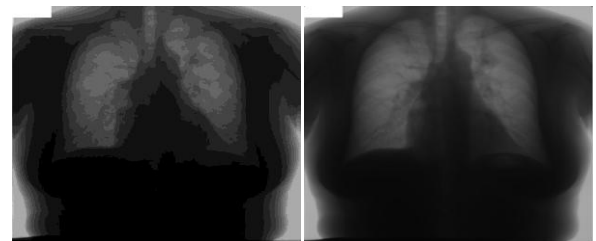


Fig. 3 c). Negative image.

Fig. 3 d). Median filter (5x5 mask).

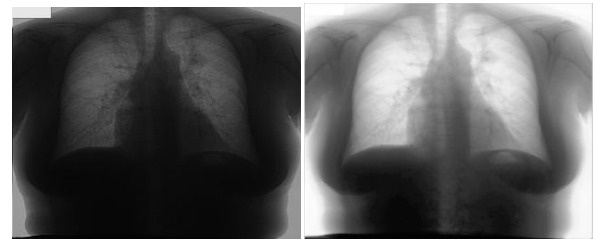


Fig. 3 e). High boost filter.

Fig. 3 f). Histogram equalized Image.

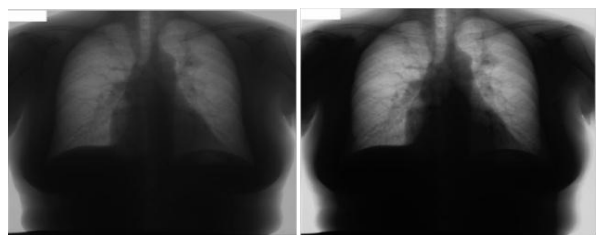


Fig. 3 g). Median filter (3x3).

Fig. 3 h). Contrast stretching.

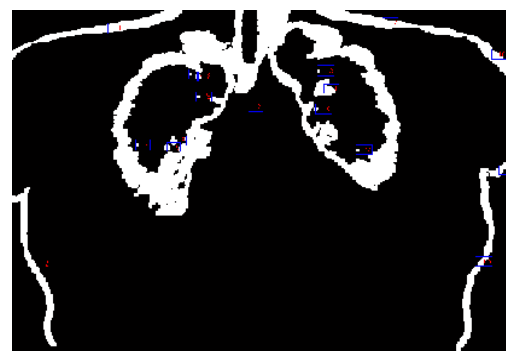


Fig. 3 i). Labeled image.

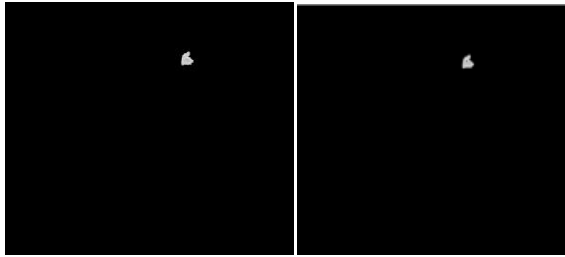


Fig. 3 j). Extracted lung nodule. Fig. 3 k). Extracted lung nodule with original gray level.

IV. FEATURE EXTRACTION

In the present study, image feature extraction is very important stage of computer based system. Feature extraction provides certain parameters, on the basis of which computer system takes decision. After the segmentation is performed on lung region, the features can be obtained from it and the diagnosis rule can be designed to detect nodules in the lung. The entire feature which are calculated from the image, convey some information regarding lung nodule. This information is very helpful in detecting lung nodule as malignant or non-malignant. In this literature, the features extracted from the X-ray image can be used as diagnostic indicators [4].

A. Area

Area of the segmented tumor is computed by counting the number of pixels which have the value '1' in the image array. The area (A) in the object is the just the count of the ones in the image array. For computing area, binary image is used [5].

$$A = n [] \quad (1)$$

where, $n []$ represents the count of number of the patterns within the parenthesis.

B. Perimeter

Perimeter property is calculating the distance between each adjoining pair of pixels around the border of the region. If the image contains discontinuous regions, then it returns unexpected results. Perimeter is calculated by counting the pixels contained only in the boundary [5].

