ISSN (Online): 2456-5717

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol.3 Issue.5 May 2017

Analysis of Classification Methods for Diagnosis of Pulmonary Nodules in CT Images

[1]S.Santhosh Baboo, [2]E.Iyyapparaj
[1]Associate Professor, P.G & Research Department of Computer Science, Dwaraka Doss Goverdhan Doss Vaishnav College, Arumbakkam-106
[2]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 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; lung cancer; support vector machine(SVM)

I. INTRODUCTION

According to the recent statistics collected by National Cancer Registry Programmes(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 it's 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. Recent

This work by IJARBEST is licensed under a Creative Commons Attribution 4.0 International License. Available at https://www.ijarbest.com

techniques used for segmentation in literatures are statistical methods, include geometrical, structural, model based, signal processing methods, spatial domain filters, Fourier domain filtering. In [7] a novel approach to extract the lung region in chest X-ray images is presented which uses adaptive contrast equalization and non-linear filtering for image enhancement. Followed by preprocessing based on morphological operations an initial estimation of lung area is obtained and then it is improved by region approach to find the accurate final contour, then for rib suppression, oriented spatial Gabor filter is used. In [8] a new method for segmenting lung CT images by combining fuzzy logic with bitplanes was proposed to locate the region of interest which consists of following three steps, namely identification, rule firing, and inference. In this paper, CC along with the region growing algorithm have 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

Sometimes lung nodules may present on the lung parenchyma region. So if the lung region is not segmented completely the lung nodule on parenchyma border may get lost and accuracy gets decreased. With this aim it is very important to separate voxels belonging to lung region from voxels belonging to surrounding area.

In [15] using an optimal thresholding and mathematical morphological method, the authors first acquired the rough image of segmented lung. Finally, they used a fast self-fit segmentation refinement algorithm to adapt for the unsuccessful left-right lung segmentation in previous stage. In [6] an efficient method for segmentation denoted as Complex-Valued Artificial Neural Network with Complex Wavelet Transform (CWT-CVANN) is proposed. This architecture is made of two stages. The first stage extract features using different levels of complex wavelet transform followed by segmentation with complex-valued artificial neural network in second stage. In [13] authors introduced an novel geometric method for segmentation of lungs using novel Adaptive Border Marching algorithm. This work models the lung segmentation as a smoothing process of contours in continuous space and exist low computational cost.

Lin DT et al [10] proposed proposed a neural network-based fuzzy model to detect lung nodules present in the CT lung images. In their work segmentation is achieved by series of techniques including thresholding, median filtering, morphological closing, and labeling. In the next step features are extracted from ROIs which are fed into neural network based fuzzy model for classification. 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 Antonelli et al. [12] for segmenting the lung region. In their work, due to the presence of noise the background was not well eliminated well.

Ozekes et al [14] author proposed a four step process. In first step lung region of the CT images is segmented 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 voxel.

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. In the first step, the extraction of region of interest and preprocessing techniques such as labeling, shrinking and expansion is done in the CT. 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,Variable N-Quoit (VNQ) filter is used to extract suspicious shadows from GGO. The suspicious shadows are then classified into their appropriate classes using feature values calculated from the suspicious shadows.

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 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).

Segmentation Via Contextual Clustering With Region Growing

Region growing [1] is an iterative approach for segmentation. This technique involves identification of ROIs that are connected depending upon some predefined rule following some connectivity.

First an initial seed point is taken then the pixel is compared with its connective neighbors for some condition if the condition is satisfied it is added to region. The whole procedure is repeated until there is no more pixels to be added to the region.

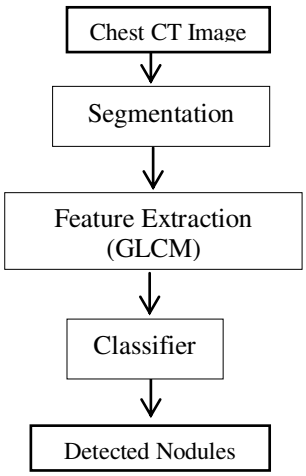

Fig. 1 Overall view of proposed method

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 in region growing is a 3x3 voxel in central slice and it was selected. Fuzzy rule is used to fix threshold in region growing.

