ISSN (ONLINE):2456-5717
International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol.4, Issue.6, June 2018

A Vital Approach to Classify Diffused Lung Disease Patterns

Vijayakumari Pushparaj
Assistant Professor (SG), ECE, Mepco Schlenk Engineering College, Sivkasi-626005

Abstract – **Lung diseases are the common type of disease in the world that affect the lungs, which is one of the most grave health problem in India. In this paper, an algorithm is developed to detect and classify five diseased patterns and normal lung. The main goal of the work is to segment, detect and classify the lung disease patterns with random forest classifier and compare the results with Multi class SVM and K-NN classifiers. It includes effective segmentation through Fuzzy Connectedness and feature extraction through second order statistics. The pre-segmentation techniques will detect the diseased patterns and the useful features are extracted from the image, then the classifiers are employed to classify the lung disease and the performance measures are obtained to validate the algorithm. The result shows that the random forest classifier provides more promising results than the other two classifiers. Experimental results show the ability, accuracy and high performance of the proposed algorithm.**

Keywords: **Pre-segmentation, Pleural effusion, Honeycomb, Cavity, Fuzzy Connectedness, Skin Boundary**

I. INTRODUCTION

Nowadays, pulmonary diseases and disorders are one of the major reasons for the deaths and hospitalization around the world. The American Lung Association estimates that about 400,000 deaths occur per year in the United States are due to lung diseases. The signs and symptoms of this disease differ by its type. For non-invasive diagnosis of lung diseases, Computed Tomography (CT) is the current standard in the routine medical field which provides more information as it has more slices for a single direction. The segmentation is an initial prerequisite method for lung disease categorization. The various issues addressed in the literature includes that aaccurate classification of pathological lung is a difficult one for CAD systems because CT scans may not diagnose the disease pattern easily due to the sudden changes in lung patterns. This can be handled by using specific image processing algorithm in the proposed work. The proposed work addressing this challenge, and provide a generic and effective solution for lung segmentation from CT scans and hence automatically classifying the five abnormal imaging patterns such as Honeycomb, Tree-in-Bud(TIB), Ground Glass Opacity (GGO), Pleural effusion ,Cavity along with normal lung as shown in Fig.1.

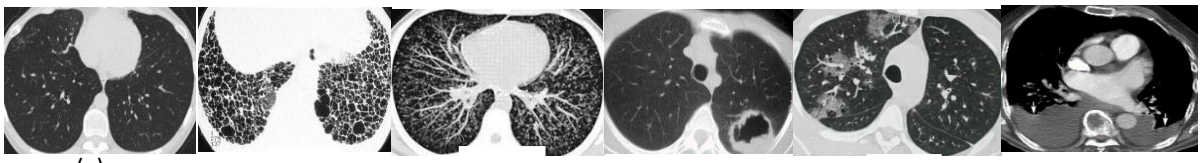


Fig. 1. Lung Disease patterns (a)Normal (b) Honeycomb (c) TIB (d) GGO (e) Cavity (f)Pleural effusion

For the more robust, fast, and flexible lung disease segmentation and classification fuzzy connectedness algorithm and random forest classifier have been chosen in this work. The paper is organised as follows: Related Works are discussed in Section 2. Section 3 describes the methodology of the proposed system. Section 4 shows the results and discussion of the system. Finally, the proposed algorithm is concluded.

II. RELATED WORKS

The research work for interstitial lung disease pattern detection was started at the fullswing from 2006 onwards.In 2010, Lauge Sorensen et al. [1] analysed the quantitative of pulmonary emphysema using Local binary patterns. For improving quantitative measures of disease patterns in CT images of lungs by texture analysis such as Local Binary Patterns and Gaussian Filter Bank are used.In 2011, the left and right lungs separation [2] was performed using 3D information of CT images and a guided dynamic programming algorithm. This algorithm separates the left and right lung accurately. In 2012, D. L. Gupta et al. [3] analysed the performance of classification tree learning algorithms using four classifiers such as J48, Random Forest (RF), Reduce Error Pruning (REP) and Logistic Model Tree (LMT).But, machine learning algorithm has gained growing interest in segmenting abnormal organs due to their strong abilities to exploit intensity, shape, and anatomy information.Yang Song etal.proposed feature based image patch approximation [4] for lung tissue classification. Features used were RGLBP texture and MCHOG gradient descriptor. SVM was used for classification. The patch approximation algorithm yields high precision rate. 3-D texture classification for lung pathologies [5] was also performed. Vijayakumari et al. [6] introduced an automatic ground glass pattern and honey combing detection with gabor filter bank .

