International Journal of Advanced Research Methodology in Engineering & Technology, ISSN 2456-6446
Volume 2, Issue 1, January 2018

A CLASSIFICATION AND ANALYSIS OF PULMONARY NODULES IN CT IMAGES USING RANDOM FOREST

Capt. Dr. S.SANTHOSH BABOO
Associate Professor
P.G & Research Department Of Computer Science
Dwaraka Doss Goverdhan Doss Vaishnav College
Arumbakkam-106

E.IYYAPPARAJ
Research Scholar
P.G & Research Department Of Computer Science
Dwaraka Doss Goverdhan Doss Vaishnav College
Arumbakkam-106

Abstract-The main aim of this work is to propose a novel Computer-aided detection (CAD) system based on a Contextual clustering combined with region growing for assisting radiologists in early identification of lung cancer from computed tomography(CT) scans. Instead of using conventional thresholding approach, this proposed work uses Contextual Clustering which yields a more accurate segmentation of the lungs from the chest volume. Following segmentation GLCM and LBP features are extracted which are then classified using three different classifiers namely Random forest, SVM and k-NN.

Keywords—***Computer aided detection(CAD);computed tomography(CT)imaging; lungcancer; Support vector machine(SVM); Random forest(RS); K-NN.***

I. INTRODUCTION

According to the recent statistics collected by National Cancer Registry Programmers(NCRP) India occupies 11th position among top 15 countries in the world with higher Age Adjusted Incidence Rate(AAR). Further it is calculated that total number of new cancer cases registered will reach 13,88,397 by 2015 and 17,34,886 by 2020. Among these lung cancer alone accounts for 10% among male and 3% among female. However survival rates is still low (<50%) in most part of India. Therefore detection of Lung cancer at earlier stages is of great importance and it can increase survival rate of cancer patients .So an computer-aided detection (CAD) system in supplement to radiologists' diagnosis has become a promising tool to serve such purpose.

On the diagnosis of lung cancer the most important and nontrivial task is the Detection of pulmonary nodules since appearance of pulmonary nodules varies in a wide range, and also nodule densities have low contrast when compared with adjacent vessel segments and other lung tissues. For nodule detection Computed tomography (CT) has been shown as the most popular and widely used imaging modality in [2], [4], because of its ability to provide reliable image textures for the detection of small nodules. Over a past few decades there has been a significant progress [5], [6] in development of lung nodule CAD systems using CT image modality. Generally, such CAD systems consist of following three steps: 1) Segmentation 2) Feature extraction and 3) Classification.

In this work once various regions in CT images are obtained by segmenting the image they can be further used for different types of analysis and interpretations. Therefore, segmentation of image mainly involves extracting important features and deriving the relevant metrics to segregate regions of homogeneous intensities. In order to achieve this, it is necessary to choose a selective region of interest by considering the application requirements. In the past, many image segmentation methods have been proposed by various researchers for performing successive image analysis. Traditionally, many researchers have used the existing thresholding techniques for segmenting the various regions of interest. In short, the most frequently used techniques for segmentation in literatures are statistical methods, geometrical, structural, model based, signal processing methods, spatial domain filters, Fourier domain filtering. Gabor and wavelet models have also been used in most works present in the literature [6, 7].In [8] a new method for segmenting lung CT images by combining fuzzy logic with bit planes was proposed to locate the region of interest which consists of following three steps, namely identification, rule firing, and inference. In the first step, bit planes that represent the lungs clearly were identified. In the second step, using triple sig num function an optimum threshold is assigned based on the grayscale values for the anatomical structure present in the medical images. Fuzzy rules are formed based on the available bit planes to form the membership table and are stored in a knowledge base. Finally, in inference process fuzzy rules are fired to assign final segmentation values. In this paper, CC

along with the region growing algorithm has been used for effective segmentation of the CT Lung image.

The remainder of this paper is organized as follows: Section 2 discusses the methods proposed in related works. Section 3 explains the method used in this work. Section 4 provides the results. Section 5 gives the conclusion on this work and also provides some possible future works.

II. RELATED WORKS

Accurate segmentation of lung region is important since, the nodules present on it may be on the boundary of the lung parenchyma. So such lung nodules may lost and this reduces the detection accuracy, if the entire lung is not segmented accurately. So the ultimate goal of lung region of interest segmentation is to separating the vowels corresponding to lung region from the vowels corresponding to the surrounding anatomy.

