Pulmonary Nodule Classification in Thoracic CT Images using Random Forest Algorithm

Utkarsh Shukla, Kshitij Srivastava, Aarav Bhati, M. Jasmine Pemeena Priyadarsini, A.Jabeena, G.K.Rajini

Abstract—In this paper, an automatic classification of thoracic pulmonary nodules with Computed Tomography Image as input is performed. We can crisply classify the nodules into two categories: Benign and Malignant. Benign nodules are the ones which do not cause any harm and even if they do, the impact is negligible. Malignant Nodules are the ones which, if not detected on time can cause severe damage to a person, even resulting in death. Henceforth, detection at early stage of lung cancer is critical. We plan to perform our analysis in 4 steps. Firstly, a noise free CT image is obtained after preprocessing. Then, we apply the improved Random Walker algorithm to perform region-based segmentation, resulting in generation of foreground and background seeds. The next step is to bring out important features of the segments. The features can be intensity, texture and geometry based. Finally we used an improved Random Forest method to generate classification trees, comprising of different class labels. Using RF Algorithm, we predict the accurate class label which corresponds to a particular type of nodule and the stage of cancer that it has developed.

Keywords: Benign; Malignant; Random; Walker; Forest; Preprocessing; Segmentation; Classification.

I. INTRODUCTION

Pulmonary nodules, also known as lung nodules, are mostly benign (non-cancerous). However, chances are there that some of them develop cancer-causing abilities. Such dangerous nodules are known as malignant nodules. They are mostly white patches on the lungs that can clearly be seen in a Computed Tomography (CT) scanned image. They are round in shape and their size varies from one another. A larger pulmonary nodule poses a higher threat of causing cancer compared to a smaller one, which is close to only a millimeter in size. Such large nodules are approximately 30 millimeters in size. If a doctor doesn't detect an increase in size of lung nodules in further CT scans, one can be mostly assured that the nodule is non-cancerous. These safe nodules can be caused by any infection in the past and do not require any treatment. However the nodules undergoing an increment in size are unsafe and must be analyzed further.

Revised Manuscript Received on October 05, 2019
Utkarsh Shukla, School of Electronics Engineering, Vellore Institute of Technology, Vellore
Kshitij Srivastava, School of Electronics Engineering, Vellore Institute of Technology, Vellore
Aarav Bhati, School of Electronics Engineering, Vellore Institute of Technology, Vellore
M. Jasmine Pemeena Priyadarsini, School of Electronics Engineering, Vellore Institute of Technology, Vellore
A.Jabeena, School of Electronics Engineering, Vellore Institute of Technology, Vellore
G.K.Rajini, School of Electrical Engineering, Vellore Institute of Technology, Vellore.

Because of inefficient detection, lung cancer still remains to be a prominent source of malignancy-connected deaths throughout the world. One of the significant reasons of lung-cancer death is that the symptoms do not show up until the disease has reached an advanced stage, which is dangerous to a point where follow-up treatment almost becomes a myth.

In normal practice, proper grouping of benign and malignant nodules is still a hot topic of discussion among the radiologists. However, these conventional methods of classification are immensely time-consuming and complex. This makes automatic classification an urgent need of the hour as doing so can track the progression of the disease, the type of nodule and based on these characteristics, it can help us predict the future of the patient. Automatic classification is much quicker to an extent that it can significantly reduce the severe effects of later-stage cancers and hence the number of deaths, thanks to its early detection and efficient classification. Even after the ability of various supervised and unsupervised methods of classifications, a lot of research is still being done on pulmonary nodules' behavior. Our focus of this domain is to come up with an efficient method of categorization, which leaves minimal room of doubt regarding the type of nodule present in the lung as well the stage of cancer.