C. Irregularity Index

Lung cancer is characterized partially by the irregularity in its tumor border. For this analysis, the irregularities in the tumor are computed by an index:

$$I = \frac{4\pi \times Area}{(Perimeter)^2} \quad (2)$$

The metric value or roundness or circularity index or irregularity index (I) is equal to 1 only for circle and it is < 1 for any other shape. Here it has been assumed that, more circularity of the object, the probability of the object being nodule is high [6].

D. Equivalent Diameter

Scalar that specifies the diameter of a circle with the same area as the region [5], [7]. Computed by formula

$$E_{diameter} = \sqrt{\frac{4 \times area}{\pi}} \quad (3)$$

E. Convex Area

The convex hull or convex envelope of a set X of points in the Euclidean plane is the smallest convex set that contains X . For instance, when X is a bounded subset of the plane, the convex hull may be visualized as the shape formed by a rubber band stretched around X . Convex area is the number of pixel in convex image. Convex image is the size of the bounding box of the region. Bounding box is convex hull.

F. Solidity

It gives the proportion of the pixels in the convex hull that are also in the region.

$$Solidity = area / convexarea \quad (4)$$

G. Gray-Level Co-Occurrence Matrix Properties

A statistical method of examining texture that considers the spatial relationship of pixels is the gray-level co-occurrence matrix (GLCM), also known as the gray-level spatial dependence matrix. The GLCM functions characterize the texture of an image by calculating how often pairs of pixel with specific values and in a specified spatial relationship occur in an image, creating a GLCM and then extracting statistical measures from this matrix.

The number of gray levels in the image determines the size of the GLCM. The gray-level co-occurrence matrix can reveal certain properties about the spatial distribution of the gray levels in the texture image. For example, if most of the entries in the GLCM are concentrated along the diagonal, the texture is coarse with respect to the specified offset.

Offsets define pixel relationships of varying direction and distance. However, a single GLCM offset might not be enough to describe the textural features of the input image. For example, a single horizontal offset might not be sensitive to texture with a vertical orientation. For this reason, multiple GLCMs offset for a single input image [8].

$$\begin{aligned} \text{Offsets} = & [0 \ 1; 0 \ 2; 0 \ 3; 0 \ 4 \dots \\ & -1 \ 1; -2 \ 2; -3 \ 3; -4 \ 4 \dots \\ & -1 \ 0; -2 \ 0; -3 \ 0; -4 \ 0 \dots \\ & -1 \ -1; -2 \ -2; -3 \ -3; -4 \ -4]; \end{aligned}$$

The Fig. 4 illustrates the spatial relationships of pixels that are defined by this array of offsets, where D represents the distance from the pixel of interest.

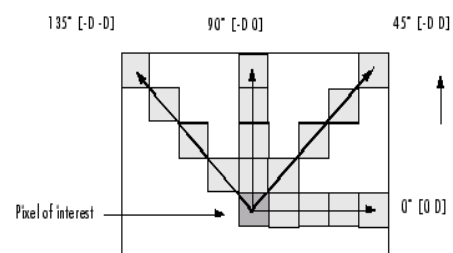


Fig. 4. Gray level co-occurrence matrix with different offsets. Shaded region at the middle indicate the pixel of interest.

Several statistics properties drive from GLCM. These statistics provide information about the texture of an image. The following are the statistics properties are listed below

H. Contrast

It measures the local variations in the gray-level co-occurrence matrix. It returns a measure of the intensity contrast between a pixel and its neighbor over the whole image. Contrast is 0 for a constant image. Formula for contrast is given by

$$\sum_{i,j} |i-j|^2 p(i,j) \quad (5)$$

where, 'i' is x coordinate and 'j' is y coordinate of the pixel. $p(i,j)$ is a pixel value at that location.