Lin DT et al [10] proposed a novel threshold based segmentation approach for segmenting lung region present in the CT lung images. In their work, during preprocessing they have used a 5x5 median filter for removing the noise present in it. The Foreground region is then separated by omitting the rim of the image along with the background regions. In [11] segmentation is done by thresholding each image by an optimal threshold derived by comparing the curvature of the lung boundary along with the ribs. A combination of background-removal operator together with iterative gray level thresholding is used by Atonally et al. [12] for segmenting the lung region. In their work, due to the presence of noise the background was not well eliminated well.

Pu et al. [9] proposes an adaptive border marching algorithm to segment the lung region and reduces the under segmentation ratio. They used the gray-level thresholding to obtain the lung regions and then follow flood-filling methodology to remove any non-lung regions present after the thresholding. Ozekes et al [11] segmented the lungs of the CT images using Cellular Neural Networks trained by genetic algorithm. In their work, the lung regions were specified using the 8 directional searches and +1 or -1 value were assigned to each vowels. In the work proposed by Cao Lei et al [8], a rough image of lung was acquired by combining optimal thresholding together with mathematical morphology. A self-fit segmentation algorithm was then applied on the segmented result to obtain a final refined output.

In [16] a novel three step segmentation process is proposed for the analysis and segmentation of lung CT images was. In this approach, the task feature extraction is left to medical doctor if the area occupied by GGO in the CT image is large. This is due to the reason that when the area of GGO is small there is a possibility of overlooking the light gray shadows present in the image. In the first step, the extraction of region of interest is done in order to segment the lung area. Preprocessing techniques such as labeling, shrinking and expansion done in the CT by employing the process of binarization in order to achieve a better segmentation accuracy. In the second step of their work, parameters such as mean value, standard deviation, and semi interquartile range are extracted from GGO shadows. In the final step, the GGO shadow's regions were extracted by the process called linear discriminate function. Variable N-Quoit (VNQ) filter is used to extract suspicious shadows from GGO.

III methodology

This section describes proposed method for lung cancer detection. The proposed method involves three stages are shown in Fig.1. Initially the CT lung images are segmented using contextual clustering along with region growing algorithm. Next stage is Feature extraction which is done by extracting GLCM and LBP features. The third stage is classification with three different types of classifiers namely k-Nearest Neighbor (k-NN), Random forest (RF) and Support vector machines (SVM).

A. Segmentation Via Contextual Clustering With Region Growing

Region growing [1] is an iterative region based segmentation technique employed to identify connected regions of interest (contiguous sets of vowels) in images, obeying some inclusion rule (generally based on threshold values), and according to the notion of discrete connectivity [2]. The first step in region growing is to choose initial seed point. Initial region includes seed point and then each pixel in

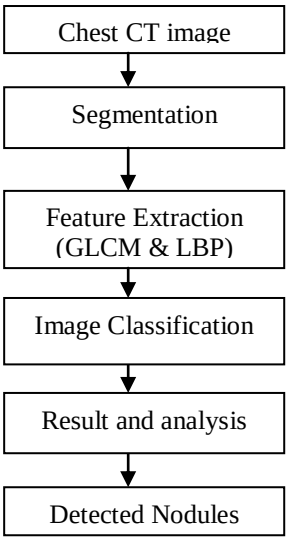

Fig . 1**:** Frame work for Analysis of pulmonary nodules in CT images

neighborhood of seed point are checked based on thresholding value for possible inclusion in region if condition is satisfied such pixels are added to region .Each included vowels becomes in turn a seed point for the following next iteration. The above process continues until all the pixels are added in the grown region based on the set of rules and threshold.

In our approach, a region growing approach along with the clustering is used to fix the threshold in order to segment the region of interest present in the CT lung images. The initial seed point is a vowels (3x3 or 5x5 pixels) belonging to the lung region, and then fuzzy rule fixes a threshold value to select vowels with intensity values lower than the threshold fixed.

This ensures that the entire lung parenchyma is connected which present in the CT lung image is starting from the bronchi, carina, and the trachea. A 3x3 pixel present in central slice of CT image is automatically chosen to be initial seed and then grows towards the entire lung region present in the image.

Recently, a lot of researchers use statistical clustering in image segmentation [3] .In a clustering technique along with the region growing, each pixel is associated with one of the finite number of threshold is grown to form disjoint regions. The contextual clustering method proposed by [4, 9] is a supervised algorithm. It uses a 3 X 3 overlapping windows of pixels to form a segmented image.

The quality of segmented image in contextual clustering depends upon following four factors 1) A defined threshold value T provided by user to choose the nearest region for segmentation (T=140 here), 2) a controlling parameter β which is in the range 0 to 1, 3) the median value of the all 9 pixels 3x3 pixel window 4) the total number of intensity values (u) inside the window, excluding the already identified median value.

