Lung Cancer Detection at Initial Stage by Using Image Processing and Classification Techniques

Bhawana Malik[1], Jaykant Pratap Singh[2], Veer Bhadra Pratap Singh[3], Prashant Naresh[4]

Student of Masters of Technology Computer Science, Department of Computer Science and Engineering, *APJKTU, Lucknow, India*[1]

Assistant Professor, Department of Computer Science and Engineering, APJKTU, Lucknow, India[2]

Assistant Professor, Department of Computer Science and Engineering, APJKTU, Lucknow, India[3]

Assistant Professor, Department of Computer Science and Engineering, APJKTU, Lucknow, India[4]

***Abstract*—** *Cancer is becoming a huge threat in human life .There are different kinds of cancer, Lung cancer is common type of cancer causing very high ephemerality rate. To effectively identify lung cancer at an early stage is an important application of image processing. In this paper, an approach is presented which will detect lung cancer at prior stage using CT scan images of Dicom format. One of the key challenges is to dislodge white Gaussian noise from the Computed Tomography scan image, for that purpose non local mean filter is lied. Otsu's thresholding is used to segment the lung image. Morphological operations such as opening, closing, edge detection and region filling are applied as post processing on the images to make image more clear for the detection of nodule. To form feature vector, the textural and structural features are extracted from the processed image. Data mining algorithms are used to the extracted features for the detection of lung cancer. In this project, three classifiers namely Support Vector Machine, Artificial Neural Network and k-NN are applied for the detection of lung cancer to find the rigor of disease (stage I or stage II) and comparison is made with Artificial Neural Network, and k-NN classifier with respect to different quality attributes such as accuracy, sensitivity, precision and specificity.*

***Keywords*— Computer aided diagnosis, CT-Scan images, Feature Extraction, Image Processing, Lung Cancer, Segmentation.**

1.INTRODUCTION

For the human beings among various diseases cancer has become a big threat, as per Indian population census data. In India the second most common disease responsible for maximum mortality is Cancer with about 0.3 million deaths per year [1]. According to GLOBOCAN 2012, 14.1 million new cancer cases estimated in which 8.2 million cancer-related deaths occurred in 2012, compared with 12.7 million and 7.6 million, respectively, in 2008. The most commonly diagnosed cancers worldwide were those of the lung (1.8 million, 13.0% of the total) [2]. Lung cancer is the main cause of cancer death worldwide. It is difficult to identify in its early stages because only in the advanced stage symptoms appear among all other types of cancer [3]. If lung nodules can be identified accurately at an early stage, the patient's survival rate can be increased by. In the modern era of automation, the field of automated diagnostic systems plays an important role. Medical Image Processing is one such field in Automated diagnostic system designs where numerous systems are proposed and still many more under conceptual design due explosive growth of the technology today [4].

Data mining provides the method for analysis the useful information from data for decision making**.** As the volume of data is growing exponentially with the increase in population, there is a greater need to extract the knowledge from the data. Predicting the outcome of a disease is a challenging tasks of data mining [5]. Data mining tool has proved to be successful in disease diagnosis [6]. Data mining has already started to find its application in the diagnosis of cancer such as cancer lesion detection [7], pulmonary nodule detection [8], and classification of cancer stage from tree-text histology report [9], breathe biomarker detection [10] and so on. Various Data Mining Tools are available for Disease Diagnosis or for the prediction of disease outcome. The data mining tasks can be broadly classified in two categories: descriptive and predictive. Descriptive mining tasks describe the general properties of the pre- stored data. A predictive mining task gives conclusion on the basis of current data. Data mining algorithms follow three different learning methods: supervised, unsupervised, or semi-supervised. The classification task can be seen as a supervised technique where each instance belongs to a class [11].

By image processing an image is converted into digital format to get a useful image or to obtain useful information from it. It is a category of signal processing in which input is image, such as video frame or photograph and output can be an image or characteristics associated with that image. Image segmentation is the process to allocate a label to every pixel in an image in such a manner that pixels with the same label share certain visual characteristics. Image processing is used in medical image processing for the Lung cancer diagnoses [3].

