# Detection of Pulmonary Nodules in Thoracic CT images using Image Processing and Data Mining Techniques

R. Geetha Ramani, PhD
Department of IST
CEG, Anna University
Chennai, India

R. Keerthana
Department of IST
CEG, Anna University
Chennai, India

B. Lakshmi
Department of IST
CEG, Anna University
Chennai, India

## ABSTRACT
Pulmonary nodules are soft tissue masses present in the lung that can be seen in thoracic CT images. When the size of the nodules is large, immediate attention is required. Automatic detection of these large nodules would be appreciated by the medical professionals as it reduces labor and time. The lung nodules are usually extracted through image processing and classification techniques. In this work, lung nodules are detected through image clustering followed by classification of candidate nodules as either nodules or non-nodules. K-Means clustering is performed to delineate the candidate nodules. The features of the candidate nodules are extracted and fed as input to bagged random tree classifier for classification of nodules as true or not. The proposed system was tested on ELCAP dataset and an accuracy of 95.31% was achieved. The system has a social cause.

## General Terms
Data Mining, Medical image analysis, Pulmonary nodules

## Keywords
Lung nodules, K-Means Clustering, Classification, Random Tree, Data Mining, Image Processing

## 1. INTRODUCTION
Computational techniques have been highly used in medical image analysis [1]. Medical images expose the organs of the human body, which can be analyzed to detect any abnormality if present. Lung nodules are a soft tissue mass located in the lungs. These can be detected by radiography imaging techniques namely MRI, CT etc. Computed Tomography (CT) [2] images are 3D scans, formed from long series of 2D images. Lung nodules can be detected by analyzing these CT scan images. Pulmonary (lung) nodules [3] do not reveal any symptoms during its early stages. They are incidently diagnosed while analyzing the scans obtained for some other disorder.

The lung nodules can be of varying size. If the size of the nodule is very small, then the chance that they are non-cancerous is high and the practioners wait to see the progress. If the size of the nodule is large, then immediate attention is required to continue with further analysis clinically. A sample of CT image that reveals the large lung nodule (annotated) is shown in Figure 1.

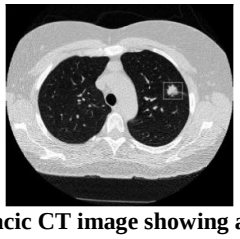

**Figure 1: Thoracic CT image showing annotated large nodule**

Image processing techniques [4] and data mining techniques [5] have been prominently utilized for automatic identification of lung nodules. In this work, image clustering and classification of candidate components is attempted to identify large nodules in thoracic CT images.

The rest of the paper is organized as follows: Section 2 presents the literature survey on existing methods to segment the nodules, section 3 highlights the proposed methodology to segment the large nodules, Section 4 discusses the results and Section 5 concludes the paper.

## 2. LITERATURE SURVEY
Detection of large lung nodules has been an interesting area of research. The existing methods in extraction of lung nodules from thoracic CT images are concisely presented here.

In 2009, Hong et. al. [6] utilized adaptive Thresholding and SVM classification to eliminate false positives. The system achieved an 89.47% with 11.9% false positive per case when tested with 44 solitary pulmonary nodules.

In 2010, Messay et. al. [7] presented a CAD system using thresholding, morphological processing and Fisher Linear Discriminant to segment, detect candidate nodules and eliminate false positives. The system obtained a sensitivity of 82.66 percentage with 3 FP per case being validated with 143 nodules (juxta vascular, solitary, ground-glass opacity and juxtapleural), with sizes from 3mm to 30mm. Again in the same year, Gomathi et. al. [8] used image processing techniques, Fuzzy C-Mean algorithm and neural classifier in the stages of preprocessing, segmentation and nodules detection respectively. This system yielded an efficiency of 76.9% and false positives being validated with 13 nodules and 8 nodules were less than 2 mm size.