Let us simple assume that contextual clustering segments a data into two different categories namely category 1 (ω_0) and category 2 (ω_1) based on the grown region. The steps in proposed method for implementing the contextual clustering to segment the lung region from CT lung images are mentioned as follows.

1) The decision parameter T (positive) and weight of neighborhood information β (positive) are chosen first.

2) Assume that the total number of data in the neighborhood be N_n. Let Z_i be the data itself with intensity 'i'.

3) The data is classified to category ω_1 when $Z_i > T\alpha$ else the data is classified to category ω_0.

4) The classification results are stored to variable C_0 and C_1.

5) The each data 'i', the total number of intensity values u_i in the neighborhood window of data 'i' belonging to class ω_1 is counted.

6) The data inside range is assigned to ω_1 and others to ω_0. The classification is stored to new variable C_2.

7) The data outside the range belong to ω_0.

8) If condition $C_2 \neq C_1$ and $C_2 \neq C_0$, is satisfied C_1 is copied to C_0 and C_2 is copied to C_1.

9) The algorithm returns back to step 3, otherwise the process is stopped and returned to C_2.

The following figure describes flow of algorithm.

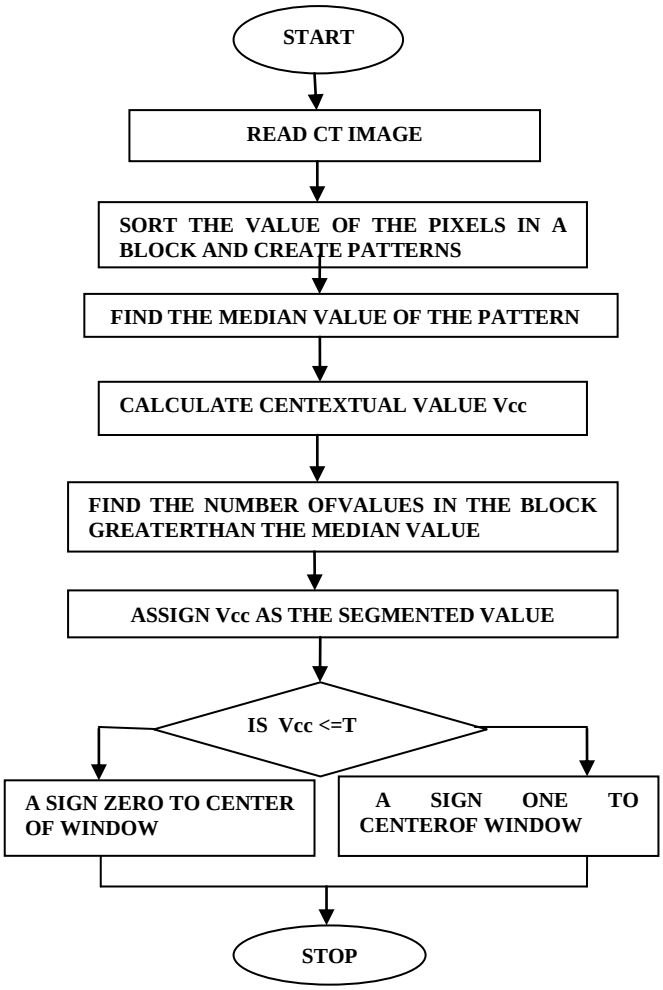

Fig. 2 Flow diagram with algorithm

$$V_{cc} = Median\ value + \frac{\beta}{threshold} * \left\{u - \frac{window\ size}{2}\right\} \quad (1)$$

B. Extraction Of GLCM Features

In statistical texture analysis, texture features are computed from the statistical distribution of observed combinations of intensities at specified positions relative to each other in the image. According to the number of intensity points (pixels) used in each combination, statistics are classified into first-order, second-order and higher-order statistics. The Gray Level Co occurrence Matrix (GLCM) method is a way of extracting second order statistical texture features. The approach has been used in a number of applications. Third and higher orders textures can also be formed by consider the relationships among three or more pixels. These are theoretically possible but not commonly implemented due to complexity and long computational time.

A GLCM is basically a matrix where the number of gray levels in image equals the number of rows and columns. The matrix element M(i, j | Δx, Δy) is the relative frequency with which two pixels with intensity value 'i' and other with 'j', are separated by a pixel distance (Δx, Δy),occur within a given neighborhood. The matrix element M(i, j | d, ɵ) contains the second order statistical probability values for changes between gray levels 'i' and 'j' at a particular displacement distance d and at a particular angle (ɵ). Using a large number of intensity levels L implies storing a lot of temporary data, i.e. a L×L matrix for each combination of (Δx, Δy) or (d, ɵ).