This paper is organized into six sections. In section 2 related work carried out in this field is described. In section 3 proposed methodology for early detection of cancer is explained. In section 4 system architecture of nodule predictor is discussed. In section 5 experimental result and discussion is explained followed by conclusion in section 6.

2.RELATED WORK

In this section, some of the works on prediction of lung cancer, pre-processing, segmentation and classification techniques have been discussed

For automatic lung nodules detection a two stage scheme in Multi-Slice Computed Tomography (MSCT) [12] scans with multiple SVMs to reduce number of false positive with accuracy of 87.82% is presented. Three SVMs classifiers for are used on preprocessed images to categorize the candidates as nodule or non-nodule. An automatic CAD system [13] of 80% accuracy is developed by analyzing LUNG CT images using several steps for early detection of lung nodule. A computerized system in [14] for lung cancer detection in CT scan images consists of two stages: a) lung segmentation and enhancement, b) feature extraction and classification. To remove background and extracts the nodules from an image threshold segmentation is applied. A feature vector is calculated for possible abnormal regions and neuro fuzzy classifier is used to classify such regions. Doing so, an accuracy of 95% was achieved.

In the image processing procedures of [15], processes such as image pre-processing, segmentation and feature extraction are discussed. Sensitivity of 95% in [16] is presented, to improve the efficiency of the diagnosis system for lung cancer, through a region growing segmentation [17] method to segment CT scan lung images. For noise removal, Linear-filtering and contrast enhancement is used as preprocessing step to prepare the image for segmentation. A system which predicts lung tumor from Computed Tomography (CT) [18] images through image processing techniques coupled with neural network classification. To segregate lung regions Optimal thresholding [19], is applied to the denoised image. Region growing method is used to segment Lung nodules. A set of textural features extracted from the extracted ROIs by the back propagation neural network with an accuracy of 86.30%.

An efficient lung nodule detection scheme of accuracy 80.36% [20] is developed to perform nodule segmentation through weighted fuzzy possibilistic [21] based clustering is carried out for lung cancer images. The RBF kernel based SVM classifier performs better than linear classifier. An automatic Computer-Aided Detection (CAD) scheme in [22] is presented having accuracy 95% that can identify the pulmonary nodule at prior stage. Simple but efficient methodology for lung nodule classification without the stage of segmentation [23] with 84% accuracy. Bayesian classification and Hopfield Neural Network algorithm [24,25] for extracting and segmenting the sputum cells is presented for the purpose of lung cancer early diagnosis which achieved an accuracy of 88.62%. Morphological processing on the segmented image improved the performance of HNN algorithm.

This study present an overview of different algorithm for classification and image processing used in the field of lung cancer prediction. Summary of various segmentation and classification techniques with their classification accuracy and sensitivity of nodule detection has been presented, based on above s it has been found that not much work has been carried out for early detection of lung cancer. Hence it has been taken up here.

3.METHODOLOGY

Methodology is composed of two phases.

1. In first phase, the CT scan image is pre-processed to remove Gaussian white noise using non-local mean filter. As the accuracy of the segmentation algorithm depends on the quality of image so, the image is cleansed and segmented using Otsu's thresholding [26] and then, the textural and structural features are extracted from the segmented image by the application of feature extraction techniques.

2. In the second phase, the SVM classifier is implemented and then it is trained and tested on sample data for the prediction of lung cancer and the output is checked for accuracy which is a measure of how accurately the classifier predicts the status of patient.

4.SYSTEM ARCHITECTURE OF NODULE PREDICTOR

Lung Nodule prediction aims to automatically predict the information of nodule presented in lung's medical images and addressed to physician, by making use of Computer Aided Diagnosis (CAD) system. A typical application of CAD System is the detection of a tumor. CAD is a relatively interconnected technology combining elements of artificial intelligence and digital image processing with image processing which play the roles of physicians and computers, whereas automated computer diagnosis is a concept based on computer algorithms only [27].