I. Correlation

It measures the joint probability occurrence of the specified pixel pairs. Returns a measure of how correlated a pixel is to its neighbor over the whole image. Range is from 1 to -1. Correlation is 1 or -1 for a perfectly positively or negatively correlated image. Correlation is null for a constant image. It is given by formula

$$\sum_{i,j} \frac{(i-\mu_i)(j-\mu_j)\bar{p}(i,j)}{\sigma_i\sigma_j} \quad (6)$$

J. Energy

Provide the sum of squared elements in the GLCM. It is also known as uniformity or the angular second moment. Returns the sum of squared elements in the GLCM. Its range is from 0 to 1. Energy is 1 for a constant image.

$$\sum_{i,j} p(i,j)^2 \quad (7)$$

K. Homogeneity

It returns a value that measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal. Its range is from 0 to 1. Homogeneity is 1 for a diagonal GLCM.

$$\sum_{i,j} \frac{p(i,j)}{1+|i-j|} \quad (8)$$

L. Statistical Properties of an Image

It is a single measure of some attribute of a sample. It is calculated by applying a function (statistical algorithm) to the values of the items of the sample, which are known together as a set of data.

M. Entropy

Measures the randomness of a gray-level distribution. The Entropy is expected to be high if the gray levels are distributed randomly throughout the image. It is given by

$$E = -\sum_i \sum_j P[i,j] \log P[i,j] \quad (9)$$

N. Mean

It provides the mean of the gray levels in the image. The Mean is expected to be large if the sum of the gray levels of the image is high. It is given by

$$\mu = \frac{1}{N \times M} \sum_{i=0}^M \sum_{j=0}^N P(i,j) \quad (10)$$

O. Variance

Variance tells us, how spread out the distribution of gray levels is. The Variance is expected to be large if the gray levels of the image are spread out greatly. It is given by

$$\sigma^2 = \frac{1}{N \times M} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} (P(i,j) - \mu)^2 \quad (11)$$

P. Standard Deviation

Standard Deviation shows much 'variation' or 'dispersion' exists from the average.

$$Std = \sqrt{\sigma^2} \quad (12)$$

V. RESULT AND DISCUSSION

Chest radiographs has always played an important role in differential diagnosis and determining the extent of disease. In developing countries like India, Pakistan, Bangladesh etc. where people cannot afford expensive detection test such as CTs, MRIs, PETs scan [15], in such countries X-ray plays a very important role for earlier detection of disease. X-ray scanning centres are easily available even in or near rural areas as compare to other scanning centre. Generating X-ray from the X-ray machine is less complicated and less time consuming. Therefore, X-ray is the first and most preferred can by which the radiologist can detect, suspect and decide whether the patient is having a suspicious tumor in the lung or not. Further testing will classify this nodule as benign or malignant. Early detection not only increases the chances of surviving the patient but it also reduces the unwanted medical expenses.

As X-ray images contain noise and irrelevant information, some time because of some internal organ overlapping, accidental marks on the body, hairs on the body, bad illumination, very higher or very lower brightness and contrast makes the task of detecting lung nodule from X-ray image very difficult by radiologist. Increased work load on interpretation of digital images (CTs and X-Rays) studies for a radiologist can be a potential source of fatigue and error in detecting suspicious nodule from the chest X-ray. In this paper it is tried to extract as many information as possible in the form of features using digital image processing on X-ray image. Feature extracted from the image is used for determining and classifying the suspicious area of the lung (nodule). Features are calculated from their formulas as given in section IV. The developed CAD system will not replace the doctor/radiologist role but it will help the doctor to take correct decision in short time with accuracy. It will act as a

second opinion to the doctor/ radiologist. The result of CAD system and the analysis of doctor together increase the accuracy of detecting malignant and benign lung tumor. If the malignant tumor is detected in its earlier stage, then the chance of surviving the patient increases. Feature extraction method suggested in this paper is less costly, less time consuming and easy to implement.

For implementing the algorithm Digital Image Processing and GUI toolbox of MATLAB software are used. All the images are randomly pickup from the database. Images which are used are of different age group, different sex and the location of lung nodule is also different.

Following images Fig. 5 a) and Fig. 6 a) are taken from lung cancer database which are certified for malignancy and benign by radiologist. Fig. 5 b) and Fig. 6 b) are the enlarged version segmented lung nodule from the original X-ray image by the use of software, which is developed by using GUI and image processing toolbox in MATLAB.



Fig. 5 a). Benign image. Fig. 5 b). Enlarge segmented nodule.



Fig. 6 a). Malignant image. Fig. 6 b). Enlarge segmented image.