Since dimension of GLCM is very large they are sensitive to the size of the texture samples on which they are estimated. To avoid this more often, the number of gray levels is reduced.

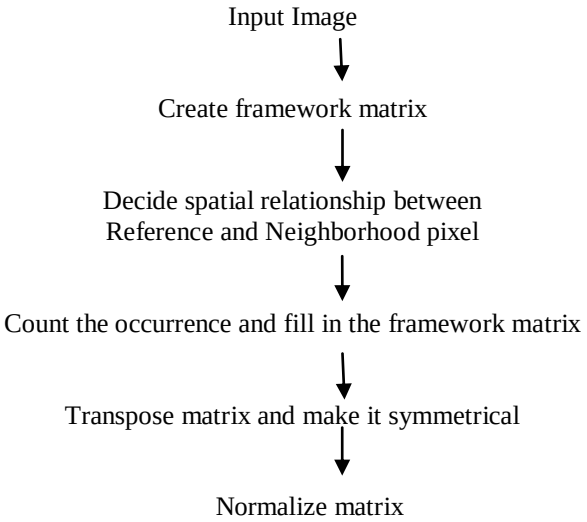

Fig.3 GLCM Feature Extraction Methodology

C. Extraction Of LBP Features

Local binary pattern (LBP) is an simple and less complex operator for texture feature extraction which uses simple comparison for feature extraction. Since LPB uses local image information for feature extraction first the whole image is divided into small fixed size blocks usually 16x16 pixels. Each pixel in the block is compared with their surrounding 8 neighborhood pixels in anticlockwise direction. Any neighborhood pixel greater than center pixel is represented by binary '1' else it is represented by binary '0'. The result of comparison a string of binary is then encoded as decimal number. So an 8 neighborhood will give decimal value up to 255. Then the feature vector of each block is represented as normalized histogram count of decimal value obtained for each pixel in that block. Local descriptors with each block are concatenated to form final feature vector.

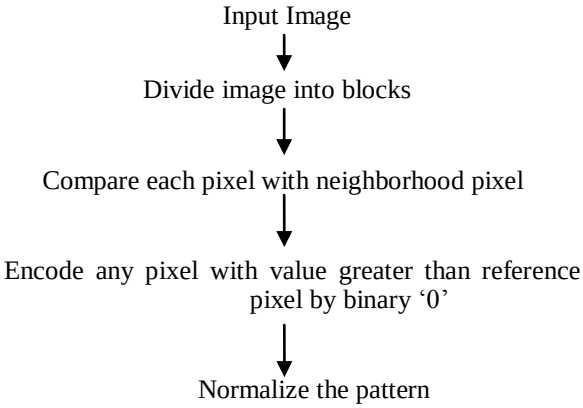

Fig.4 LBP Feature Extraction Methodology

Sometimes different number of sampling points and radius are used for improved accuracy and less computation. Since LPB uses local points for feature extraction they are invariant to illumination changes and also to rotation effects. Further since feature extraction uses simple comparison it can be implemented directly in hardware and also consume less memory.

D. Classifier

Three types of classifiers are used for classification namely Random forest, SVM and k-NN.

1) SVM: Support vector machines are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification .Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class

memberships. The basic SVM algorithm takes a set of input data along with set of predicts for classifying the data to the classes. From given set of training examples, an SVM training algorithm builds a model that assigns new examples into one category or the other.

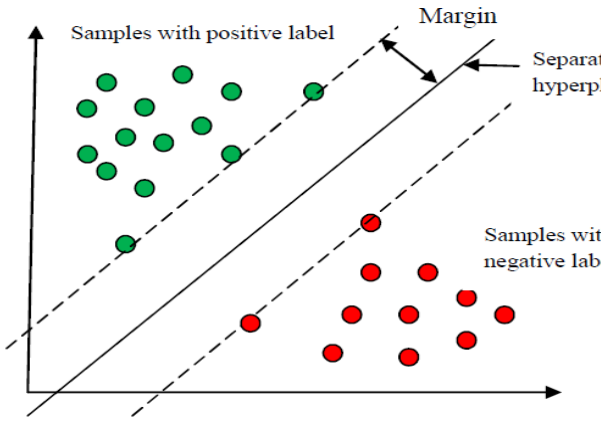

Fig.5. Maximum margin classifier

In the proposed method we are using linear classifier for classifier. Since our aim is to find is to find a best hyper plane that represents the largest separation or margin between the two classes we choose the hyperplane so that the distance from it to the nearest data point on each Side of hyperplane is maximized. If such a hyper plane exists, it is known as the maximum margin hyperplane and the linear classifier is defined as maximum classifier, which is shown in Fig. 2.