Computer Aided Diagnosis (CAD) system takes CT scan images of lung cancer patients as input and provides status of patient as output on the basis of classifiers. In lung CT image segmentation process, Gaussian noise is removed from CT scan image which is most common type of noise present in medical images. After that, segmentation is done using Otsu's thresholding to segment the lung part in an image. post processing enhancement is done to get clear image for detection of nodule (tumor) by detecting boundary in image using canny edge detection. Then, two largest regions are filled to remove extra muscle part from an image except lungs. In nodule's feature extraction module, output of post processing is given as input to extract textural and structural feature of nodule after that SVM classifier is trained and tested on the basis of those features to provide final output i.e severity of the disease. The figure.1 depicts the system architecture of CAD system.

4.1Pre-processing

In order to extract the information from millions of pixels in medical (CT, X-ray, etc.) images, all components in the CAD system are designed to reduce the amount of data.As an important step, image and data pre-processing serve the purpose of extracting regions of interest and reducing noise from the images, so that they can be efficiently processed by the Feature Extraction step.

Pre-processing is the method to correct different kind of errors in images, done before processing. It is needed in order to improve the quality of the image and make it available for next phases. The importance of the pre-processing stage of a Computer Aided Diagnosis (CAD) system for prediction of lung cancer lies in its ability to remedy some of the problems that may occur due to some factors

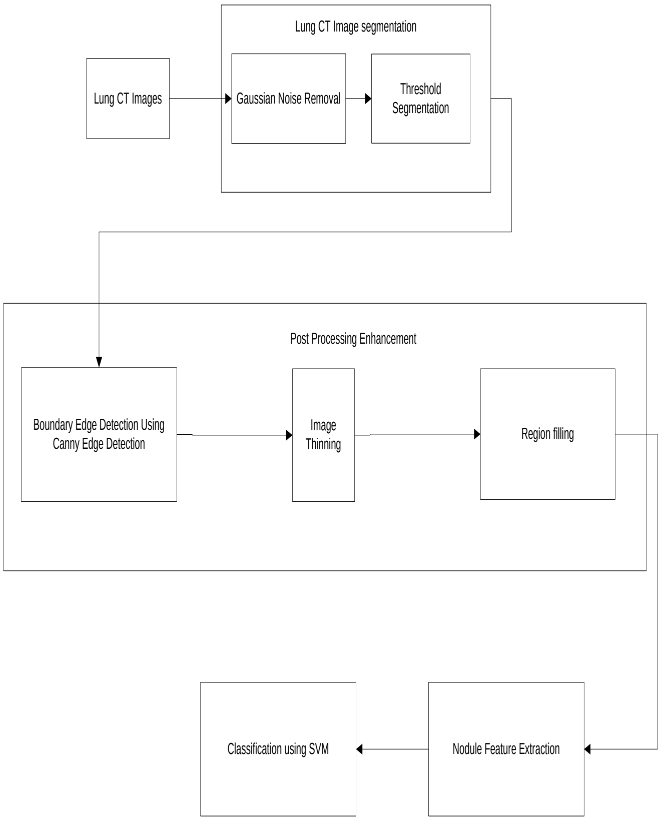

Lung CT Image segmentation
Lung CT Images
Gaussian Noise Removal
Threshold Segmentation
Post Processing Enhancement
Boundary Edge Detection Using Canny Edge Detection
Image Thinning
Region filling
Classification using SVM
Nodule Feature Extraction

Fig-1: System Architecture of CAD System

4.2Image Segmentation

The term image segmentation of an image involves the separation of lung nodule from other part of the CT scan images and then enhancement of the resultant image to get details.The goal in many tasks is for the regions that represent meaningful area of an image. Thresholding, clustering, comparison based, histogram based, edge detection and region growing are several general-purpose techniques have been developed for image segmentation.

4.3Post processing

The post processing means filling and thinning. The series of operations evolved in enhancement after segmentation are Morphological opening, Morphological closing, Morphological thinning, Morphological filling is applied on thresholded image for the enhancement. Morphological opening eliminates the small objects inside and outside the lungs. Morphological closing is then applied on the image. It enhances borders and fills the gaps in the border. After Morphological operations boundary of the enhanced image is detected. Morphological thinning is then applied on the boundary extracted image. After the thinning process Morphological filling is applied on the image to get the final post-processed image.Thinning brings down the width of the line. While Filling gets rid of small breaks and holes in the contour, remove extra part from image and make image more clear for nodule detection.