II. METHODOLOGY

Let us begin with the description of some of the key processes used in this paper. They are **Image preprocessing, Region-based image segmentation, feature extraction and Image classification.** Our proposed method can be expressed in the form of a flowchart as follows:

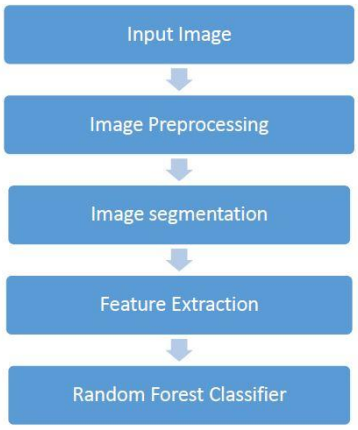
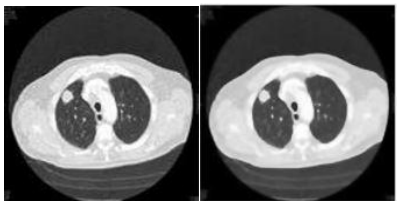


Fig.1 Flowchart of our proposed method

1. Input image: It is a CT scan image which is fed as the input. In case if it is colored, it is converted into appropriate gray image before it can be processed further. Best

resolution is 256*256 pixels for the input image.

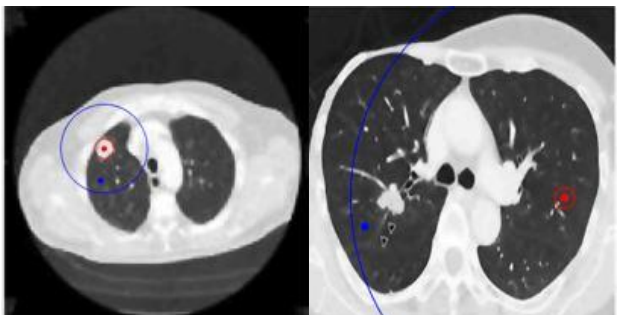
2. Image Preprocessing: Before we can perform segmentation on the pulmonary nodules, preprocessing the image is a key step. If there are noises present in image, this step suppresses them and still manages to preserve the nodule boundaries. Here we use an **anisotropic non-linear diffusion filter** to perform the preprocessing step. This filter removes noise without disturbing or blurring the nodule boundaries.



(a) (b)

Fig 2. (a) Input image. (b) Clearer image after preprocessing.

3. Region based pulmonary nodule segmentation: Before the image can be considered for feature extraction, the segmentation is very important. In normal practice, we use **Random Walker** algorithm for performing segmentation. Although a variety of segmentation processes [1-10] are available, precisely segmenting the pulmonary nodules is still a challenging aspect. Region based segmentation involves separating the image into various foreground and background regions. But using the normal Random Walker algorithm, it is difficult to group the foreground and background regions which have the same intensity. For this reason, we use an improved RW method, which results in automatic segmentation of pulmonary nodules using seed acquisition method. Here, the foreground and background seeds are sampled using a circle of radius R and 4-times-R respectively. The reason for performing this sampling is that in later stages, it helps us in determining a probability if a node has enlarged so as to reach the background region from foreground. The shortest possible distance [11] is used to obtain the nodule centers. Random Walker algorithm is a learning algorithm (supervised), which is used to obtain a partial labelling of the seeds. In a way, this process treats the image like a graph and minimize certain energy functions on that graph to result in segmentation.



(a) (b)

Fig.3 Images (a) and (b) showing the region based segmentation areas. The foreground region falls under the red circle, while the background under the blue circle.

Here, the Local Binary Pattern (LBP) description [12] method is implemented for assigning different labels and for obtaining different texture features. The labelling of pixels is done in such a way, that the neighboring pixel is set as a threshold and the final result is obtained as a binary number (0 or 1). In the below equation, b(i) is a prior-assigned label of the node i, which can defined as follows:

$$b(i) = -1, \text{ if } i \in V_M \text{ OR } 1, \text{ if } i \in V_{MB}$$

Here V_M and V_{MB} indicate the Maximum Response(MR) that would be obtained on an LBP Histogram.

4. Pulmonary Nodule Feature Extraction: After the CT images are completely segmented by using the improved Random Walker process, the next step is extracting its intensity, geometric and texture features before being given to RF classifier for training. Intensity features include factors like skewness, kurtosis, maximum intensity, etc. Texture features include contrast, cluster prominence, energy, entropy, maximum probability, sum of squares, sum average, sum variance etc. Finally, geometric features include eccentricity, circularity of a pulmonary nodules and many such features can be obtained if we know the lengths of semi-major axis as well as semi-minor axis of the nodules.