2) Random forest(RF): The random forest (RF) algorithms form a family of classification methods that are formed by combining decision trees (Fig.3).An particular important characteristic of such Ensembles of Classifiers is their decision tree components are grown from a certain amount of randomness. Based on this idea, RF is also defined as a generic principle of randomized ensembles of decision trees [19]. The basic unit of RF is a binary tree constructed using recursive partitioning .

The basic unit of RF tree is typically grown using the CART [20] methodology, in which binary splits recursively partition the tree into homogeneous or near homogeneous terminal nodes. According tom this method a good binary split must push data from a parent tree node to its two daughter nodes so that the ensuing homogeneity in the daughter nodes is improved from the parent node. RF is often a collection of hundreds to thousands of trees, where each tree is grown from original data by bootstrap sampling. RF trees differ from CART due to the fact that they are grown none deterministically using a two-stage randomization procedure. In addition to the randomization introduced by bootstrap sampling of the original data, a second layer of randomization is introduced at the node level when growing the tree.

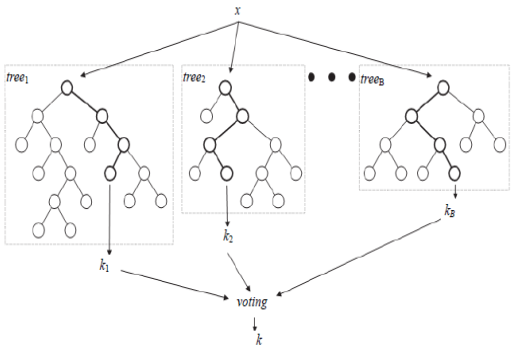

Fig.6. Random Forest Classifier

Rather than splitting a tree node using all variables (predictors), RF selects only a random subset of variables at each node and use them as candidates to find the best split for the node. The main aim of this two step randomization is to decorrelate decision trees so that the forest ensemble will have low variance. The Breiman's approach to build random forest generally consist of following main steps:

- Draw n-tree bootstrap samples from the original data.
- For each bootstrap data set grow a tree. At each node of the tree, randomly select m variables(predictors) for splitting. Continue growing the tree so that each terminal node has no fewer noes than nodesize cases.
- Aggregate information from the n-tree for classification.
- Using the data not in bootstrap sample compute an out-of-bag (OOB) error rate.

*3) k-Nearest Neighbor(k-NN):*The k-nearest neighbor algorithm (k-NN) proposed by Cover and Heart in 1968[4] is a non parametric method used for classification and regression. k-NN makes prediction from using training set directly. predictions are made by new vector for by searching through entire dataset for finding k most similar neighbours and summarizing the output of those k values. Incase of classification this might mode class value and for regression this might be mean output variable.

To determine which of k vectors in dataset are close to given input some kind of metrics is used. Normally for real valued data Euclidean distance is

widely apart from these hamming, manhattan, minkowski distance are also used. Euclidean is used when input data are of same type. Manhattan distance is used when inputs are not of similar data type. The computation complexity of k-NN increases with increase in dataset size. There also several other forms of k-NN namely instance based learner, lazzy learner, non-parametric learner. k-NN when used for classification the class with highest frequency from k similar instances is calculate as output. Class probabilities are calculated as normalized frequency of samples that belong to set of k class with similiarity.When number of class is odd choose k as an even number when number of class is even choose k as an odd number.

III. RESULTS

The CAD system is implemented in MATLAB 2015b and was validated using one of the largest publicly available database namely Lung Image Database Consortium image collection (LIDC-IDRI)[21]. The entire dataset contains CT images from a total of 1018 patients and the complete data along with annotated results can be downloaded from the website http://cancerimagingarchive.net.

Figures 4-13 depict results obtained from proposed method. Fig.8 show the lungs segmented from their background. Fig.10 represent the output of SVM similarly Fig.11 represent the output of k-NN and Fig.12 represent the output of RF classifier.