4.4Feature Extraction

Cancer nodule usually has large number of features. It is important to identify and extract interesting features form it. Aim of feature extraction is to find a set of features that define the shape of nodule as precisely and uniquely. Not all the attributes of segmented nodule are useful for knowledge extraction. Extraction of certain features that characterize the nodule, but excludes the insignificant attributes is the way of describing nodule.

There are two main types of feature descriptors, namely textural features and structural features. The textural is concerned with the spatial (statistical) distribution of gray tones, e.g. mean, standard deviation, energy etc [28]. The structural features consider the structure of the objects in the image, e.g. number of connected components in an image, area of the object, compactness etc [29]. It Computes the structural features value of nodule i.e. Area, Convex Hull Area, Equiv Diameter and Solidity.

4.5Classification

After features are extracted, algorithms are used for classifying the data into the categories. These algorithms are also known as classifiers. There are two kinds of classification: supervised classification and unsupervised classification. If the feature vectors are given with known labels (the corresponding correct outputs), then the training is called to be supervised. If the classifier categorizes the data automatically without any use of class labels, then it's known to be unsupervised classification. Examples for supervised classifiers are Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), Artificial Neural Networks (ANN), Classification Trees etc. and unsupervised classifiers include k-means clustering, mixture models, hierarchical clustering, Principal Component Analysis, Singular Value Decomposition etc [30].

5.EXPERIMENTAL RESULTS AND DISCUSSION

The Performance Metrics for the test data are shown in Table1. The formulations are shown in percentage, each column indicates the neural networks Classifier used and the rows indicate the Metric value respectively.

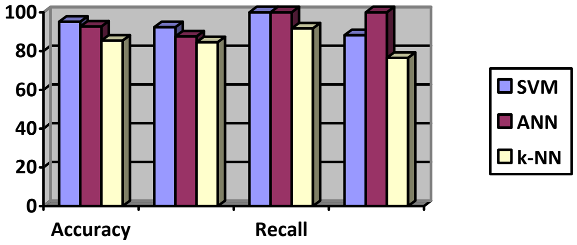


Fig-2: Performance metrics of Classifiers

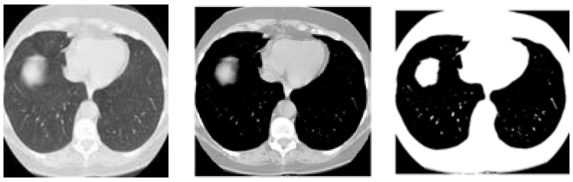
From table 1 it is shown that accuracy of SVM is 95.12% which is better that ANN classifier (92.68%) and k-NN classifier (85.37%). The graph of the Table 1 is given in Figure 2.

These results shown in figure 2 with 24 images of stage I and 17 images of stage II is used as test dataset. Value of accuracy, precision, recall and specificity is in percentage (%).

Table-1: Performance Metrics In Percentage For Test Data

		Classifier		
		SVM	ANN	KNN
Metrics	Accuracy(%)	95.12	92.68	85.37
	Precision(%)	92.31	87.50	84.62
	Recall(%)	100.00	100.00	91.67
	Specificity(%)	88.24	100.00	76.47

For experimentation of the technique, the CT images are obtained from a NIH/NCI Lung Image Database Consortium (LIDC) dataset. This data consists of 1000 lung images. Those images are progressed to this system. The diagnosis rules are then produced from those images and these rules are progressed to the classifier for the learning process. After learning, a lung image is progressed to the proposed system. Then the proposed system will execute its processing and finally it will detect whether the input image is having cancer or not. The proposed CAD system is capable of detecting lung nodules with diameter ≥ 2.5 mm, which means that the system is capable of detecting lung nodules when they are in their early stages. Thus facilitating early diagnosis will improve the patients' survival rate.