Our primary focus on this paper has been on working towards the texture features. Various methods can be employed for texture feature extraction, such as GLCM (Grey Level Co-occurrence Matrix) [13-16], Gabor Filter [17] and Local Binary Pattern [18-20]. Oftentimes, Gabor Filter and LBP methods are combined in order to ameliorate the discriminating strength of texture features. One intensity texture that we have calculated is as follows:

$$\Delta I(x,y) = (G_x^2 + G_y^2)^{1/2}$$

This is kind of the distance between two neighboring nodules, which can help us in extracting better features. LBP have brought about a revolution in feature extraction because of its simplicity and efficiency and can be used to detect lesions in breast cancer[21] as well.

The process of feature extraction helps us in obtaining a unique discrimination vector, which can be provided to an RF classifier for further training.

5. Random Forest Classification: The concept of Random Forest Algorithm is based on creating multiple decision trees and being able to predict a stable and accurate output. It is a supervised learning algorithm that creates an ensemble of decision trees and uses the bagging method in order to train these trees. A major significance of RF algorithm is that it can be used both for classification as well as regression. Below we can actually visualize a Random Forest with 2 trees.

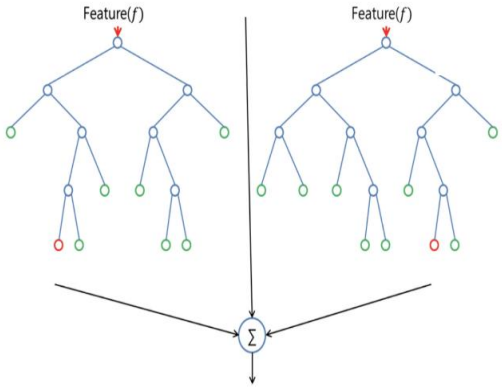


Fig 4. An example of a 2-tree Random Forest.

In our paper, we use the Random Forest Algorithm to classify the Benign and Malignant pulmonary nodules. RF is made to perform a prediction of a class label. At each node of the tree, a feature-subset is chosen in order to find the best break, which is mostly generated by making use of the bootstrap method.

A distinctive feature of the Random Forest process is its sensitivity to the number of trees. More the number of trees, more the number of distinctions that can be made and hence more accurate will be the obtained output. A confusion matrix is created based on the number of samples present. In our analysis, the increase of decision trees upto 60 is beneficial. Beyond that, there isn't any significant increase or improvement in accuracy. We can prove that by observing that the out-of-bag-probability becomes constant upon reaching 60 in any of our cases. Out of bag probability is an indicator of whether a sample is a part of a particular decision tree or not.

III. EXPERIMENTAL OBSERVATIONS

The explanation of concepts has already been covered in the Methodology section. Let us now have a look at the observations in our analysis. We have used MATLAB 2018 version for generating these outputs.

Three input images can be shown as follows:

Fig 5(a) Input images

Now all the further images will be in corresponding order of the input images only. The noise-free images after applying the anisotropic non-linear diffusion filter are as follows:

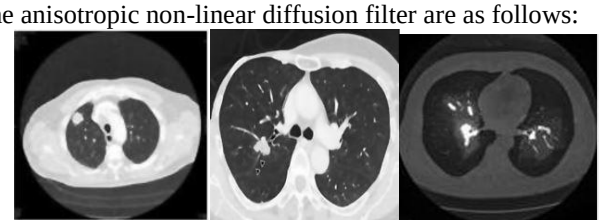


Fig 5(b) Anisotropic non-linear diffusion filtered images.

Then, we generate the texture features for above three images as follows:

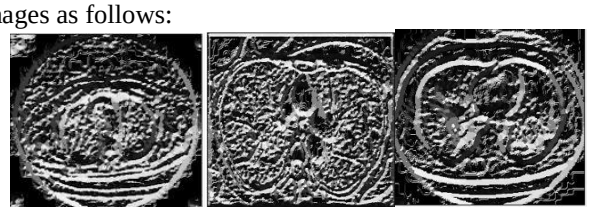


Fig 5(c) Texture Features

Then upon segmentation, we obtain foreground and background regions as follows:

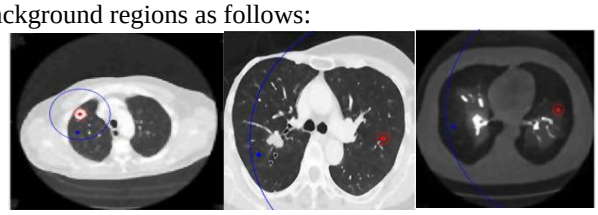


Fig 5(d) Foreground and background regions.