These approaches mostly concentrae on extracting appropriate features such as shape and/or texture for a classifier such as support vector machines, random forests, neural networks etc. A number of feature sets for classification of lung disease patterns have been proposed: 3-D Adaptive Multiple Feature Method (AMFM), texton-based approach [7], intensity-based features [8], Gray Level Co-occurrence Matrix (GLCM) [9], wavelet and Gabor transform, shape and context-based attributes [9], [10], and Histogram of Gradients (HOG) [11],[12]. The most interesting aspect of these approaches is the selection of appropriate feature set, which is an active area of research. In 2014, AwaisMansoor et al. [13] analysed a novel method for fully

automated lung segmentation with and without abnormalities by applying fuzzy connectedness (FC) algorithms.

III. PROPOSED WORK

The proposed model is evolved for the classification of lung diseases by image processing techniques using computed tomography images for obtaining the desired results. The software used is MATLAB version 10.11.0 (R2015a). Fig.2 defines the steps followed in the proposed work. It involves five major steps for designing an automated lung disease pattern segmentation and classification. The following steps are involved in the proposed work :For the accurate classification, the input images are pre-segmented by the FC algorithm to efficiently detect the lung disease patterns. Then, the second order statistical features are used to extract the appropriate features from the pre-segmented images.In the next step,the diseased patterns are classified by RFC, MSVM and k-NN tocategorize the six classes as normal lung image, Ground Glass Opacity (GGO), Cavity, Pleural effusion, Tree-in-Bud (TIB) and Honeycomb. Finally, the performance measures like accuracy, specificity Sensitivity, Precision, G-Mean, F-Measure, Error rate and ROC are evaluated.

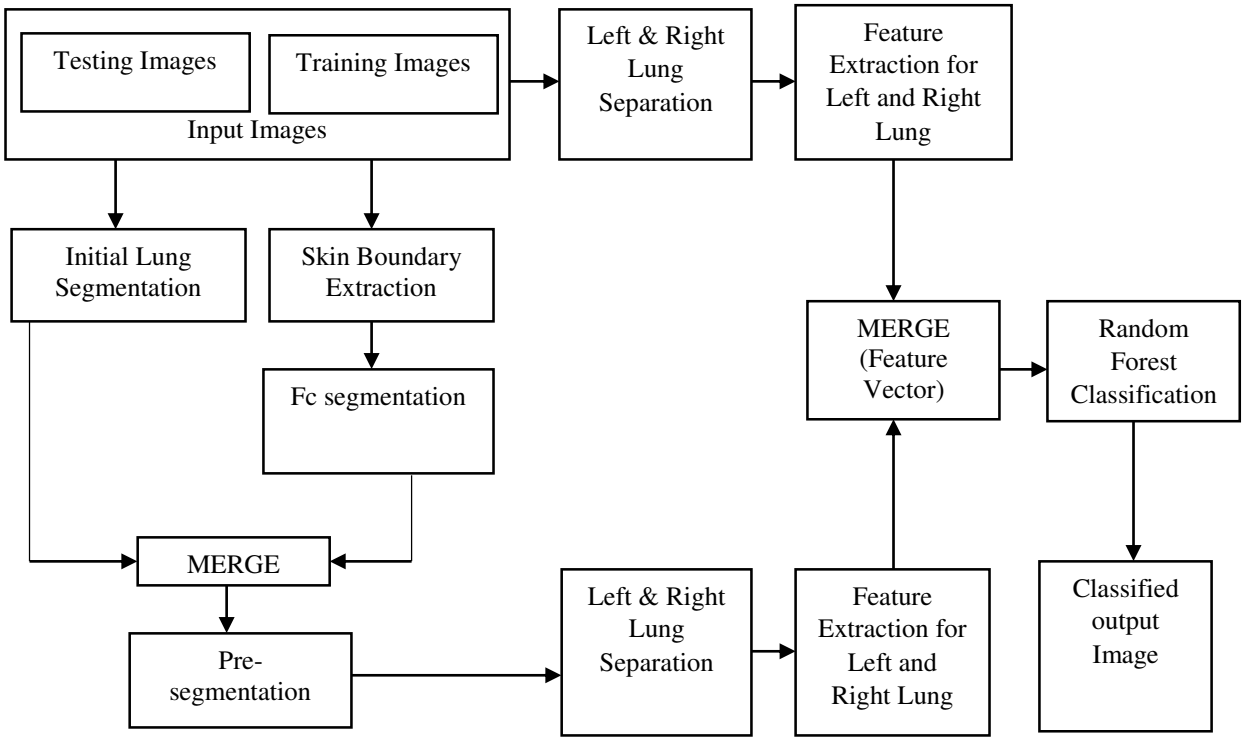


Fig 2. Block Diagram of the proposed work

A. Pre-segmentation

Image pre-segmentation can significantly increase the reliability of an optical assessment. For this, Fuzzy Connectedness algorithm is used and it is explained in our earlier work [14]. For qualitative feature extraction, the pre-segmented image is overlaid with the given input image for different pathologies.