In 2011, Amal Farag et. al. [9] proposed the lung nodule detection through extraction of geometric feature extraction and classification in low dose CT. Geometric features such as SIFT, LBP and SURF were extracted and K-NN classification was performed. The methodology yielded a sensitivity of 85% on ELCAP dataset. In 2012, Cascio et. al. [10] made use of a neural classifier, region growing technique with morphological filter and Mass-spring models to eliminate false nodules. The system achieved a performance of 97% with 6.1% FP per case being validated with 148 internal and juxtapleural nodules.

In 2015, Suganya et. al. [11] presented a survey on classification techniques of lung nodules. Generally, thresholding and Robust segmentation techniques are used in the segmentation process. Then, feature set containing MR8 (Maximum Response) +LBP (local Binary Patterns), Sift descriptor and MHOG (Multiorientation Histogram of Oriented Gradients) are used for SVM classification on ELCAP public database. In 2016, Mayuri et. al. [12] proposed

lung nodule detection through concentric level partition, feature extraction and classification. Concentric level partition was constructed by an improved quick shift superpixel formulation, FS3 feature set included SIFT, MR8+LBP and multi orientation HOG was generated. Support vector machine classifier and Probabilistic latent semantic classifier were designed to classify the lung nodule type. This methodology was done by ELCAP public database with improved performance accuracy of 81.45%.

From the previous studies, the lung nodule detection has been carried out using image pre-processing, thresholding, false nodule detection and classification. The next section highlights the proposed methodology.

## 3. PROPOSED METHODOLOGY

Automatic lung nodule detection is attempted in this work. The proposed framework consists of image pre-processing phase, lung segmentation and nodule detection, feature extraction and classification as nodules and non-nodules. Each step is explained in detail in the following paragraphs. The proposed framework is depicted in Figure 2.