(a) (b) (c)

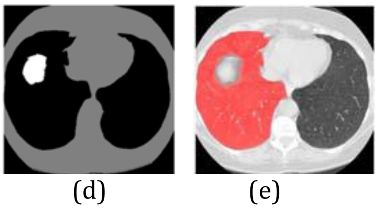

(d) (e)

Fig-3: Segmentation steps: (a) Original, (b) Pre-processing (c) Thresholding
(d) Segmenting Lung Region, (e) Showing Cancerous Nodule

6.CONCLUSION

The field of Disease Diagnosis is a continuously evolving and very active field for research. The main focus of the current approach is to predict the status of patient for initial stage detection of lung cancer. A novel approach for predicting Lung cancer nodule at prior stage using SVM Classifier is proposed here. The Structural and Textural Features have been used for reporting the nodule. The results got are very stimulating, data was tested on Support Vector Machine Classifier with RBF kernel obtained an accuracy of 95.12%. The classification rates obtained for the SVM, ANN and k-NN Classifier are 95.12%, 92.68% and 85.37%.

ACKNOWLEDGMENT

Initially, we would like to thank our almighty in the success of completing this work. We would extend our gratitude to all the experts for their critical comments.

REFERENCES

[1] Imran Ali, Waseem A. Wani and Kishwar Saleem, "*Cancer Scenario in India with Future Perspectives*", Cancer Therapy, vol. 8, 2011, pp. 56-70..

[2] Ferlay J, Soerjomataram I, Ervik M, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray, F (2013). GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11, Lyon, France: International Agency for Research on Cancer.

[3] S.Shaik Parveen, C.Kavitha, "*Detection of lung cancer nodules using automatic region growing meth*od", Proceedings of the 4th International Conference on Computing, Communications and Networking Technologies (ICCCNT), 2013, pp. 201-206.

[4] Guruprasad Bhat, Vidyadevi G Biradar , H Sarojadevi Nalini, "*Artificial Neural Network based Cancer Cell Classification (ANN – C3)*", Computer Engineering and Intelligent Systems, vol. 3, (2), 2012, pp. 116-119.

[5] Juliet R Rajan1, Jefrin J Prakash, "*Early Diagnosis of Lung Cancer using a Mining Tool*", Proceedings of the National Conference on Architecture, Software systems and Green computing-2013, pp. 87-91..

[6] Ada, Rajneet Kaur, "*A Study of Detection of Lung Cancer Using Data Mining Classification Techniques*", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, (3), 2013, pp. 67-70.

[7] T. Jia , Y. Wei, D. Wu, "*A Lung Cancer Lesions Detection Scheme Based on CT Image*", Proceedings of the 2nd International Conference on Signal Processing Systems (ICSPS), 2012, pp. 45-50.

[8] L. Yang, Y. Jinzhu , Z. Dazhe, "*A Method of Pulmonary Nodule Detection utilizing multiple support Vector Machine*", Proceedings of the International Conference on Computer Application and System Modeling, 2010, pp. 203-207.

[9] M. Iain , M. Darren, F. Mary-Jane, "*Classification of Cancer Stage from Free-text Histology Reports*", Proceedings of the 28th IEEE EMBS Annual International Conference New York City, USA, Aug 30-Sept 3, 2006, pp.156-159..

[10] D. Siqi, H. Tianlin , S. Yang, L. Chun, H. Yuanqing, "*Detection of Lung Cancer with Breath Biomarkers Based on SVM Regression*", Proceedings of the Fifth International Conference on Natural Computation 2009, pp. 93-96.

[11] Sunita Beniwal, Jitender Arora, "*Classification and Feature Selection Techniques in Data Mining*", International Journal of Engineering Research & Technology (IJERT), vol. 1, (6), 2012, pp. 94-97.

[12] Yang Liu, Jinzhu Yang, Dazhe Zhao, Jiren Liu, "*A Method of Pulmonary Nodule Detection utilizing multiple Support Vector Machines*", Proceedings of the International Conference on Computer Application and System Modeling (ICCASM 2010), 2010, pp. 118-121.

[13] Disha Sharma, Gagandeep Jindal, "*Identifying Lung Cancer Using Image Processing Techniques*", Proceedings of the International Conference on Computational Techniques and Artificial Intelligence (ICCTAI), 2011 pp. 115-120.