B. Lung Separation

In order to get refined and strengthened feature vector, the input and segmented images are further split into left and right lungs. The left and right lung is separated by creating the masks for defining Region of Interest (ROI).The original image is overlaid with the masked region

C. Feature Extraction

To reduce the input vector size of images, features are extracted for describing a large set of data accurately. The feature sets commonly used in various lung CAD systems are GLCM, Gray Level Run Length Matrices (GLRLM), and HOG. A GLCM is a two dimensional matrix where the number of rows and columns is equal to the number of gray levels, G, in the image. The GLCM features used in this work are Energy, Contrast, Variance, Correlation and Homogeneity. The energy is obtained as,

$$Energy = \sqrt{ASM} \quad \text{- --- (1)}$$

where, $ASM = \sum_i \sum_j p(i,j)^2$

Here,*i*and *j* are the horizontal and vertical cell coordinates and *p*is the cell value. The contrast parameter *C*is,

$$C = \sum_i \sum_j (i-j)^2 p(i,j) \quad \text{------ (2)}$$

The Variance is obtained using, $V = \sum_i \sum_j (i-\mu)^2 p(i,j)$ ------ (3)

where μ is the mean value of *P*. The correlation is,

$$Corr = \sum_i \sum_j p(i,j) \frac{(i-\mu_x)(j-\mu_y)}{\sigma_x \sigma_y} \quad \text{------- (4)}$$

where μ_x ; μ_y and σ_x ; σ_y are the mean and standard deviations and are expressed as:

$$\mu_x = \sum_i \sum_j ip(i,j) \text{ and } \quad \mu_y = \sum_i \sum_j jp(i,j)$$

$$\sigma_x = \sqrt{\sum_i \sum_j (i-\mu_x)^2 p(i,j)} \text{ and } \sigma_y = \sqrt{\sum_i \sum_j (i-\mu_x)^2 p(i,j)}$$

and homogeneity is, $$H = \sum_{i,j} \frac{1}{1-(i-j)^2} p(i,j) \quad \text{------- (5)}$$

For higher order statistical texture measures in an image, GLRLM features are extracted. A set of consecutive pixels with the same gray level, collinear in a given direction, organize the graylevel run [15]. The run length is the number of pixels in the run and the run length value is the number of times such a run occurs in an image. The descriptors, which are classically extracted from the run-length matrices in this work are short Run Emphasis (SRE), Long Run Emphasis (LRE), Run Percentage (RP), Gray-Level Non-Uniformity (GLNU), and Run-LengthNon-Uniformity (RLNU). These parameters are observed using the follwing relations:

$$SRE = \frac{1}{n} \sum_{i,j} \frac{p(i,j)}{j^2} \quad \text{---- (6)}$$

$$LRE = \frac{1}{n} \sum_{i,j} j^2 p(i,j) \quad \text{---- (7)}$$

$$RP = \sum_{i.j} \frac{n}{p(i,j)j} \quad \text{---- (8)}$$

$$GLN = \frac{1}{n} \sum_i \left\{ \sum_j p(i,j) \right\}^2 \quad \text{---- (9)}$$

$$RLN = \frac{1}{G} \sum_j \left\{ \sum_i p(i,j) \right\}^2 \quad \text{---- (10)}$$

The Histogram of Oriented Gradients (HOG) is a feature used for the purpose of object detection in computer vision and image processing [16]. It is a local statistics of the orientation of the image gradients. It is considered by its invariance to rotation and illumination changes.

D. Classifiers

After the feature extraction process, feature vectors are obtained and then it is fed to a classifier in order to be classified. Many different classification methods have been used but for this work,

the three classifiers such as Random forest classifier (RFC), Support vector machine (SVM), and KNN are used and their performance measures are evaluated.

Ensemble-learning algorithms are receiving more attention in the field of classification. Random Forest fits to this ensemble method category; they resemble on combination of decision tree-type classifier, such as each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. To classify a new input vector, each tree "votes" for a class and the forest elects the class having the majority of votes over all the trees in the forest. The RFC algorithm is summarized as follows:The *N* bootstrap samples from the training data areobtained at first.Then for each of the bootstrap samples, unpruned classification tree is grown. Finlly, the predictions of the N trees are aggregated for predicting new data based on majority votes.

k-NN (k-Nearest Neighbors) is a modest machine learning algorithm that stores all available class and classifies new class based on a similarity measure[17]. In k-NN classification, a class membership is the output. An object is classified by a majority vote of its neighbors. Here, Euclidean distance is a distance metric for continuous variable.