TABLE I: PARAMETER COMPARISON TABLE FOR MALIGNANT AND BENIGN LUNG CANCER CASE

PARAMETER	BENIGN	MALIGNANT
Area	72	349
Perimeter	34.3848	167.196
Irregularity index	0.7654	0.1569
Equivalent Diameter	9.5746	21.0799
Convex area	82	739
Solidity	0.8780	0.47222
Mean	0.0191	0.0927
Variance	4.68E+06	2.27E+07
Standard deviation	2.16E+03	4.76E+03
Entropy	0.0011	0.0047

TABLE II: GLCM PROPERTIES CALCULATED IN DIFFERENT DIRECTION FOR BENIGN CASE

Benign				
	[-1,-1]	[-1,0]	[-1,1]	[0,1]
Contrast	0.0015	0.0011	0.0016	0.0012
Correlation	0.7917	0.8472	0.7778	0.8333
Energy	0.9998	0.9998	0.9998	0.9998
Homogeneity	1	1	1	1

The features which are calculated from the image Fig. 5 a)

and Fig. 5 b) are listed in a Table I, Table II, and Table III.

TABLE III: GLCM PROPERTIES CALCULATED IN DIFFERENT DIRECTION FOR MALIGNANT CASE

Malignant				
	[-1,-1]	[-1,0]	[-1,1]	[0,1]
Contrast	0.0076	0.0061	0.0078	0.0046
Correlation	0.7879	0.828	0.7822	0.871
Energy	0.9991	0.9991	0.9991	0.9992
Homogeneity	0.9999	0.9999	0.9999	0.9999

VI. CONCLUSION

In benign tumor, area and perimeter value are large as compare to malignant tumor. But the irregularity index for benign is more. Irregularity index gives an idea about, how much the boundary of the nodule is irregular. Malignant tumor are more irregular that why irregularity index is near to zero. Solidity value for malignant is less because of its irregular surface. For benign tumor, mean, variance and standard deviation values is lower as compared to malignant tumor. Entropy value is higher for malignant tumor. Equivalent diameter show higher value in malignant tumor. In GLCM properties, the average of contrast, correlation, energy and homogeneity values are calculated in different directions. Calculated properties gives a distinguish values for malignant or benign tumor, which helps in deciding the given image is of malignant tumor or benign tumor.

ACKNOWLEDGMENT

We would like to express our deepest appreciation to the Japanese Society of Radiological Technology (JSRT) in cooperation with the Japanese Radiological Society (JRS) for providing clinical images for research purpose.

REFERENCES

- [1] D. Sharma and G. Jindal, "Computer aided diagnosis system for detection of lung cancer in CT scan images," *International Journal of Computer and Electrical Engineering*, vol. 3, no. 5, pp. 714-718, Oct. 2011.
- [2] J. J. Dignam, L. Huang, L. Ries, M. Reichman, and E. Feuer, "Estimating breast cancer statistic and other-cause mortality in clinical trial and population-based cancer registry cohorts," vol. 115, pp. 5272-5283, Nov. 2009.
- [3] S. N. Deepa and B. A. Devi, "A survey on artificial intelligence approaches for medical image classification," *Indian Journal of Science and Technology*, vol. 4, no. 11, pp. 1583-1595, Nov. 2011.
- [4] K. P. Aarthy and U. S. Ragupathy, "Detection of lung nodule using multiscale wavelets and support vector machine," *International Journal of Soft Computing and Engineering (IJSCE)*, vol. 2, issue 3, July 2012.
- [5] R. C. Gozalez and R. E. Woods, *Digital Image Processing Using Matlab*, 2nd ed, Gatesmark, USA, 2002, ch. 12, pp. 642-654.
- [6] S. A. Patil and V. R. Udpi, "Chest x-ray features extraction for lung cancer classification," *Journal of Scientific and Industrial Research*, vol. 69, pp. 271-277, April 2010.
- [7] A. K. Jain, *Fundamental of Digital Image Processing*, 1st ed., Englewood Cliffs: Prentice Hall, 1989, ch. 7, pp. 241-250.
- [8] M. S. Ahmad, M. S. Naweed, and M. Nisa, "Application of texture analysis in the assessment of chest radiograph," *International Journal of Video & Image Processing and Network Security (IJVIPNS)*, vol. 9, no. 9, pp. 291-297.
- [9] S. Jayaraman, S. Esakkirajan, and T. Veerakumar, *Digital Image Processing*, 3rd ed., Patel nagar, New Delhi: Tata McGraw Hill, 2009, ch. 5, pp. 245-273.
- [10] S. A. Patil, "Texture analysis of TB x-ray images using image processing technique," *Journal of Bio-Medical and Bioengineering*, vol. 3, issue 1, pp. 53-56, March 2012.
- [11] A. Pathan and B. K.saptalkar, "Detection and classification of lung cancer using artificial neural network," *International Journal on Advanced Computer Engineering and Communication Technology*, vol. 1, issue 1, pp. 62-67, 2012.