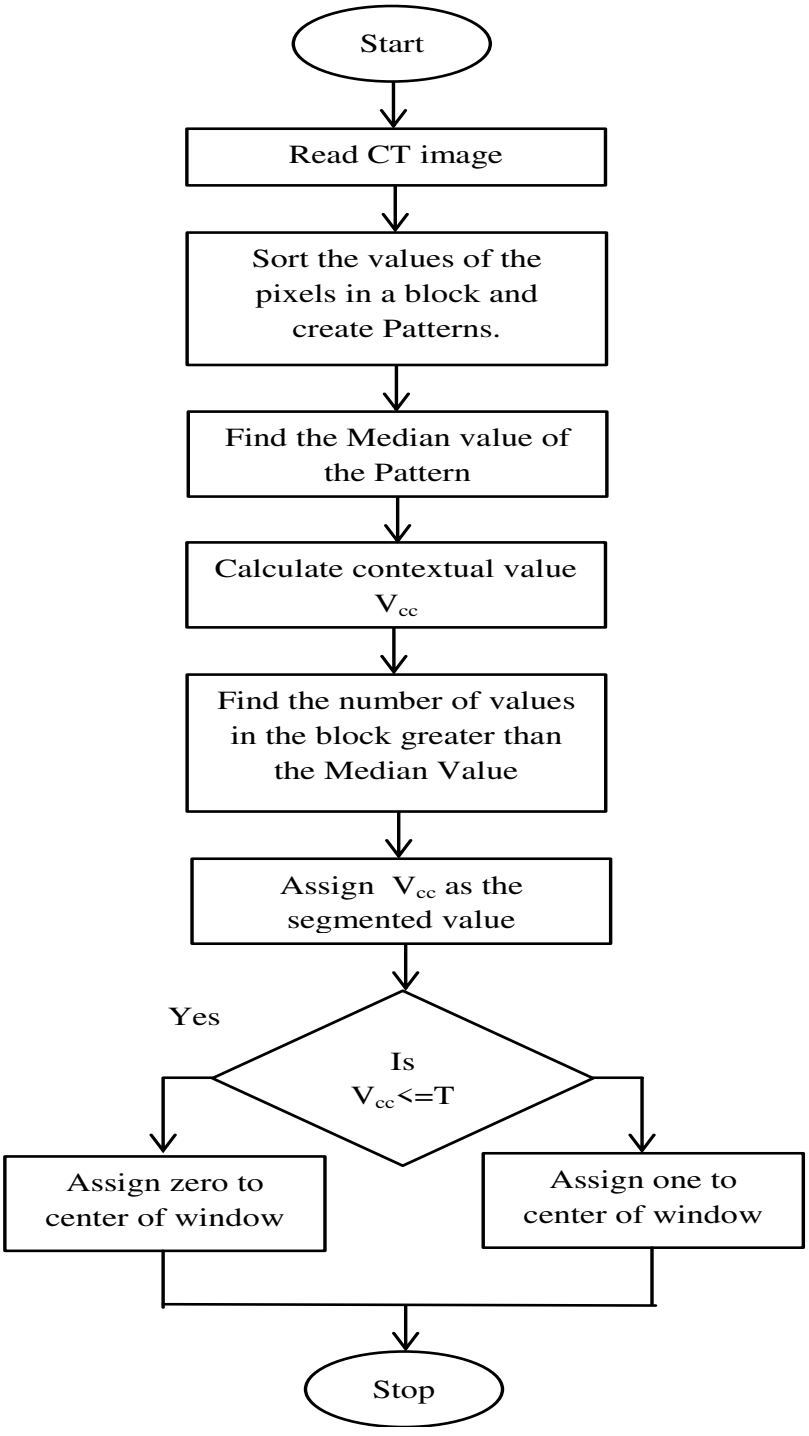

Fig. 2 Flow diagram of proposed algorithm

Recently, a lot of researchers use statistical clustering in image segmentation [3]. Contextual clustering [4,9] is a supervised algorithm for segmentation which uses spatial