Several techniques have been developed to deal multi class problems. Among the various methods, one-against-rest is considered here. Multi-class SVM formulation [5] with K-class is defined as:

$$\underset{w_1,\ldots..w_k}{\text{minimize}} \ \frac{\lambda}{2}\sum_{c=1}^{K} w_c^T w_c + \sum_{i=1}^{N}\xi_i \qquad \text{------ (11)}$$

$$\text{subject to } (w_{yi} - w_c)^T x_i \geq 1 - \xi_i - \delta_{yi,c} \qquad \text{------ (12)}$$

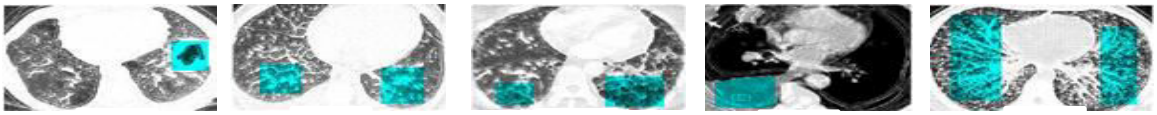
$\forall i = 1,\ldots...., N, \quad c = 1,\ldots..., K$, Here c refers to each class and w_c is its weight vector, and each of the N training vectors x_i has label y_i. The indicator function $\delta_{yi,c}$ =1 if $y_i = c$ and 0 otherwise. The variable ξ_i refers to slack variables for each data item, in such a way that the margin between correct class and most confusing class is penalized.

IV. Results And Discussion

In this section, experimental results of our algorithm are presented. The real time database is collected from radiologists of Tanjore medical college hospital and Aarthi scan centre, Rajapalayam. It includes five lung disease patterns like Ground Glass Opacity (GGO), Cavity, Pleural effusion, Tree-in-Bud (TIB), Honeycomb and normal lung images. These images were obtained from 90 subjects (#40 Male, #50 Female) with an average age group of 45 to 70 years. A user friendly algorithm was developed to separate left and right lungs .This algorithm is

evaluated with a database of 140 CT scan left and right lung images. Each segmented image is a matrix of size 220x220 with 16 bit resolution and of DICOM standard.

After separating the lung regions, pre-segmentation was done by fuzzy connectedness algorithm using sigmoidal membership function and skin boundary extraction. Fuzzy segmentation was done for disease pattern detection. Skin boundary extraction was performed to get the perfect detection of disease patterns. Finally, the output of two stages such as initial lung segmentation using fuzzy and skin boundary extraction based fuzzy edge detection are merged to produce pre-segmented output. The images obtained by merging are overlaid with the given input image for different pathologies for qualitative feature extraction. The pre-segmented images using FC is shown in Fig.3.



(a) (b) (c) (d) (e)

Fig.3. Pre-segmentation (a) Cavity (b) GGO (c) Honeycomb(d) Pleural effusion (e) TIB

For accurate classification and strengthened features, the left and right lung is separated in pre-segmented images. They are separated using Binary Singleton Expansion method. The resultant images are shown in Fig 4. Features like GLCM, GLRLM and HOG are extracted from the separated left and right lungs. A snapshot of the extracted features is shown in Fig 5. In these three features, HOG features look more similar for most of the images and GLCM, GLRLM gives most prominent values. For the 60 images, features are extracted and a feature vector has been created. Finally the images are classified using the three classifiers. The accountability of these classifiers are analysed with their performance measures. The classified output of RFC is shown in Fig 6.

