International Journal of Advanced Engineering and Technology
ISSN: 2456-7655; Impact Factor: RJIF 5.54
Received: 01-06-2019; Accepted: 05-07-2019
www.newengineeringjournal.com
Volume 3; Issue 3; July 2019; Page No. 72-76

Detection of lung nodules in CT images using random forest classifier

M Mary Adline Priya[1], Dr. S Joseph Jawhar[2]
[1] Research Scholar, Faculty of Information and Communication Engineering, Arunachala College of Engineering for Women, Kanya Kumari, India
[2] Professor, Faculty of Electrical and Electronics Engineering, Arunachala College of Engineering for Women, Kanya Kumari, India

Abstract
The classification and identification of the disease in medical images were helpful in biomedical applications. Lung nodule detection becomes the crucial part in the lung cancer diagnosis. The accurate segmentation of lung nodules from computerized tomography scans is important for lung cancer diagnosis and research. 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: feature extraction, computed tomography, lung cancer, lung segmentation, computer aided diagnosis, random forest, SVM, k-NN

Introduction
Lung Cancer is one of the most serious human body problems in the world. The death rate of lung cancer is highest of all other types of cancer. Up to 10 million patients in the world will die of lung cancer by 2030 in terms of the report from the World Health Organization. The survival rate of lung cancer is very smallest among all types of cancer. So, there is a need to design a computational intelligence based approaches to detect the lung cancer because the survival from lung cancer is directly related to its growth at its detection time. If we detect lung cancer at early stage, then there are more possibilities to survive the patients.
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 lose 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 surrounding anatomy. With the hypothesis that deep analysis of radiographic images can inform and quantify the microenvironment and the extent of intra-tumoral heterogeneity for personalized medicine, analysis of large numbers of image features extracted from computed tomography (CT) with high throughput can capture spatial and temporal genetic heterogeneity in a non-invasive way, which is better than invasive biopsy, based molecular assays. It will be useful for medical research, computer-aided diagnosis, radiotherapy and evaluations of surgery outcome as well. For this purpose, accurate segmentation of lung nodules is the pre-requisite.

Related Works
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.
Lee, Kouzani, and Hu (2010) used a Random Forest based classification aided by clustering for detection of lung nodule. This method used a hybrid random forest algorithm to classify the lung nodules and non-nodules. The images were obtained from 32 patient scans in the LIDC-IRDI. The authors achieved 98.33% sensitivity and 97.11% specificity. The clustering techniques chosen were k-means and expectation maximization algorithms, which require parameterization. The number of clusters used was 2 in the classification aided by clustering method.
Namin, Moghaddam, Jafari, Esmaeil-Zadeh, and Gity (2010) present a methodology with two steps, segmentation and classification of lung nodules in malignant and benign using fuzzy k-nearest neighbor (FKNN). FKNN is used in both steps. The extracted features are textures based on the intensities of the voxels (Hounsfield units) and the geometry. Accuracy of 88% and a mean of 10.3 false positives per exam were reported. The author also presented that the nodules with small size and/or irregular shapes were the main shortcomings of the segmentation method.
Tartar, Kilic, and Akan (2013) suggest an approach that

classifies lung nodules in CT scans using hybrid features. Four techniques are explored: principal component analysis (PCA) and minimum redundancy maximum relevance (mrMR), statistical features extraction, geometrical features extraction and an hybrid method. They evaluated statistical values from PCA in two dimensions and mrMR with geometrical features. The results were 90.7% accuracy, 89.6% sensitivity and 87.5% specificity. The methodology adopts a dataset of 2D CT slices evaluated by radiologists, divided in 95 nodules and 75 non-nodules from 63 patients. The uses of two dimensional images for shape-based features extraction may raises the sensitivity, but prejudice the specificity of the third and fourth method because of the limited analysis of the whole nodule property on the original 3D CT.

Zhang *et al*. (2013) proposed an approach to classify lung nodules in four categories: well-circumscribed, vascularized, juxtapleural and pleural tail. The method describes both the lung nodule and the surrounding context information to perform the characterization in two distinct ways: 1) super pixel labeling, which labels the pixels as foreground or background, and 2) context curve calculation. The image database resource was the early lung cancer action program (ELCAP), which contains 50 low-dose scans. This method resulted in 82.5% accuracy. Their database contained 379 unduplicated lung nodules with locations indicated by annotations. The nodules images were cropped with a 31 x 31 pixels window for extracting the context curves. As super pixel labeling results in texture characteristics, there is no guarantee that the nodule image sliced comprises all fundamental features of the nodule texture.

The methodology developed by Choi and Choi (2014) uses Hessian matrices to calculate two angular histograms. The angular histograms of the normal surface are used as features to eliminate structures linked to lung parenchyma and to classify the remainder as nodules or non-nodules. Using 84 exams from the LIDC, they achieved 97.4% accuracy, 97.2% sensitivity and 97.7% specificity. The method multi-scale dot enhancement filtering is applied into 3D shape of the object, but lacks of pre-processing techniques. It was used 84 exams with a total of 148 isolated, juxtapleural and juxtavascular nodules.