[14] Anam Tariq, M. Usman Akram and M. Younus Javed, "*Lung Nodule Detection in CT Images using Neuro Fuzzy Classifier*", Proceedings of the Fourth International Workshop on Computational Intelligence in Medical Imaging (CIMI), 2013, pp. 49-5

[15] Anita chaudhary, Sonit Sukhraj Singh, "*Lung cancer detection on CT images by using image processing*", Proceedings of the International Conference on Computing Sciences, 2012, pp. 143-146.

[16] Atiyeh Hashemi, Abdol Hamid Pilevar, Reza Rafeh, "*Mass Detection in Lung CT Images Using Region Growing Segmentation and Decision Making Based on Fuzzy Inference System and Artificial Neural Network*", I.J. Image, Graphics and Signal Processing, 2013, pp. 16-24.

[17] J. Quintanilla-Dominguez, B. Ojeda-Magaña, M. G. Cortina-Januchs, R. Ruelas, A. Vega-Corona, and D. Andina, "*Image segmentation by fuzzy and possibilistic clustering algorithms for the identification of microcalcifications,*" Sharif University of Technology Scientia Iranica, vol. 18, 2011, pp. 580–589.

[18] S.K. Vijai Anand, "*Segmentation coupled Textural Feature Classification for Lung Tumor Prediction*", Proceedings of the International Conference on Computing, Communications and Networking Technologies ICCCCT, 2010, pp. 518-524.

[19] Shiy ingH u, EricA Huffman, and Jospe h M. Reinhard t, "*Automatic lung segementation for accurate quantitiation of volumetric X-Ray CT images*", IEEE Transactions on Med ical Imaging, vol. 20 , (6), June 2001, pp. 490 -498.

[20] S.Sivakumar, Dr.C.Chandrasekar, "*Lung Nodule Detection Using Fuzzy Clustering and Support Vector Machines*", International Journal of Engineering and Technology (IJET), vol. 5, (1), Feb-Mar 2013, pp. 179-185.

[21] S.Sivakumar and C.Chandrasekar, "*Lung Nodule Segmentation through Unsupervised Clustering Models", Procedia Engineering*, vol. 38, pp. 3064-3073.

[22] JIA Tong, ZHAO Da-Zhe, YANG Jin-Zhu,WANG Xu, "*Automated Detection of Pulmonary Nodules in HRCT Images*", IEEE, 2007, pp. 38-41.

[23] Hiram Madero Orozco, Osslan Osiris Vergara Villegas, "*Lung Nodule Classification in CT Thorax Images using Support Vector Machines*", Proceedings of the 12th Mexican International Conference on Artificial Intelligence, 2013, pp. 277-283.

[24] Fatma Taher, Naoufel Werghi and Hussain Al-Ahmad, "*Bayesian Classification and Artificial Neural Network Methods for Lung Cancer Early Diagnosis*", IEEE, 2012, pp. 773-776.

[25] R. Duda, P. Hart,"Pattern Classification", Wiley-Interscience 2nd edition, October 2001.

[26] N. Otsu, "*A Threshold Selection Method from Gray-level Histograms*", IEEE Trans. Syst. Man Cybernetics, vol. 9, (1), 1979, pp. 62-66.

[27] Kunio Doi, "*Computer-aided diagnosis in medical imaging: Historical review, current status and future potential*", Computerized Medical Imaging and Graphics 31, 2007, pp. 198–211.

[28] Robert M. Haralick, K. Shanmugam, and Its'Hak Dinstein, "*Textural Features for Image Classification*", IEEE Transactions on Systems, Man and Cybernetics, vol. SMC-3, (6), November 1973, PP. 610-621.

[29] Cheriet, M., Kharma, N., Cheng-Lin, Suen, C.Y., "*Feature Extraction, Selection, and Creation*", in Character Recognition Systems, 1st ed, New Jersey, John Wiley & Sons, ch 4, sec 4.1, 2007, pp. 129-131.

[30] David Barber, "*Machine Learning Concepts*", in Bayesian Reasoning and Machine Learning, 1st ed, Cambridge University Press, United Kingdom, ch 13, sec 13.1, 2013, pp. 289-292.