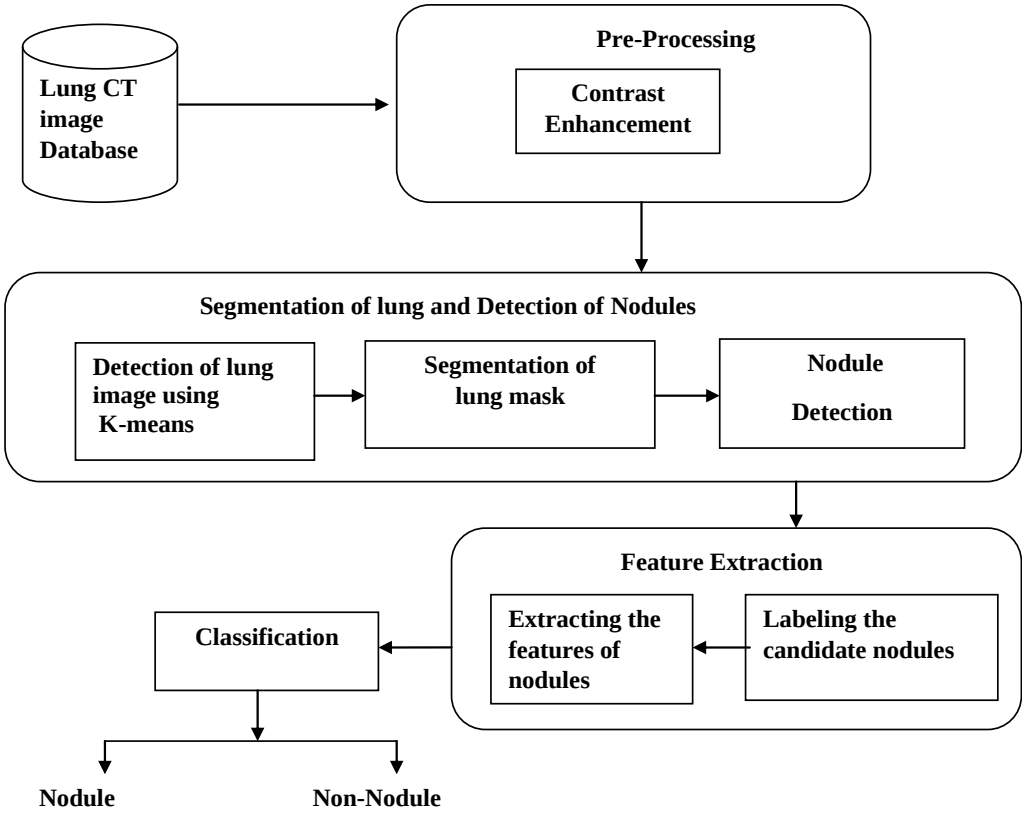


**Figure 2: Proposed framework for lung nodule detection**

Image pre-processing is used to improve the quality of the image. Generally, the quality of image is affected by different artifacts due to non-uniform intensity, variations, motions, shift, and noise. Thus, the pre-processing of image aims at selectively removing the redundancy present in scanned images without affecting the details, the quality of the image is enhanced by improving the contrast of the image through Contrast Limited Adaptive Histogram Equalization (CLAHE) [13]. Contrast enhancement is performed to improve the quality of the image and reveal the structures more clearly.

Then, K-Means Clustering [14] is performed on the contrast enhanced image. K-Means clustering is one of the most popularly used unsupervised clustering algorithms to group similar entities together. K-Means utilizes Euclidean distance metric to group similar entities such that the intra cluster distance is low and inter cluster distance is high. It requires the number of clusters as input. The number of clusters is set to 3 in this work. The first cluster obtained contained the lungs. The second and third cluster obtained consists of the nodules. Hence, the second and the third cluster were grouped. The grouped cluster reveals the lung region and the nodules. The sample of K-Means output is shown in Figure 3.

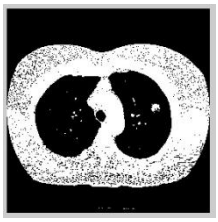

**Figure 3: Sample of K-Means Outcome**

The next step proceeds with removal of the lung region, so that the nodules are only present in the image. The removal of lung region is done through series of morphological operations [15] involving complement of the obtained cluster followed by suppression of the surroundings, Opening of the resultant image, performing dilation, filling and closing on the resultant image. Having segmented the lungs, now the nodules inside this region along are considered as candidate nodules. The segmented lungs and the candidate nodules are shown in Figure 4.

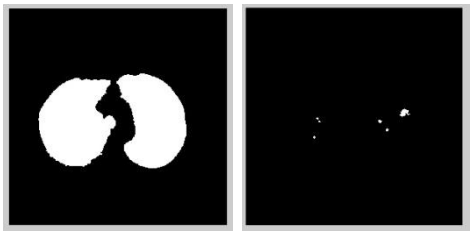

**Figure 4: Sample of lung segmented and nodule detected image**

Since there would be many candidate nodules, the properties of all the obtained nodules are elicited and provided for classification. The features extracted from the binary image of the nodule [16] include the area, perimeter, eccentricity and diameter. The features extracted from the input image [17] include the sum of intensities in the nodule, mean, standard deviation, entropy and range.

The extracted features are given as input to various classification algorithms. The classification algorithms attempted include Alternation Decision Tree [18], C4.5 [19] and Random Tree [20]. Since prediction from bagged classifiers [21] performs better than prediction from single classifier, ensembling using bagging is also attempted for the mentioned classifiers. It was seen that Bagged Random Trees performed higher than the other algorithms. The procedure for bagged random tree is shown in Figure 5.

*Step 1: For I= 1 to num_trees*
*Step 1a: Create bootstrapped data from original data*
*Step 2a: Perform Random Tree Classification as shown in Figure 6.*
*End*
*Step 2: For an instance whose label is not known, aggregate prediction of all trees.*
*Step 3: Final prediction is the class which has been predicted by maximum number of trees.*

**Figure 5: Procedure for Bagged Random Tree**

The following figure 6 explains the procedure of individual random trees.