Carvalho Filho *et al*. (2014), proposed the classification of lung nodules and non-nodules using taxonomic diversity indices. The computation of the indices was based on phylogenetic trees, which were applied to the characterization of the nodule candidate. The SVM classifier and the radial-basis function (RBF) kernel were used. The LIDC (833 scans) resulted in 97.55% mean accuracy, 85.91% mean sensitivity and 97.7% mean specificity.

In this work, a new methodology is proposed that overcomes those inconveniences. The majority of those works, with the exception of Carvalho Filho *et al*. (2014), Shen *et al*. (2015) and Tan *et al*. (2011), utilized few exams for their study cases. This work takes the advantage of use a relevant number of exams, inspired by Carvalho Filho *et al*. (2014) work. Shen *et al*. (2015) chain-code approach is very sensitive to noises, differently from the proposed method of Artificial Crawlers, which is low sensitive to them, due to its evolutionary properties.

Methodology

Block diagram of proposed method:

- The main goal of this work is to construct a consistent methodology for the classification of candidates into nodules or non-nodules.
- The proposed method involves three stages are shown in figure1.

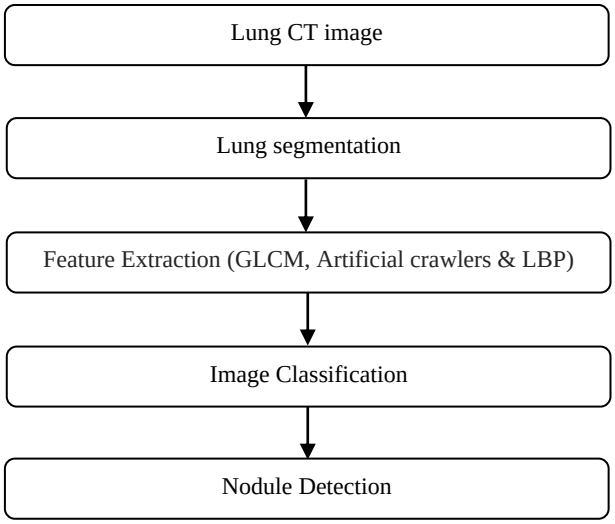

Fig 1: Block diagram of proposed work

- There are four phases: image acquisition, feature extraction, image classification and nodule detection.
- 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 Neighbour (k-NN), Random forest (RF) and Support vector machines (SVM).
- The last phase is the nodule detection.

A. Segmentation via Contextual Clustering with Region Growing

Region growing 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. The first step in region growing is to choose initial seed point. In this 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.

B. Extraction of GLCM Features

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. A GLCM is basically a matrix where the number of gray levels in image equals the number of rows and columns. 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.

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 centre 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.

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.

In the proposed method linear classifier is used for classification. Since the aim is to find a best hyper plane that represents the largest separation or margin between the two classes we choose the hyper plane so that the distance from it to the nearest data point on each side of hyper plane is maximized. If such a hyper plane exists, it is known as the maximum margin hyper plane and the linear classifier is defined as maximum classifier.

2) Random forest (RF): The random forest (RF) algorithms form a family of classification methods that are formed by combining decision trees. A 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 tree. 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 methodology, in which binary splits recursively partition the tree into homogeneous or near homogeneous terminal nodes. According to 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. Rather than splitting a tree node using all variables (predictors), RF selects only a random subset of variables at each node and uses 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 approaches 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 nodes than node size 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 Neighbour (k-NN): The k-nearest neighbour algorithm (k-NN) proposed by Cover and Heart (1968) 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. In case 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 is 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, and nonparametric 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 similarity. When number of class is odd choose k as an even number when number of class is even chooses k as an odd number.

Results and Discussion

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). 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 2 to 4 depict results obtained from proposed method. Figure5 shows the lungs segmented from their background.