information by considering number of activated neighbors of voxel and following this a rule is designed for clustering. The quality of segmented image in contextual clustering depends upon following four factors 1) A defined threshold value T provided by user for segmentation(T=140),2) a controlling parameter β in the range 0 to 1, 3) the median value of the all pixels in the pixel window 4) the total number of intensity values inside the window.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) Choose the decision parameter T,β first. Assume the neighbors to be 8-connected and set N to 8. Let V_i be intensity of a particular voxel.
2) For each voxel in image if $V_i > T$ label the voxel to ω_1 and store the result in variable G_1 , else label the voxel to ω_0 and store the result in variable G_0.
3) For each voxel 'i' compute the number of neighbors N_i who belongs to ω_1 using the variable G_1.
4) Re-label a voxel V_i to ω_1 if $V_i + \frac{\beta}{T}\left(N_i - \frac{N}{2}\right) < T$ else label it to ω_0.
5) Go back to step 3 and repeat the procedure until the labeling assigned in previous iteration equals the one got in current iteration.

The following figure describes flow of proposed algorithm.

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

IV. Extraction Of GLCM Features

Usually when texture features are extracted by texture analysis from image two types of approach may be followed one based on statistics and another based on structure. In this work we use statistical approach for texture analysis and extract Grey Level Co-occurrence Matrix(GLCM) features. GLCM also known as Grey Tone Spatial Dependency Matrix is a simple tabulation of how frequent different combination of gray levels occur in particular position with respect to other in the image. Depending upon number of gray levels in each combination statistics may be first-order, second-order or third-order etc. Third and higher order statistics considers the relation between three or more pixels and these statistics are theoretically possible but they are difficult to implement due long time taken for computation and interpolation. So GLCM uses second order statistics and this approach is used in large number of applications.

A GLCM is basically a square matrix with number of rows and columns equal to number of gray levels in the image. An element in GLCM matrix is denoted by P(i, j| Δx,Δy) which denotes the relative frequency that two pixels with intensity 'i','j' separated by distance of Δx,Δy in x &y direction lie within the given neighborhood. Similar to P matrix element M(i,

j | d, ɵ) denote the second order statistical probability for changes between gray levels 'i' and 'j' with respect to displacement distance (d) and angle (ɵ).

Using a large number of gray levels implies storing a lot of temporary data but when dimension of GLCM is very large they become more sensitive to the size of texture samples from which they are extracted. To eliminate with disadvantages total number of gray level is reduced to nominal one.

V. CLASSIFIER

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

1) SVM: After the features are extracted the next step is to classify whether tumor is present or not. Usually two types of approach is used for classification either supervised model or unsupervised one. In this work we use a supervised learning model for classification namely Support vector machine(SVM). Basically SVM models each feature vector of size N extracted as a point on a N-dimensional plane. Given such points on a N-dimensional plane the SVM tries to find a hyper-plane that best separates the point belonging to two classes. During testing the weights of hyperplane obtained is used to classify whether tumor is present or not. The feature vectors that nearer to hyperplane and satisfy their classification are called support vectors.

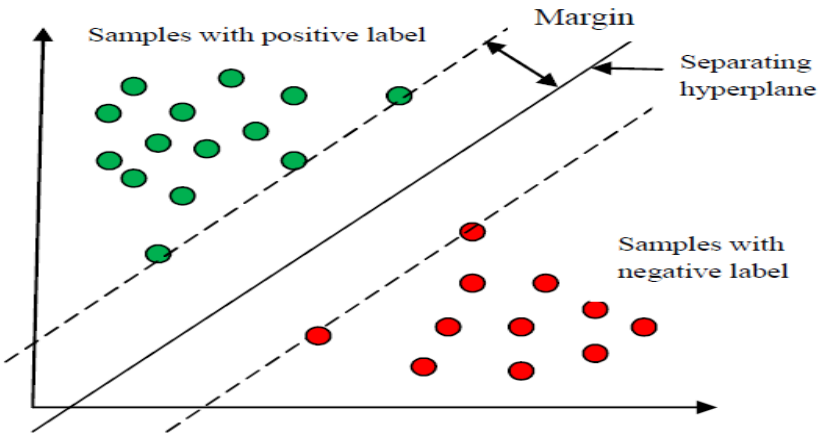

Fig. 3 Maximum margin classifier

In the proposed method we are using linear kernel for SVM classifier. Since our main target 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 such a linear classifier is defined as maximum classifier, which is shown in Fig. 3.