*Step 1: Create a root node RN.*
*Step 2: If all instances in D belong to the same class C, then*
*Return RN as the leaf node labeled with class C.*
*Step 3: Randomly select $log_2M$ attributes from feature set F with M attributes.*
*Step 4: Compute Information gain for all the selected attributes.*
*Step 5: Select attribute with maximum information gain and test the split criterion.*
*Step 6: For each outcome of the splitting criterion,*
*Create a child node*
*If dataset is non-empty, recursively build the random tree.*
*Else assign leaf node with majority class in dataset*

**Figure 6: Procedure for Random Tree**

The performance is evaluated through leave-out-out cross validation. Leave-one-out cross validation [22] evaluates by considering n-1 instances as training and the remaining one instance as test. The procedure is repeated for n number of times. The results are reported through the metric accuracy, which is defined as the num of correctly classified nodules and non-nodules to the total number of candidate nodules. The next section discusses the results.

# 4. RESULTS AND DISCUSSION

The proposed framework is evaluated on ELCAP dataset [23] with large nodules. Each image in the dataset is a three dimensional image comprising of numerous 2D slices. Each slice in the image is processed and the candidate nodules of the entire image are aggregated to form the data for classification. The implementation is done through Matlab2013a and Weka Data Mining software. The pre-processing, lung segmentation, nodule detection and feature extraction steps are carried out in Matlab. The data mining classification is performed in Weka. This section reports the accuracy (%) with which the candidate nodules are categorized as true nodules and non-nodules.

Initially, the nodules were classified using individual classifiers. Table 1 tabulates the accuracy with which the classification procedures categorized the candidate nodules.

**Table 1: Performance of Classifiers on nodule detection**

| Classifier | Accuracy (%) |
|---|---|
| Random Tree | 93.61 |
| C4.5 | 94.56 |
| AD Tree | 94.52 |

Table 1 reveals the performance of random tree, C4.5 and AD Tree in detecting large nodules. As bagging provides more accuracy in prediction when compared to individual predictions, the categorization was attempted with application of bagging. Table 2 reports the performance of bagged classifiers in detecting large nodules.

**Table 2: Performance of Bagged Classifiers on nodule detection**

| Bagged Classifier | Accuracy (%) |
|---|---|
| Bagged Random Tree | 95.31 |
| Bagged C4.5 | 94.94 |
| Bagged AD Tree | 95.09 |

Table 2 reveals the higher performance achieved by Bagged Random Tree in detection of large nodules in the thoracic CT images. The accuracy is justifiable to be used as a decision support system for medical practitioners.

# 5. CONCLUSION

Lung nodule detecting has been an interesting area of research in the recent past. Nodules are of varying size, out of which large nodules have to be treated immediately. Automated nodule detection is of great interest to the medical practitioners. In this work, the CT image is first pre-processed, then clustering through K-Means clustering. The outcome of clustering was morphologically processed to segment the lungs and detect the nodules. The nodules were detected for each slice of the image. Features for each of the nodules were extracted and provided for Bagged Random Tree classification. The outcome of the classification categorized the nodules as either true nodules or not with an accuracy of 95.31%. The proposed system has a social relevance.

# 6. REFERENCES

[1] Han, J., Pei, J. and Kamber, M. 2011. Data mining: concepts and techniques. Elsevier.

[2] Gurcan, M.N., Sahiner, B., Petrick, N., Chan, H.P., Kazerooni, E.A., Cascade, P.N. and Hadjiiski, L. 2002. Lung nodule detection on thoracic computed tomography images: Preliminary evaluation of a computer- aided diagnosis system. Medical Physics, 29(11), pp.2552-2558.

[3] Remy-Jardin, M., Remy, J., Giraud, F. and Marquette, C.H. 1993. Pulmonary nodules: detection with thick-section spiral CT versus conventional CT. Radiology, 187(2), pp.513-520.

[4] Gonzalez, R.C. and Woods, R.E. 2007. Image processing. Digital image processing, 2.

[5] Geetharamani, R. and Lakshmi, B. 2015. Automatic segmentation of blood vessels from retinal fundus images through image processing and data mining techniques. Sadhana, 40(6), pp.1715-1736.