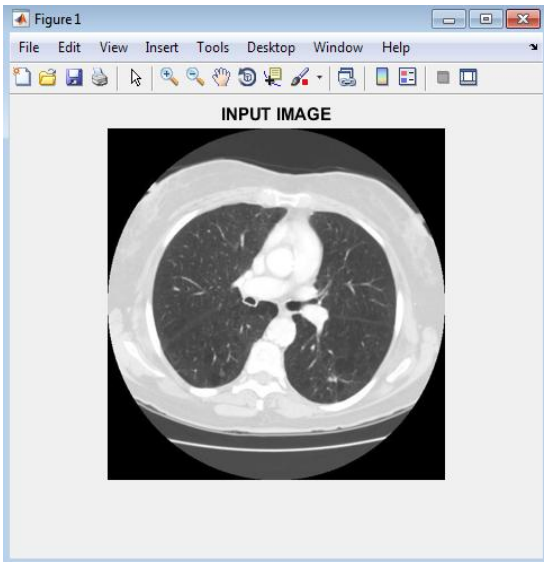

Fig. 7. Input CT image

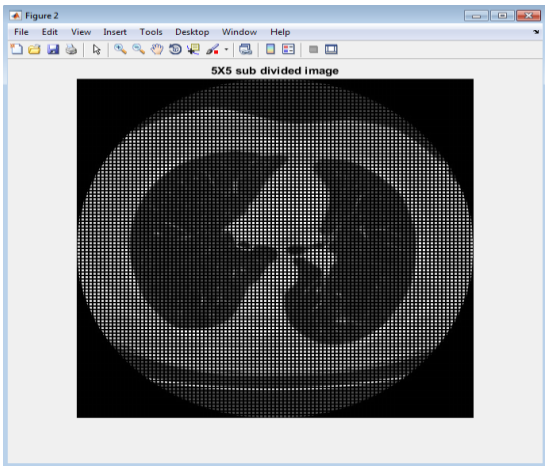

Fig.8. Input image divided into 5x5 blocks

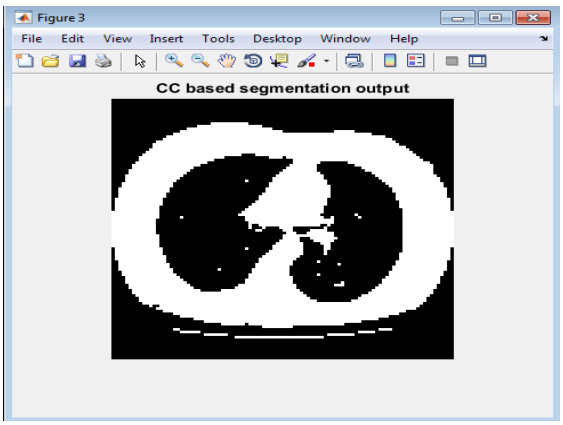

Fig. 9. Output of CC based segmentation

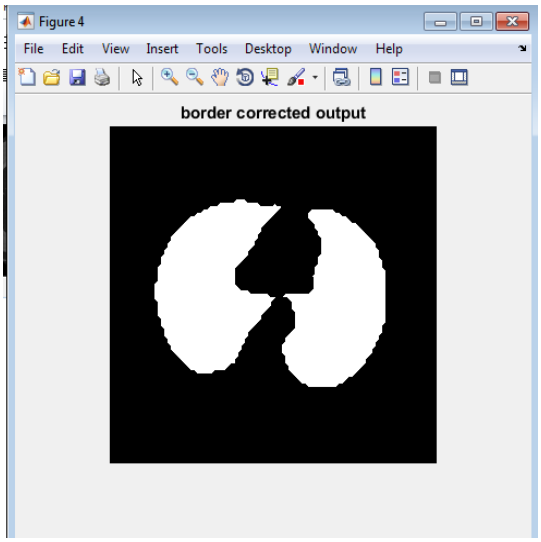

Fig. 10. Output with border connected

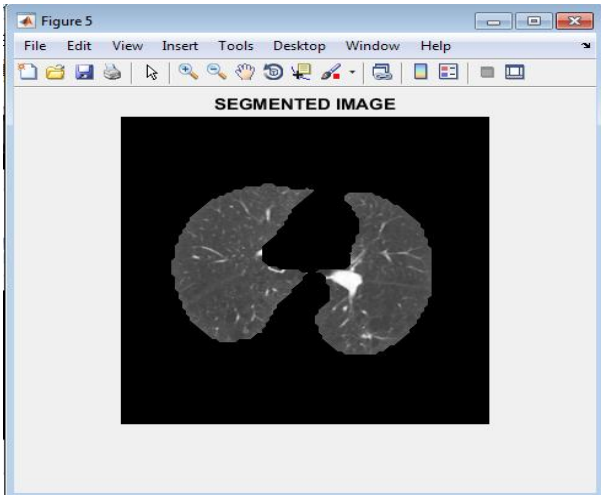

Fig. 11. Segmented Lungs

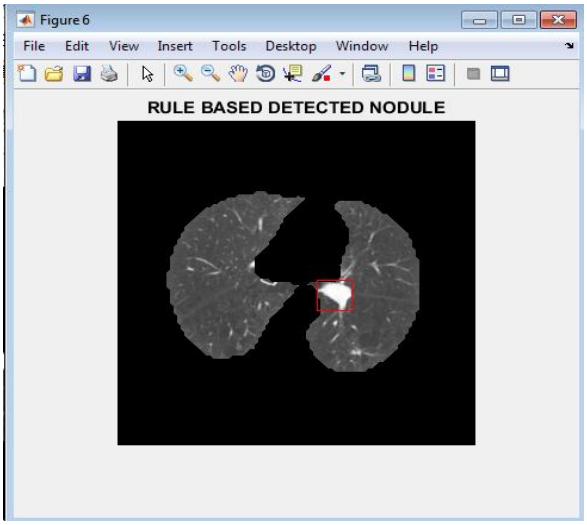

Fig.12 Detected nodule

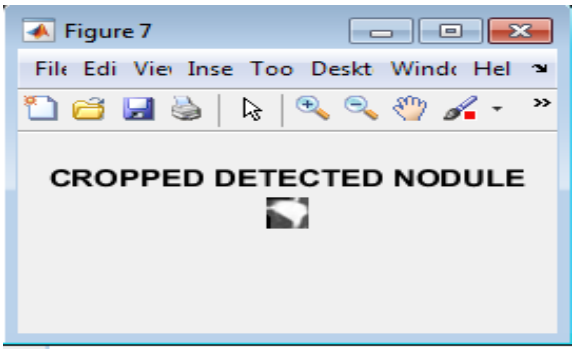

Fig. 13. Detected nodule region

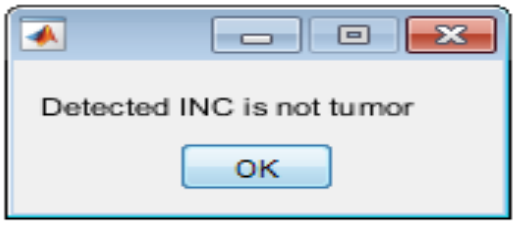

Fig. 14. Output of SVM classifier

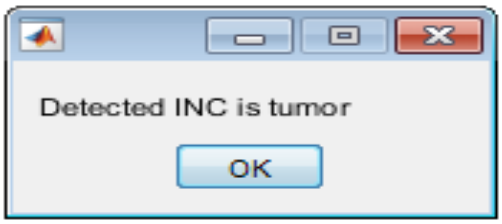

Fig. 15 Output of k-NN classifier

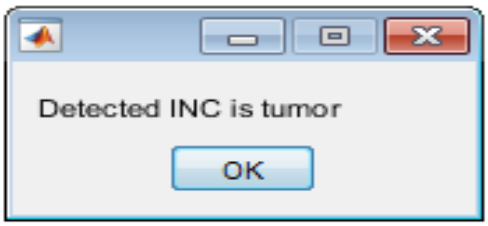

Fig. 16 Output Of Random Forest Classifier

TABLE I. PERFORMANCE METRICS

Metrics	Classifier		
	SVM	***RF***	***k-NN***
Accuracy	0.76	0.98	0.92
Sensitivity	0.825	0.975	0.95
Specificity	0.50	1	0.8
Precision	0.868	1	0.95
Recall	0.825	0.975	0.95
F_Measure	0.846	0.987	0.95
Gmean	0.642	0.987	0.872