2) Random forest(RF): Random forests is one an ensemble learning method which is mainly used for the task of classification and regression. Random forest is created by combining several decision tree (Fig.4) by exploiting certain amount of randomness. Jeng et.al in [20] proposed CART method for building RF tree using binary split approach to create homogenous and near homogenous nodes.

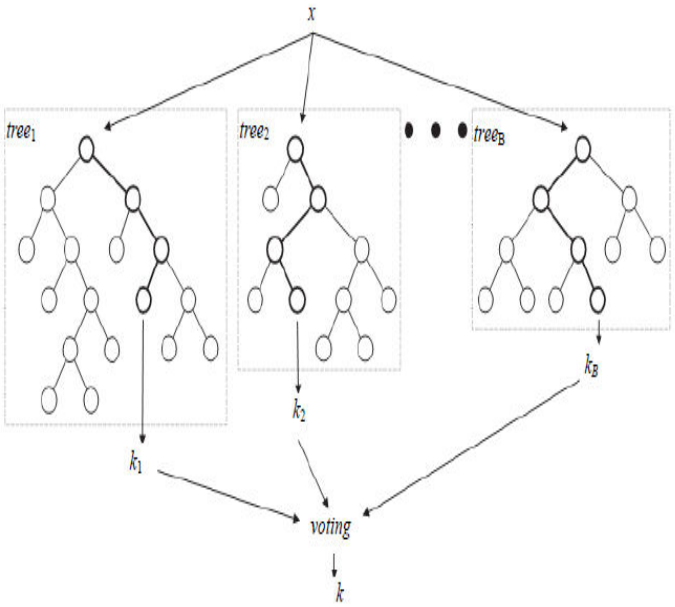

Fig. 4 Random Forest Classifier

According to this approach a good binary split must create a split such that data in the daughter nodes tends to be homogenous. Conventional random forest is very different from CART and is made of hundreds to thousands trees by bootstrap sampling of original samples. In addition to above mentioned difference another important difference is that a RF tree is built by a two-stage randomization procedure. In the first stage randomization bootstrap sampling of original sample is used. In second stage randomization rather than splitting a tree node using all predictors, only a random subset of predictors are selected at each node and used as candidates to find the best split for the node. This two stage randomization results in decorrelation of decision trees giving a low variance for whole forest ensemble. 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.

VI. RESULTS

The proposed 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.