Fig 2: Input image

Fig 3: Output of CC based segmentation

Fig 4: Border corrected output

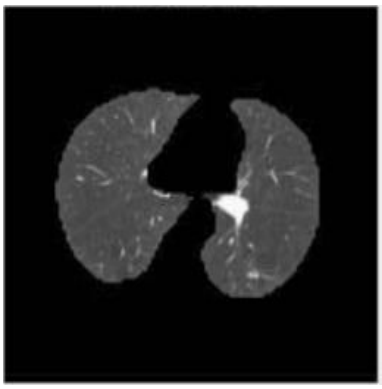

Fig 5: Segmented lungs

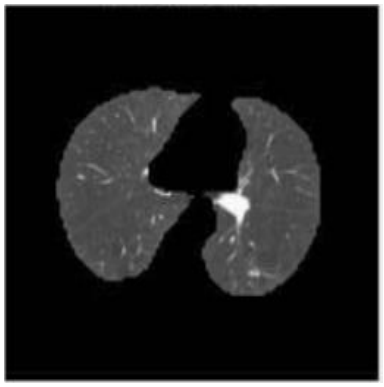

Fig 6: Detected nodule

Table 1: 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

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

Conclusion

In this paper a novel Computer-aided detection (CAD) system for classification of lung 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. Rebecca L Siegel, Kimberly D Miller, Ahmedin Jemal. Cancer statistics, 2017", CA Cancer J Clin, Jan. 2017.
2. Organization WH. Description of the global burden of NCDs their risk factors and determinants", Geneva Switzerland: World Health Organization, 2011.
3. Vigneshwar R, Karthic K, Karthick A, Senthamizhselvi R. Lung Lesion Extraction Using Improved Toboggan Based Algorithm. International Journal of Advance Research, Ideas and Innovations in Technology. 2017; 3(2):774-779.
4. Suvadip Mukherjeey, Xiaojie Huang, Roshni R Bhagalia. Lung nodule segmentation using deep learned prior based graph cut, *IEEE Trans. Image Process.* 2017; 20(7):1205-1208.

5. Xia Li, Yang Xiong, Song Jia. Pulmonary Nodules Detection Algorithm based on Robust Cascade Classifier for CT Images, IEEE Trans. Image Process. 2017; 20(7):231-235.
6. Setio Traverso, Bel Berens, Bogaard Cerello, Chen Dou, Fantacci Bram *et al*. Validation, comparison, and combination of algorithms for Automatic detection of Pulmonary nodules in CT images: the LUNA 16 challenge. Medical Image Analysis, 2017; pp. 1-19.
7. Netto Silva, Lopes Paiva, Nunes Gattass. Statistical tools for the Temporal Analysis and classification of Lung Lesions. Computer Methods and Programs in Biomedicine. 2017; p. 1-20.
8. Hassen DB, Taleb H. Automatic detection of lesions in lung regions that are segmented using spatial relations. Clinical Imaging, 2013; 37:498-503.
9. Wang L, Lin H, Huang X, Wang B, Chen Y. A 3D segmentation and visualization scheme for solid and non-solid lung lesions based on Gaussian filtering regularized level set. International conference on 3D vision. 2014; pp.67-73.
10. Santos AM, de C Filho AO, Silva AC, de Paiva AC, Nunes RA, Gattass M. Automatic detection of small lung nodules in 3D CT data using Gaussian mixture models, Tsallis entrophy and SVM. Engg. Applications of AI, 2014; 36:27-39.
11. Rossi F, Abd Rahni AA. Combination of low level processing and active contour techniques for semi-automated volumetric lung lesion segmentation from thoracic CT images. IEEE Trans. On Biomedical engineering and sciences, 2015; 15:26-30.
12. Lim J, Seelan L, Padma Suresh, Veni SHK. Automatic extraction of lung lesion by using optimized toboggan based approach with feature normalization and transfer learning methods. IEEE Trans. on emerging tech. trends, 2016; l(16):18-27.
13. Yosefina Finsensia Riti, Hanung Adi Nugroho, Sunu Wibirama, Budi Windarta, Lina Choridah. Feature extraction for lesion margin characteristic classification from CT scan lungs Image, IEEE Trans. on 2016 1st International Conference on Information Technology, Information Systems and Electrical Engineering, 2016; 16:54-58.
14. Campos DM, Simões A, Ramos I, Campilho A. Feature-Based Supervised Lung Nodule Segmentation. 2014; Ci:23-26.
15. Diciotti S, Lombardo S, Falchini M, Picozzi G, Mascalchi M. Automated segmentation refinement of small lung nodules in CT scans by local shape analysis. IEEE Trans. Biomed. Eng. 2011; 58(12):3418-3428.
16. Candemir S, Jaeger S, Palaniappan K, Musco JP, Singh RK, Xue Z, *et al*. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. IEEE Trans. Med. Imaging, 2014; 33(2):577-590.
17. Farag AA, El Munim HEA, Graham JH. A novel approach for lung nodules segmentation in chest CT using level sets. IEEE Trans. Image Process. 2013; 22(12):5202-5213.
18. Sun S, Guo Y, Guan Y, Ren H. Juxta-Vascular Nodule Segmentation Based on the Flowing Entropy and Geodesic Distance Feature. Scientia Sinica (Informationis). 2013; 61:1136-1146.
19. Gu Y, Kumar V, Hall LO, Goldgof DB, Li CY, Korn R, *et al*. Automated delineation of lung tumors from CT images using a single click ensemble segmentation approach. Pattern Recognit. 2013; 46(3):692-702.
20. Bendtsen C, Kietzmann M, Korn R, Mozley PD, Schmidt G, Binnig G. X-Ray computed tomography: Semi automated volumetric analysis of late-stage lung tumors as a basis for response assessments. Int. J. Biomed. Imaging, 2011.
21. Kubota T, Jerebko AK, Dewan M, Salganicoff M, Krishnan A. Segmentation of pulmonary nodules of various densities with morphological approaches and convexity models. Med. Image Anal. 2011; 15(1):133-154.
22. Mansoor A, Bagci U, Xu Z, Foster B, Olivier KN, Elinoff JM. A generic approach to pathological lung segmentation. IEEE Trans Med Imaging, 2014; 33:2293-2310.
23. World Health Organization. World Cancer Report 2017, Lyon: International Agency for Research on Cancer, 2017.