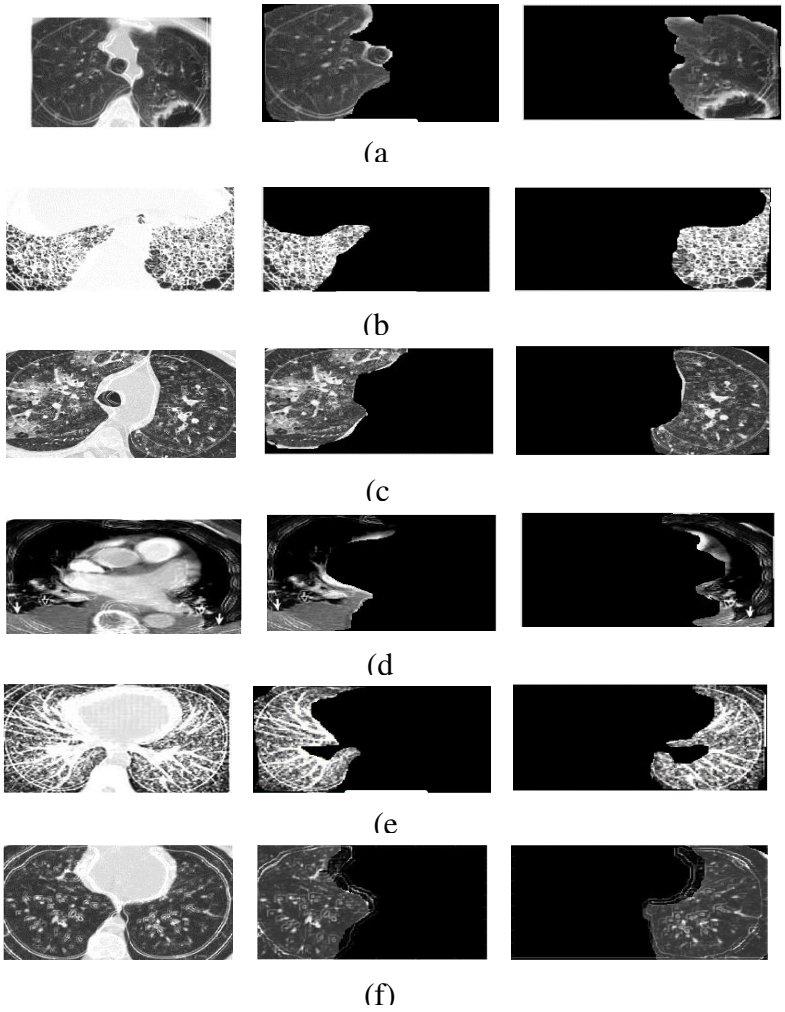
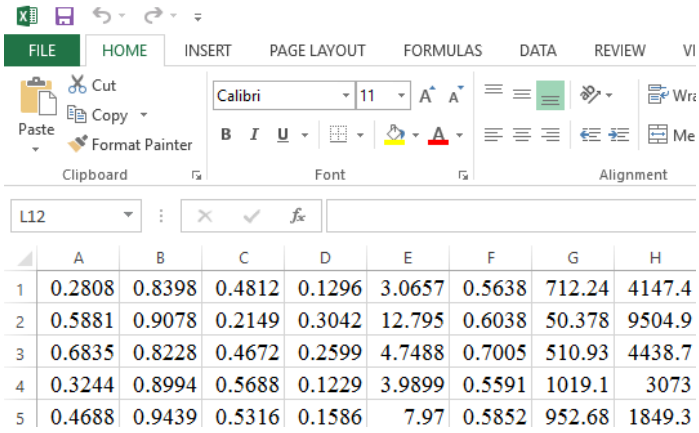


Fig. 4 Separation of left and right lung for pre-segmented image.
(a) Cavity (b) Honeycomb (c) GGO (d) Pleural effusion (e) TIB (f) Normal



	A	B	C	D	E	F	G	H
1	0.2808	0.8398	0.4812	0.1296	3.0657	0.5638	712.24	4147.4
2	0.5881	0.9078	0.2149	0.3042	12.795	0.6038	50.378	9504.9
3	0.6835	0.8228	0.4672	0.2599	4.7488	0.7005	510.93	4438.7
4	0.3244	0.8994	0.5688	0.1229	3.9899	0.5591	1019.1	3073
5	0.4688	0.9439	0.5316	0.1586	7.97	0.5852	952.68	1849.3