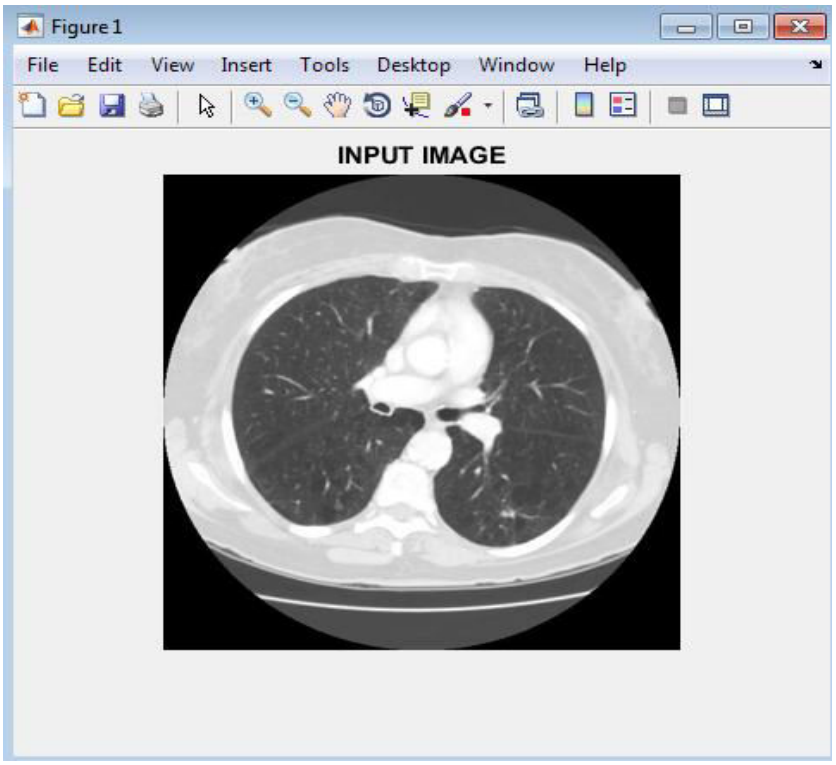

Fig. 5 Input CT image

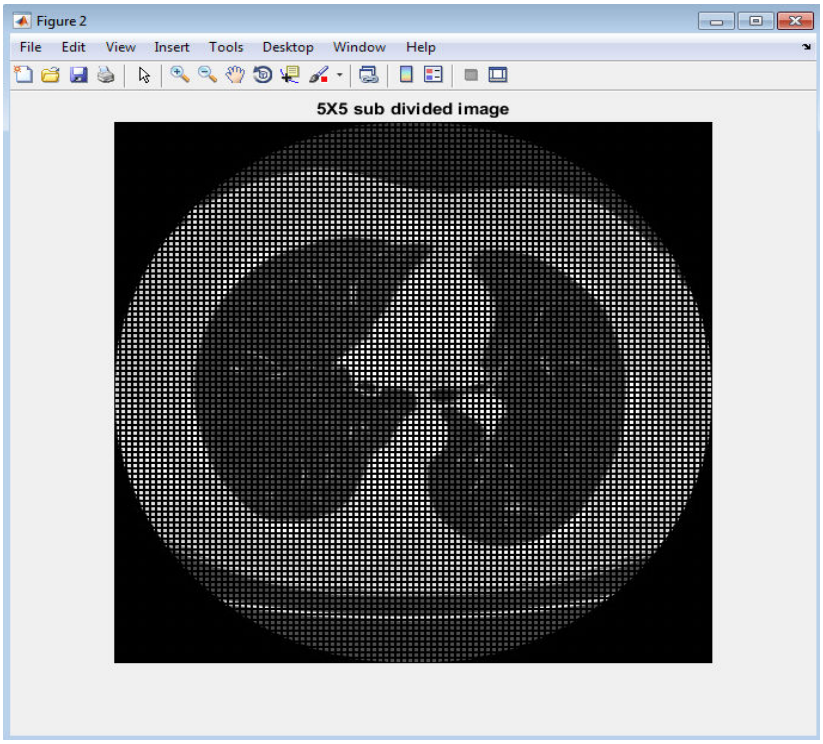

Fig. 6 Input image divided into 5x5 blocks

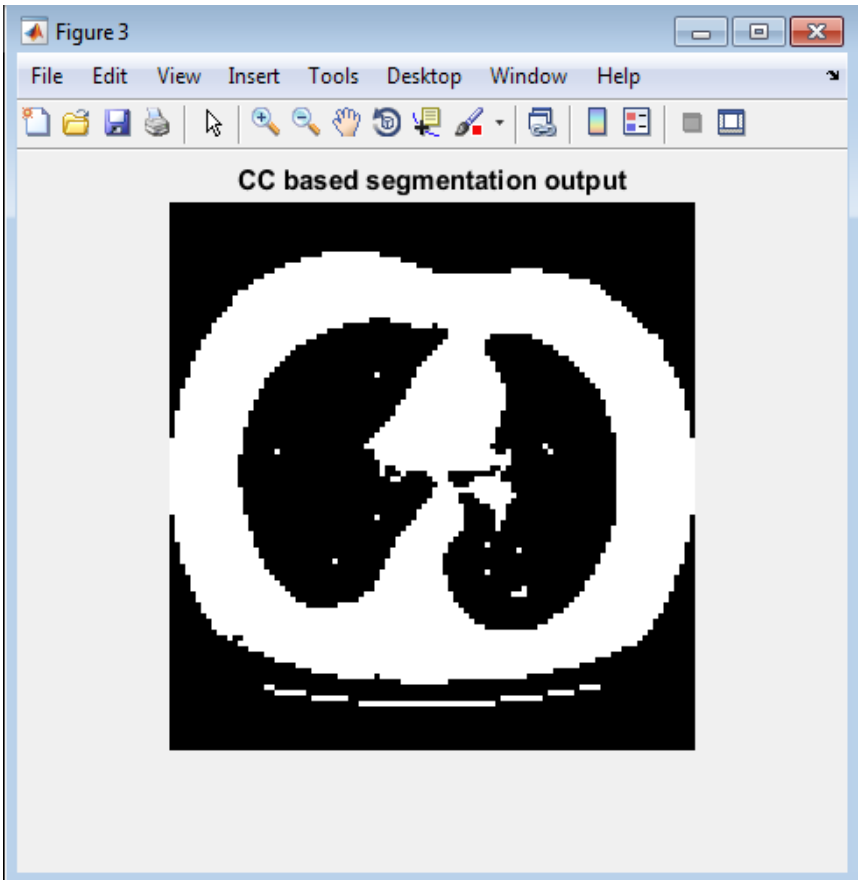

Fig. 7 Output of CC based segmentation

ISSN (Online): 2456-5717