- [12] A. Amutha and R. S. D. Wahidabanu, "A novel method for lung tumor diagnosis and segmentation using level set- active contour modeling," *European Journal of Scientific Research*, vol. 90, no. 2, pp. 175-187, Nov. 2012.
- [13] L. Boroczky, L. Zhao, and K. P. Lee, "Feature subset selection for improving the performance of false positive reduction in lung nodule CAD," *IEEE Transaction On Information Technology In Biomedical*, vol. 10, no. 3, pp. 504 – 511, July 2006.
- [14] F. Taher and R. Sammouda, "Morphology analysis of sputum colour images for early lung cancer diagnosis," in *Proc. 10th International Conference on Information Science, Signal Processing and their Application*, Malaysia, 10-13 May 2010, pp. 296-299.
- [15] M. N. Mughal and W. Ikram, "Early lung cancer detection by classifying chest CT images: a survey," *Multitopic Conference, 2004, Proc. INMIC 2004. 8th International*, 24-26 Dec. 2004, pp.67-72.
- [16] K. Le, "A design of a computer-aided diagnostic tool for chest x-ray analysis," *International Journal of Computer Science and Information Technology (IJCSIT)*, vol. 3, no 2, pp. 212-223, April 2011.
- [17] A. Chaudhary and S. S. Singh, "Lung cancer detection using digital image processing," *International Journal of Research in Engineering and Applied Sciences (IJREAS)*, vol. 2, issue 2, pp.1352-1359, Feb. 2012.
- [18] D. Vivekanandan and S. R. Raj, "A Feature extraction model for assessing the growth of lung cancer in computer aided diagnosis," *IEEE*, pp. 953-958, Sep. 2011.
- [19] P. Bhayan and G. Jindal, "A Segmented morphological approach to detect tumor in lung images," *IJCST*, vol. 2, issue 3, pp. 499-502, Sep. 2011.



Nitin S. Lingayat was born in Shahapur (Thane), India in 1971. He received the B.E. degree in electrical engineering from the University of Poona, Pune in 1992 and the M.Tech degree from Indian Institute of Technology Bombay, Mumbai in 1998. He obtained Ph.D. degree in electrical engineering from J.N.V. University, Jodhpur, India in 2008.

He is a head, Electrical Engineering Department, Institute of Petrochemical Engineering of Dr. B .A. Technological University, Lonere - Maharashtra (India) from January 2004. He has a teaching experience of 21 years. His research interest includes biomedical signal and image processing, energy management, statistical signal processing, non-conventional energy sources etc.

Dr. Lingayat is a member of IAENG, Institute of Engineers (India), Indian Society for Technical Education (ISTE). He has published many papers in the National and International Journals.



Manoj R. Tarambale was born in Jalgaon District, India, in 1970. He received graduate degree (B.E.) in electrical engineering from BVCOE, Pune-43, University of Pune, India, in 1992 and post graduate degree (M.E.) in Control System from WCOE, Sangli, Shivaji University, Kolhapur, India in 2002. Currently, he is pursuing his Ph.D. degree from PACIFIC University, Udaipur, Rajasthan, India.

He has one year industrial experience and twenty years teaching experience. At present, he is Assistant Professor and Head of electrical engineering department of Marathwada Mitra Mandal's College of Engineering, Pune-52, University of Pune, India. His main research interests are bio-medical image processing, electronic instrumentation, process control instrumentation and medical diagnosis.

Prof. Tarambale is a member of Institute of Engineers (India) and Indian Society for Technical Education (ISTE).