Fig. 5 Extracted Features

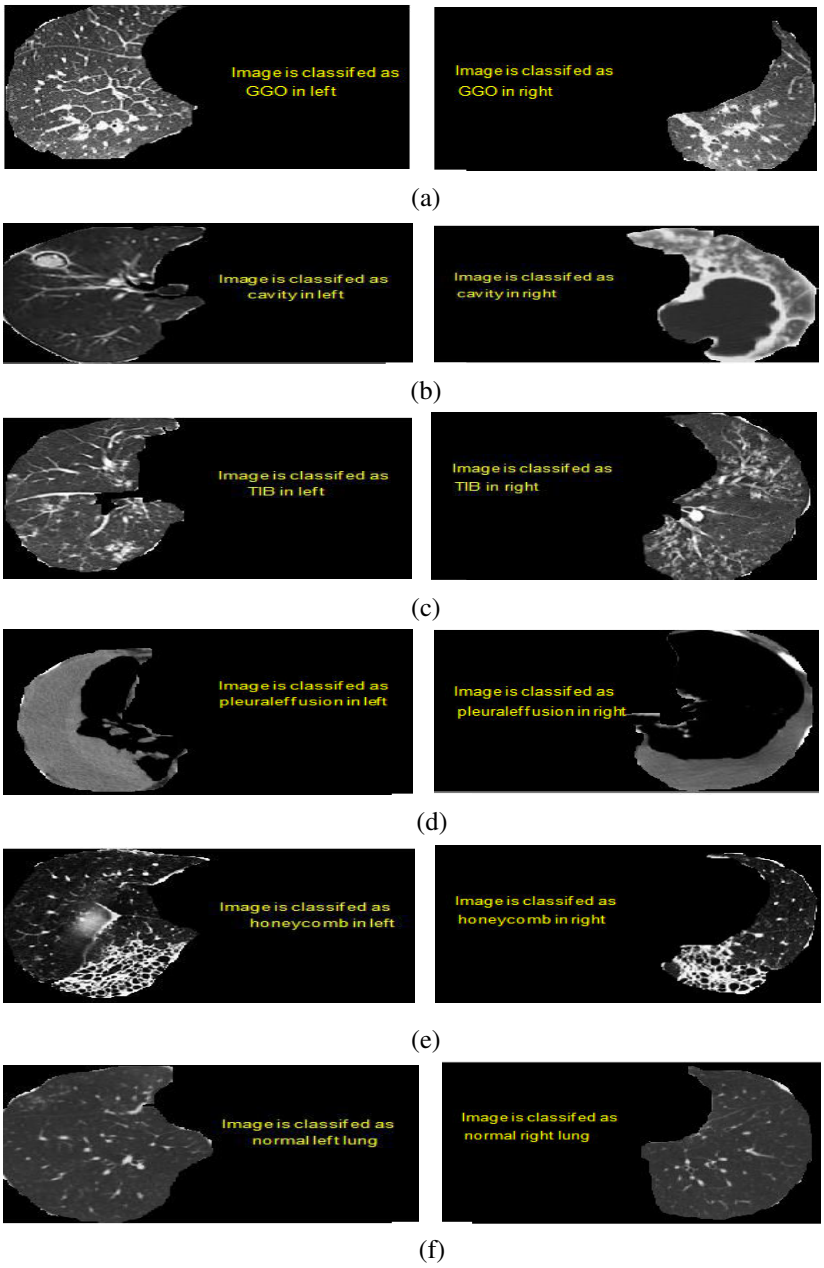


Fig. 6 Classified Results of RFC (a) GGO (b) Cavity (c) TIB (d) Pleural effusion (e) Honeycomb (f) Normal

The performance of these three classifiers are evaluated using the standard measures of classifiers like Classification Accuracy, Sensitivity, Specificity, Precision, G-Mean, F-Measure and error rate. Sensitivity of the algorithm is measured by knowing individual diseased region labeled in the original image and the result of classifier algorithm. The specificity of the algorithm can be obtained by one minus probability of an image being labeled as the disease pattern when there is no disease pattern present. Accuracy of the algorithm measures validity of the proposed algorithm. Precision is the ratio of correctly classified images to the number of entire images classified fault-prone. It is proportion of images correctly predicted as faulty. The

geometric mean is another measure used here. FM is a combination of recall and precision. It is also defined as harmonic mean of precision and recall.Error rate is finding the value of classification accuracy through misclassification.performance of the system is surveyed by validating correct and incorrect disease patterns. The individual performance of each diseased pattern and normal lung for RFC classifier is shown in Fig. 7. Each tree in the forest chooses the appropriate class of the input data for classification. The most frequent value generated by all the trees determined the output of the final classifier.

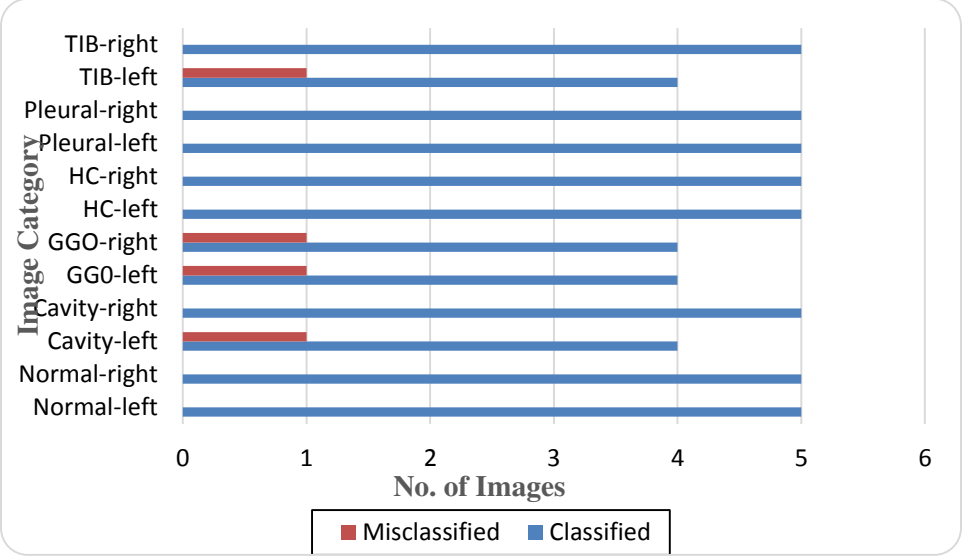


Fig.7 Performance of RFC Classifier for each image patterns.

The performance of RFC is analysed with other classifiers. The algorithm is analysed with three classifier and their corresponding computational results are presented in Table1.