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol.3 Issue.5 May 2017

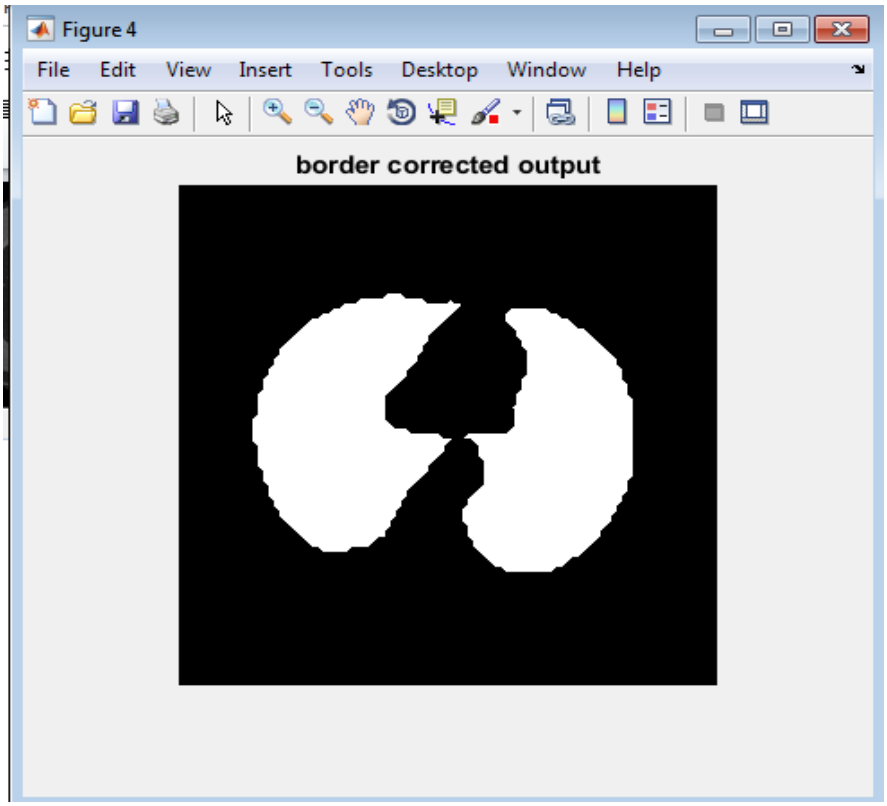

Fig. 8 Output with border connected

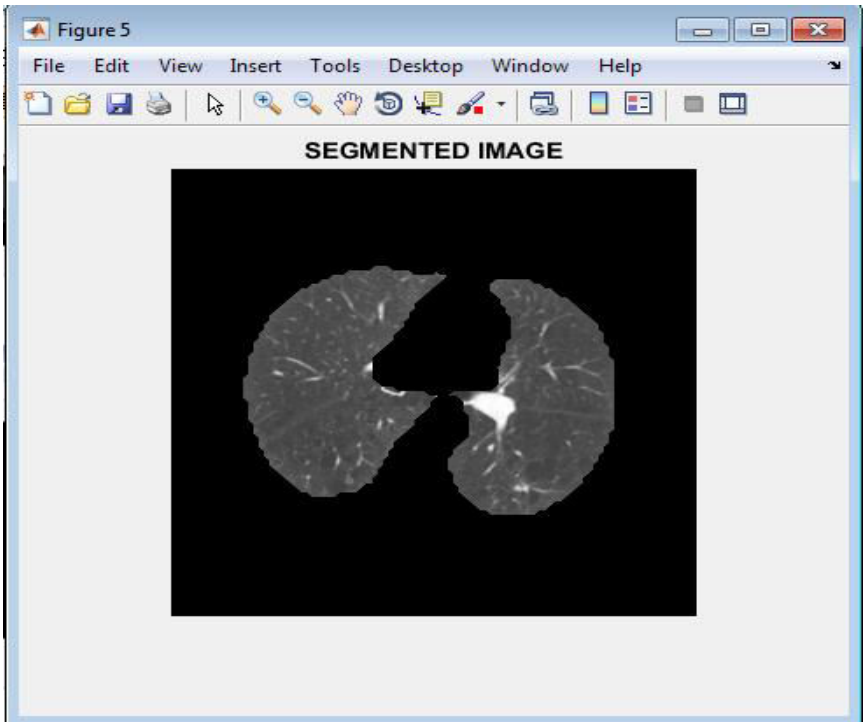

Fig. 9 Segmented Lungs

ISSN (Online): 2456-5717

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol.3 Issue.5 May 2017