Inorder to evaluate the preformance of different classifiers the the metrics accuracy, sensitivity, specificity, precision, recall, f_measure, gmean are calculated on the whole database and results are tabulated in Table I.

IV. CONCLUSION

In this paper a novel Computer-aided detection (CAD) system for classification of pulmonary

nodules in CT images is proposed. The proposed system uses contextual clustering based region growing for segmentation followed by GLCM and LBP features extraction. The extracted features are classified using three different classifiers. From performance metrics obtained it is found that Random Forest based classifier outperforms other classifiers.

REFERENCES

[1] B. S. Morse, Lecture 18: Segmentation (Region Based), 1998-2000.

[2] Giorgio De Nunzio, Eleonora Tommasi, Antonella Agrusti, Rosella Cataldo, Ivan De Mitri, Marco Favetta, Silvio Maglio, Andrea Massafra, Maurizio Quarta, Massimo Torsello, Ilaria Zecca, Roberto Bellotti, Sabina Tangaro, Piero Calvini, Niccolò Camarlinghi, Fabio Falaschi, Piergiorgio Cerello, and Piernicola Oliva, "Automatic Lung Segmentation in CT Images with Accurate Handling of the Hilar Region", Journal of digital imaging, Vol 24, No 1, pp 11-27, 2011.