[6] Shao, H., Cao, L. and Liu, Y. 2012. A detection approach for solitary pulmonary nodules based on CT images. In Computer Science and Network Technology (ICCSNT), 2012 2nd International Conference on (pp. 1253-1257). IEEE.

[7] Messay, T., Hardie, R.C. and Rogers, S.K. 2010. A new computationally efficient CAD system for pulmonary nodule detection in CT imagery. Medical image analysis, 14(3), pp.390-406.

[8] Gomathi, M. and Thangaraj, P. 2010. A computer aided diagnosis system for detection of lung cancer nodules using extreme learning machine. International Journal of Engineering Science and Technology, 2(10), pp.5770-5779.

[9] Farag, A., Ali, A., Graham, J., Farag, A., Elshazly, S. and Falk, R. 2011, March. Evaluation of geometric feature descriptors for detection and classification of lung nodules in low dose CT scans of the chest. In Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on (pp. 169-172). IEEE.

[10] Cascio, D., Magro, R., Fauci, F., Iacomi, M. and Raso, G. 2012. Automatic detection of lung nodules in CT datasets based on stable 3D mass–spring models. Computers in Biology and Medicine, 42(11), pp.1098-1109.

[11] Suganya, A., Mohanapriya, N. and Kalaavathi, B. 2015. Lung Nodule Classification Techniques for Low Dose Computed Tomography (LDCT) Scan Images as Survey. International Journal of Computer Applications, 131(14), pp.12-15.

[12] Zhang, F., Song, Y., Cai, W., Lee, M.Z., Zhou, Y., Huang, H., Shan, S., Fulham, M.J. and Feng, D.D. 2014. Lung nodule classification with multilevel patch-based context analysis. IEEE Transactions on Biomedical Engineering, 61(4), pp.1155-1166.

[13] Pizer, S.M., Amburn, E.P., Austin, J.D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J.B. and Zuiderveld, K. 1987. Adaptive histogram equalization and its variations. Computer vision, graphics, and image processing, 39(3), pp.355-368.

[14] Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R. and Wu, A.Y. 2002. An efficient k-means clustering algorithm: Analysis and implementation. IEEE transactions on pattern analysis and machine intelligence, 24(7), pp.881-892.

[15] Gurcan, M.N., Sahiner, B., Petrick, N., Chan, H.P., Kazerooni, E.A., Cascade, P.N. and Hadjiiski, L. 2002. Lung nodule detection on thoracic computed tomography images: Preliminary evaluation of a computer- aided diagnosis system. Medical Physics, 29(11), pp.2552-2558.

[16] Ramani, R.G., Lakshmi, B. and Jacob, S.G., 2013, August. ROC Analysis of classifiers in automatic detection of diabetic retinopathy using shape features of fundus images. In Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on (pp. 66-72). IEEE.

[17] GeethaRamani, R, Lakshmi, B. and Shomona, G.J. 2012. Automatic prediction of Diabetic Retinopathy and Glaucoma through retinal image analysis and data mining techniques. In Machine Vision and Image Processing (MVIP), 2012 International Conference on (pp. 149-152). IEEE.

[18] Quinlan, J.R. 2014. C4.5: programs for machine learning. Elsevier.

[19] Freund, Y. and Mason, L. 1999, June. The alternating decision tree learning algorithm. In icml (Vol. 99, pp. 124-133).

[20] Breiman, L. 2001. Random forests. Machine learning, 45(1), pp.5-32.

[21] Breiman, L. 1996. Bagging predictors. Machine learning, 24(2), pp.123-140.

[22] Ramani, R.G., Lakshmi, B. and Jacob, S.G. 2012. Data mining method of evaluating classifier prediction accuracy in retinal data. In Computational Intelligence & Computing Research (ICCIC), 2012 IEEE International Conference on (pp. 1-4). IEEE.

[23] ELCAP public lung image database. www.via.cornell.edu/databases/lungdb.html.