Table1. Performance measures for three classifiers

Classifiers / Measures	RFC	k-NN	MSVM
Accuracy	0.95	0.8833	0.8667
Sensitivity	1	0.9574	0.9565
Specificity	0.7692	0.6154	0.5714
Precision	0.94	0.9	0.88
G-Mean	0.8771	0.7676	0.7393
F-Measure	0.9691	0.9278	0.9565

From the evaluation and values of specificity, sensitivity, accuracy, Precision, G-Mean and F-Measure Random Forest Classifier gives better accuracy rate of 95%, and the other two classifiers like k-NN and MSVM gives the accuracy rate of 86% and 83% respectively as shown in table 1.

For the improved clarification, error rate is calculated for each examined algorithm which is mentioned in Table 2. Hence it is also concluded from the error rate that RFC is showing minimum than the other two classifiers because of its random behavior. It is also noticeable that MSVM classifier is showing maximum error rate.

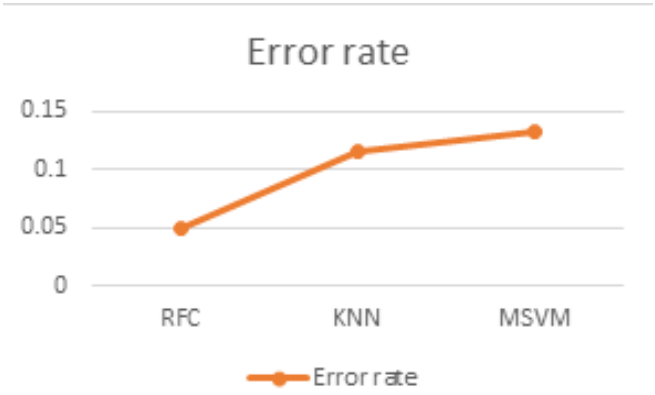


Fig.8. Error rate performance.

Other common evaluation measure used in classification problems is the Receiver Operating Characteristics (ROC), which relates sensitivity and specificity.ROC curves obtained for three classifiers are shown in Fig.9. The ROC curve is plotted between the True Positive Rates (TPR) vs. the False Positive Rates (FPR) by varying the threshold value. It is clearly observed that the blue curve is better than the other two curves because blue curve is closest to the true positive rate. As the blue curve lies along the axis, it produces highest accuracy than the other curves. The inference from the ROC curve, table 1 & 2 and Fig.8 & 9 indicates that , RFC gives better results than the other classifiers because it has an effective method for estimating missing data and maintains accuracy even when a little proportion of the data are missing.

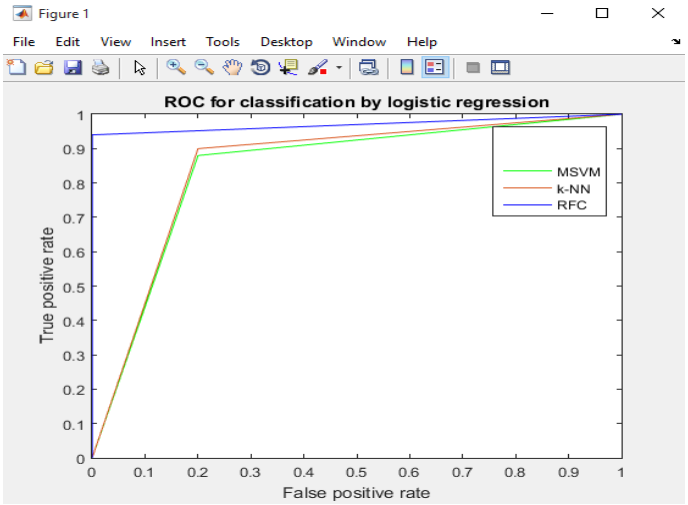


Fig. 9.ROC curve for the three classifiers.

V. CONCLUSION

A novel method for fully automated lung segmentation and classification with various abnormalities is presented in this work. In this paper, a method for classifying five lung disease patterns and normal lung based on fuzzy connectedness algorithm and random forest classification has been introduced. The algorithm was tested on five different lung pathologies. The proposed method has magnificently classified all the images. The robustness and the effectiveness of this proposed method was tested on different lung scans acquired through various sources containing a wide range of abnormalities. RFC provides maximum accuracy and minimum error rate. The exhaustive testing confirms the accuracy and the effectiveness of the proposed method. The experimental results proved the superior performance and efficiency of the proposed method achieving the best results with an average F-score of 96% and better ROC curve.