[3] 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, pp. 580–589, Received 21 July 2010; revised 26 October 2010 accepted 8 February 2011.

[4] Eero Salli, Hannu, J, Aronen, Sauli Savolainen, Antti Korvenoja & Ari Visa 2001, 'Contextual Clustering for Analysis of Functional MRI Data', IEEE transactions on Medical Imaging, vol. 20, no. 5, pp.403-414, 2001.

[5] Linda G. Shapiro and George C. Stockman (2001): "Computer Vision", pp 279-325, New Jersey, Prentice-Hall, ISBN 0-13- 030796-3.

[6] Murat Ceylan, Yuksel Ozbay, O. Nuri Ucan, Erkan Yildirim, 2010, A novel method for lung segmentation on chest CT images: complex-valued artificial neural network with complex wavelet transform , Turk J Elec Eng & Comp Sci, Vol.18, No.4, pp. 613-623.

[7] Elaheh Soleymanpour, Hamid Reza Pourreza, Emad ansaripour and Mehri Sadooghi Yazdi, 2011, Fully Automatic Lung Segmentation and Rib Suppression Methods to Improve Nodule Detection in Chest Radiographs, JMSS, Vol. 1, No. 3, pp. 44-52.

[8] Faizal Khan, Z & Kannan, "Intelligent Segmentation of Medical images using Fuzzy Bitplane Thresholding", Measurement science and Review, Vol 14, No 2, pp-94-101, 2014.

[9] Faizal Khan, Z & Kavitha, V 2012, 'Estimation of objects in Computed Tomography Lung Images using Supervised Contextual Clustering', Research Journal of Applied Sciences, vol. 7, no. 9- 12, pp.494-499.

[10] Lin DT, Yan CR, Chen WT, "Autonomous Detection of Pulmonary Nodules on CT Images with a Neural Network-Based Fuzzy system", Comp Medical. Imaging and Graphics, Vol. 29, pp. 447-458, 2005.

[11] Prasad, M.N., Brown, M.S., Ahmad, S., Abtin, F., Allen, J., Da Costa, I., Kim, H.J., McNitt-Gray, M.F., Goldin, J.G. (2008). Automatic segmentation of lung parenchyma in the presence of diseases based on curvature of ribs. Academic Radiology, 15 (9), 1173-1180.

[12] Antonelli M, Lazzerini B, Marcelloni F "Segmentation and Reconstruction of the Lung Volume in CT Images". 20th annual ACM symposium on applied computing, vol I. Santa Fe, New Mexico, pp. 255-259, March 2005.

[13] JiantaoPu, Justus Roos, Chin A. Yi, Sandy Napel, Geoffrey D. Rubin, David S. Paik, "Adaptive Border Matching Algorithm: Automatic lung segmentation on chest CT images", Comp Medical Imaging and Graphics vol. 32, pp. 452-462, 2008.

[14] Ozekes S, Osman O, Ucan ON, "Nodule detection in a lung region that's segmented with using genetic cellular neural networks and 3D template matching with fuzzy rule based Thresholding", Korean Journal of Radiology, Vol. 9, pp. 1-9, 2008. [15] Cao Lei, Li Xiaojian, Zhan Jie,ChenWufan,

[15] "Automated Lung Segmentation Algorithm for CAD System of Thoracic CT", Journal of Medical Colleges of PLA, Volume 23,Issue 4, pp. 215-222, August 2008.

[16] Hyoungseop Kim, Seiji Mori, Yoshinori Itai, Seiji Ishikawa, Akiyoshi Yamamoto and Katsumi Nakamura, 2007, Automatic Detection of Ground-Glass Opacity Shadows by Three Characteristics on MDCT Images, World congress on medical physics and biomedical engineering 2006, IFMBE Pro2.

[17] Breiman, L. (2001) Random forests. Machine Learning Journal Paper, 45, 5-32.

[18] Wu, X.D. and Kumar, V. (2009) The top ten algorithm in data mining. Chapman & Hall/CRC, London.

[19] Biau, G., Devroye, L. and Lugosi, G. (2008) Consistency of random forests and other averaging classifiers. Journal of Machine Learning Research, 9, 2015-2033.

[20] Jeng-Shyang Pan, Shyi-Ming Chen, Ngoc Thanh Nguyen;November 2010;Computational Collective Intelligence. Technologies and Applications; Taiwan;Springer.

[21] S. G. Armato, G.McLennan, L. Bidautb,et.al"The lung image database consortium(LIDC) and image database resource initiative (IDRI): A completereference database of lung nodules on CT scans," Med. Phys., vol.38,pp.915–931,2011