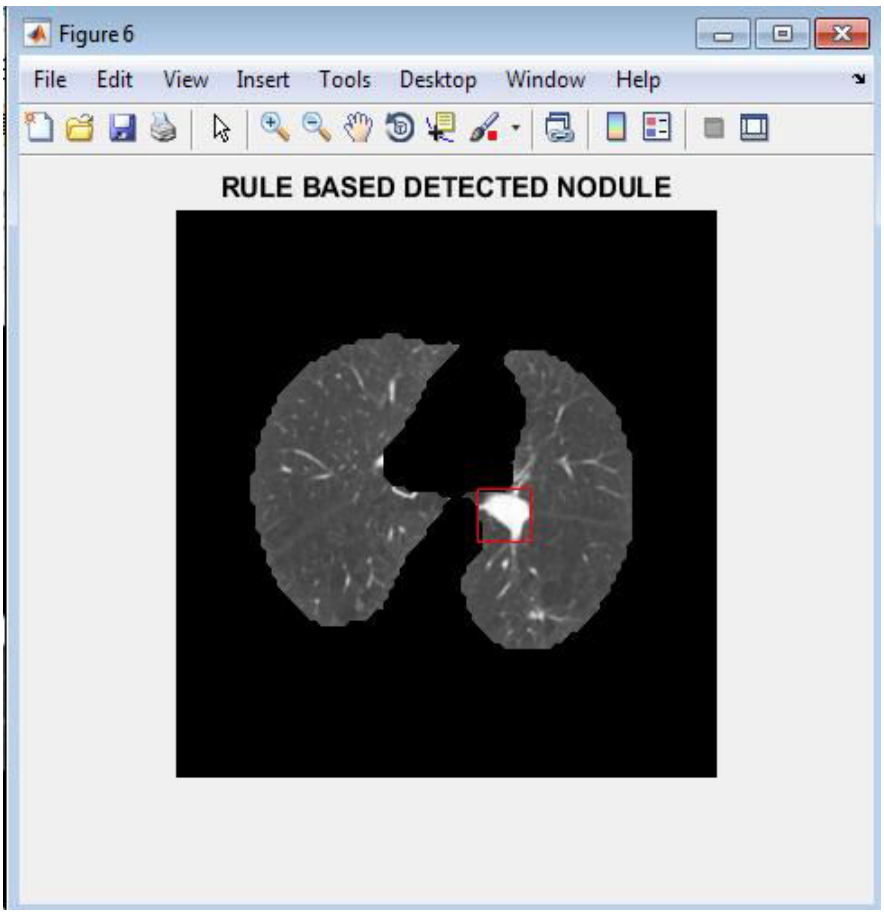

Fig. 10 Detected Nodule

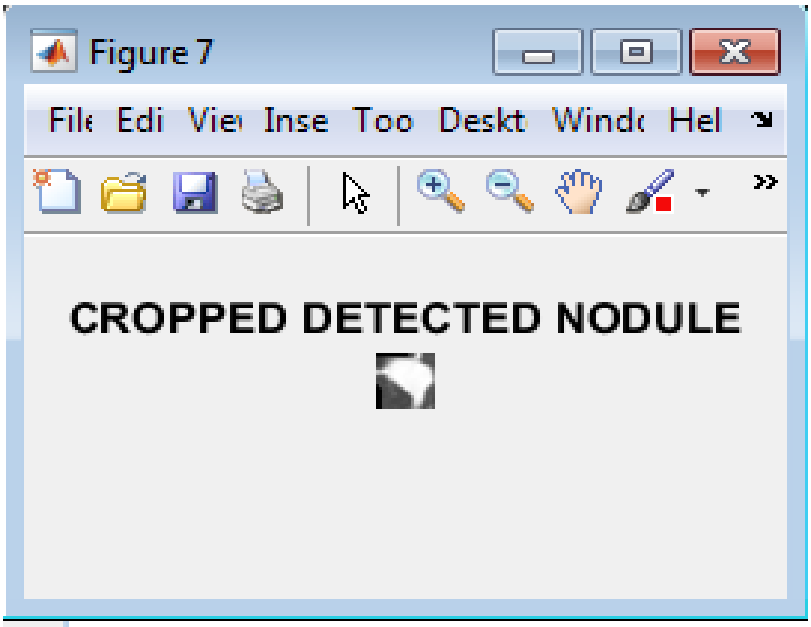

Fig. 11 Detected nodule region

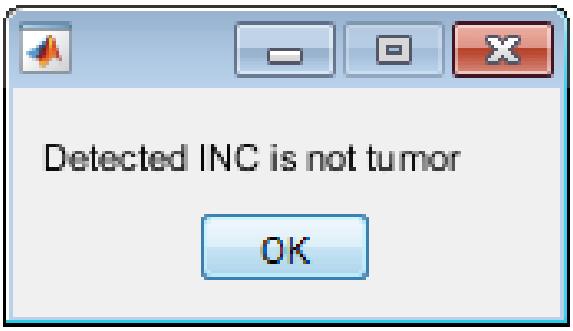

Fig. 12 Output of SVM classifier

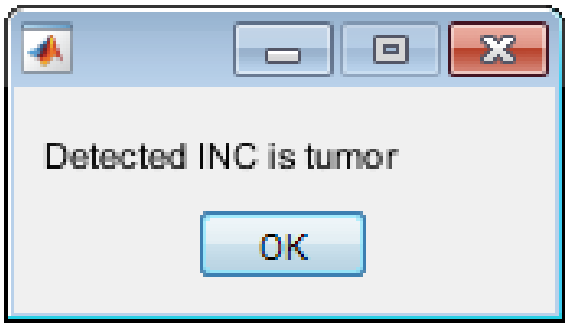

Fig. 23 Output of k-NN classifier

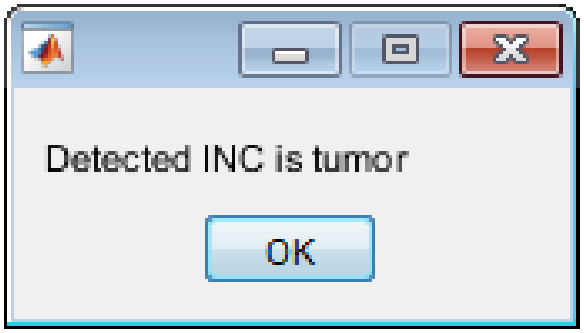

Fig. 34 Output of Random Forest Classifier

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.

VII. 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 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. 915–931,20.