REFERENCES

[1] Sørensen,L., Shaker,S. B., and De Bruijne,M. , 2010, "Quantitative analysis of pulmonary emphysema using local binary patterns," *IEEE Trans. Med. Imag*., 29(2), pp. 559–569.
[2] Sang Cheol Park, Joseph Ken Leader, Jun Tan, Guee Sang Lee, SooHyung Kim, In Seop Na, and Bin Zheng,2011, "Separation of left and right lungs using 3D information of sequential CT images and a guided dynamic programming algorithm",NIH public access,35(2).
[3] Gupta,D. L.,MalviyaSatyendra Singh,A. K., 2012, "Performance Analysis of Classification Tree Learning Algorithms" *IJCA*, 55(6), pp. 975-8887.
[4] Gupta,D. L.,MalviyaSatyendra Singh,A. K., 2012, "Performance Analysis of Classification Tree Learning Algorithms" *IJCA*, 55(6), pp. 975-8887.
[5] Song,Y., Cai,W., Zhou, Y., and Feng,D., 2013, "Feature-based image patch approximation for lung tissue classification," *IEEE Trans. Med. Imag*., 32(4), pp. 797–808.

[6] Xu,Y., Sonka,M., McLennan,G.,Guo,J., and Hoffman,E. A.,2006, "MDCTbased 3-D texture classification of emphysema and early smoking related lung pathologies," *IEEE Trans. Med. Imag*., 25(4), pp. 464–475.
[7] Vijakumari,B.,Geetha,A.,Haribaskarraj,D.,Senrayaperumal,R.,Saronsamraj,R.,2008, "Automatic ground glass pattern detection in lung diseases using gabor filter" *IETE Journal of Research*,54(3), pp. 249-254.
[8] Gangeh,M. J.,Sørensen,L., Shaker,S. B., Kamel,M. S., De Bruijne, M., and Loog,M., 2010, "A texton-based approach for the classification of lung parenchyma in CT images," in Medical Image Computing and Computer- Assisted Intervention—*MICCAI* 2010. New York: Springer, pp. 595–602.
[9] Yao,J.,Dwyer, A. R., M. Summers, and Mollura, D. J., 2011,"Computeraided diagnosis of pulmonary infections using texture analysis and support vector machine classification," *Acad. Radiol*., 18(3), pp. 306–314.
[10] Bagci,U. Yao,J. , Wu,A.,Caban,J.,Palmore,T. N.,Suffredini, A. F.,Aras,O. , and Mollura,D. J., 2012, "Automatic detection and quantification of tree-in-bud (TIB) opacities from CT scans," *IEEE Trans. Biomed. Eng*., 59(6), pp. 1620–1632.
[11] Sluimer, I. C., Prokop, M., Hartmann, I.,and van Ginneken, B., 2006, "Automated classification of hyperlucency, fibrosis, ground glass, solid, and focal lesions in high-resolution CT of the lung," *Med. Phys*., 33,pp. 2610.
[12] Song,Y. ,Cai,W.,Kim, J.,and Feng,D. D., 2012, "A multistage discriminative model for tumor and lymph node detection in thoracic images," *IEEE Trans. Med. Imag*., 31(5), pp. 1061–1075.
[13] Ye, X. , Lin,X.,Beddoe,G., and Dehmeshki, J.,2007,"Efficient computer aided detection of ground-glass opacity nodules in thoracic CT images," *IEEE Eng. Med. Biol. Soc*., pp. 4449–4452,
[14] AwaisMansoor, Member, IEEE, UlasBagci, Member, IEEE, ZiyueXu, Brent Foster, Kenneth N. Olivier, "A Generic Approach to Pathological Lung Segmentation" ,*IEEE*. 33(12), pp. 2293-2310.
[15] Vijayakumari,B., Nivetha,M., Prabin Jose, J., and Ganga Devi,J., 2016, "Analysis of lung Diseased Pattern Segmentation using Fuzzy Connectedness", *International Journal of Applied Engineering Research*
[16] Galloway, M.M.,1975, "Texture Analysis Using Gray Level Run Lengths," *Computer Graphics Image Processing*, 4, pp. 172–179.
[17] Dalal,N.,and Triggs,B., 2005, "Histograms of Oriented Gradients for Human Detection", Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2, pp. 886-893.
[18] Min-Ling Zhang and Zhi-HauZhou,"2005, A k-Nearest neighbour based algorithm for multi-label classification", IEEE, 2, pp.718-721.
[19] Yao, J., Dwyer, A., Summers, R. M. and Mollura,D. J., 2011, "Computer aided diagnosis of pulmonary infections using texture analysis and support vector machine classification," *Acad. Radiol*., 18(3), pp. 306–314.

AUTHOR BIOGRAPHY

Vijayakumari Pushparaj completed her BE from Institute of Road and Transport Technology, Erode during the year 1997, received her ME (communication systems) from Thiagarajar College of Engineering, Madurai, during the year 2006, and completed full-time PhD in medical image processing at Thiagarajar College of Engineering, Madurai. She is currently working as an Assistant Professor (Selection Grade) in the department of ECE, Mepco Schlenk Engineering College, Sivakasi. Her research area includes medical image processing and image analysis. Her research publication includes few national and international journals and conferences.