In the next page, we show the output masks obtained for the 3 images. The purpose of output mask is to help us focus on only that area of the nodules which needs diagnosis, meanwhile removing the unnecessary parts of the image that need not be examined.

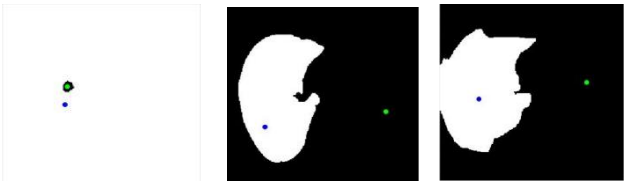


Fig 5(e) Output Masks.

Then we obtained the Outlined masks, which recombine the noise-free image and the output mask.

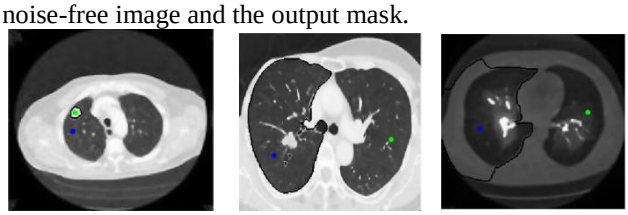


Fig 5(f) Outlined Masks.

Next, we obtained the boundary of Improved Random Walker Process, which can be shown as follows:

Fig 5(g) Boundary of Improved Random Walker

We then performed the necessary differential operations to obtain the Foreground and background close-ups.

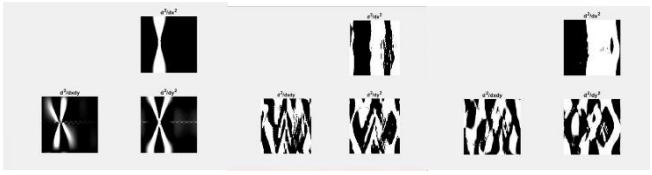


Fig 5(h) Images with Foreground and Background Close-ups.

Then we obtain the respective steerable Riesz wavelets, that tell us about the number of decomposition levels by taking into consideration the Number of Samples and Riesz order.

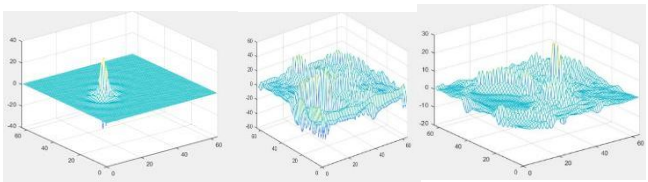


Fig 5(i) Steerable Riesz Wavelets.

Then we display the graph of Out of bag error probability against the number of trees. As mentioned earlier, the graph becomes constant, by the time we reach 60 trees. The 3 graphs for the initial 3 input images can be shown as follows:

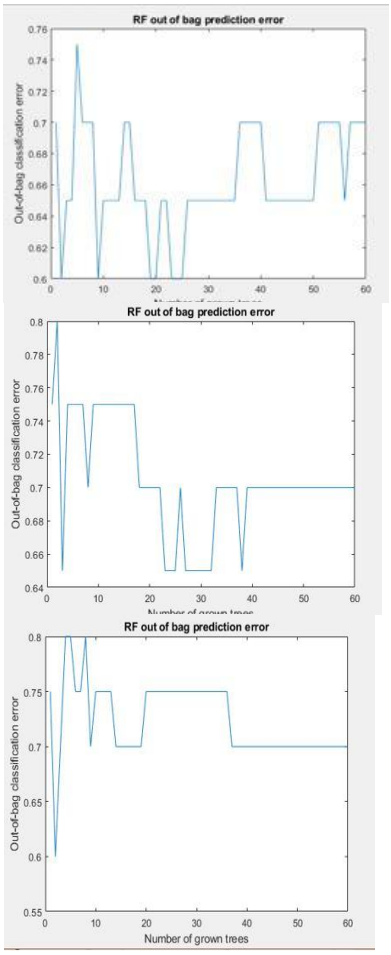
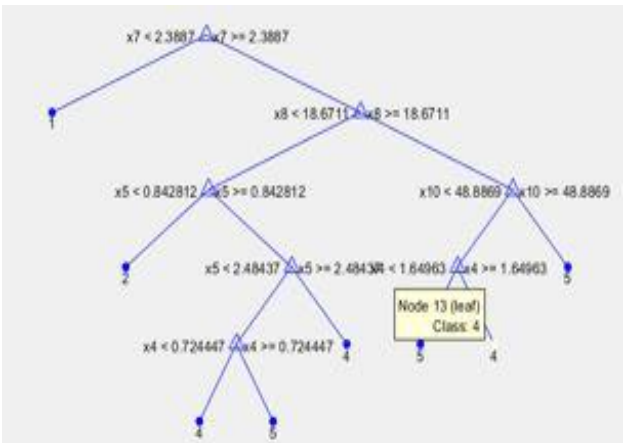
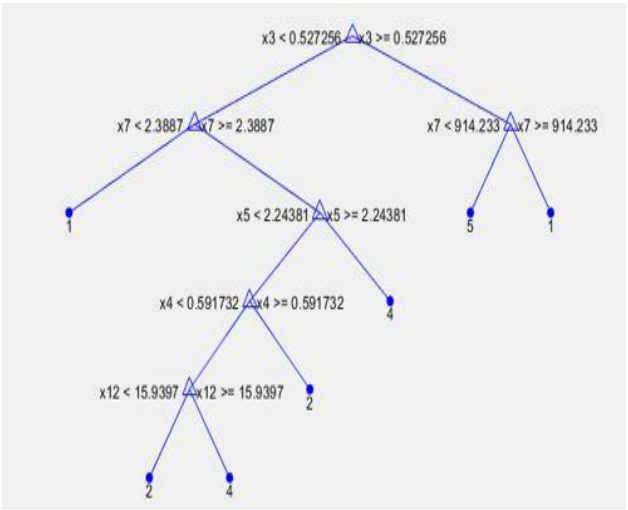


Fig 5(j) Out of bag error probability graphs against number of given trees.

Finally we obtain the classification tree for the 3 images as follows. These trees are generated using Random Forest Algorithm. The labelled values are the parameterized threshold values that distinguish one class label from another.





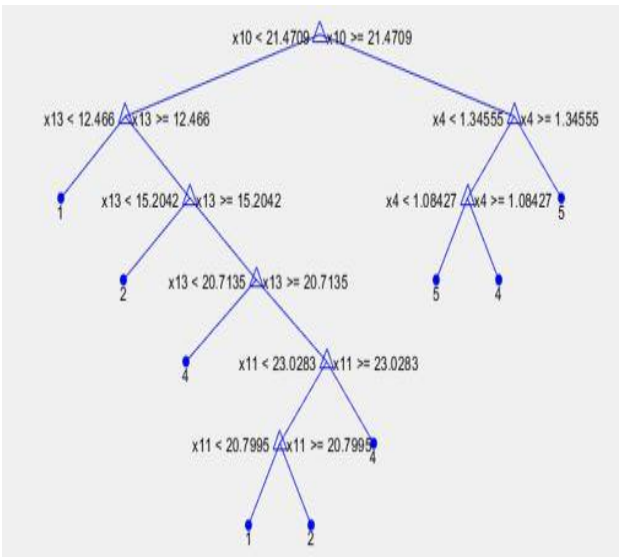


Fig 5(k) Classification Tree obtained for the three images.

IV. RESULTS AND CONCLUSION

For the three input images in Figure 5(a), we obtain the following respective results:

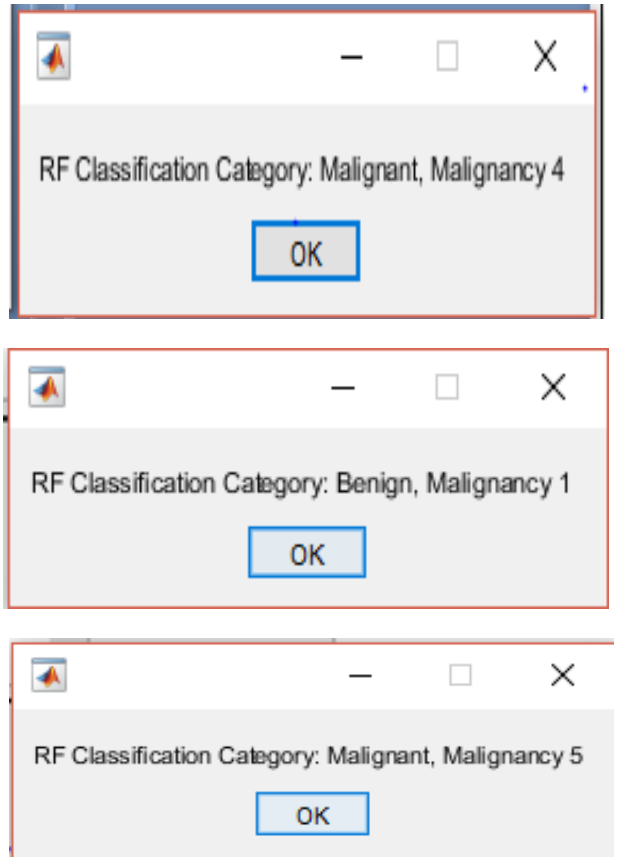


Fig 6. Final results obtained for the three input images.

Hence, we display the class label as output, which signifies the kind of pulmonary nodule and the stage of cancer that the patient is currently in (indicated by indices from 1 to 5). A higher index indicates more dangerous stage, while a lower index indicates an early stage.

FUTURE SCOPE

In this paper, we were successfully able to perform an automatic segmentation using Random Walker, feature extraction and finally automatic classification of pulmonary nodules using Random Forest Algorithm. In future, improvements can be made on the classification performance of pulmonary nodules and to optimize the proposed model. In addition, extended work can be done on grading the images based on the degree of the malignancy of pulmonary nodules, which is of valuable importance for the diagnosis and treatment of lung cancer in clinic applications.

REFERENCES

1. Kostis, W.J., Reeves, A.P., Yankelevitz, D.F., et al.: 'Three-dimensional segmentation and growth-rate estimation of small pulmonary nodules in helical CT images', IEEE Trans. Med. Imaging, 2003, 22, pp. 1259–1274.
2. Kuhnigk, J.M., Dicken, V., Bornemann, L., et al.: 'Morphological segmentation and partial volume analysis for volumetry of solid pulmonary lesions in thoracic CT scans', IEEE Trans. Med. Imaging, 2006, 25, pp. 417–434.
3. Diciotti, S., Picozzi, G., Falchini, M., et al.: '3-D segmentation algorithm of small lung nodules in spiral CT images', IEEE Trans. Inf. Technol. Biomed., 2008, 12, pp. 7–19.
4. Dehmeshki, J., Amin, H., Valdivieso, M., et al.: 'Segmentation of pulmonarynodules in thoracic CT scans: a region growing approach', IEEE Trans. Med.Imaging, 2008, 27, pp. 467–480.
5. Kubota, T., Jerebko, A.K., Dewan, M., et al.: 'Segmentation of pulmonary nodules of various densities with morphological approaches and convexity models', Med. Image Anal., 2011, 15, pp. 133–154.
6. Farag, A.A., Abd El Munim, H.E., Graham, J.H., et al.: 'A novel approach for lung nodules segmentation in chest CT using level sets', IEEE Trans. Image Process., 2013, 22, pp. 5202–5213.
7. Netto, S.M.B., Silva, A.C., Nunes, R.A., et al.: 'Automatic segmentation of lung nodules with growing neural gas and support vector machine', Comput.Biol. Med., 2012, 42, pp. 1110–1121.
8. Chen, K., Li, B., Tian, L.F., et al.: 'Vessel attachment nodule segmentation using integrated active contour model based on fuzzy speed function and shape-intensity joint Bhattacharya distance', Signal Process., 2014, 103, pp.273–284.
9. Sun, S.S., Guo, Y., Guan, Y.B., et al.: 'Juxta-vascular nodule segmentation based on flow entropy and geodesic distance', IEEE J. Biomed. Health Inf., 2014, 18, pp. 1355–1362.
10. Messay, T., Hardie, R.C., Tuinstra, T.R.: 'Segmentation of pulmonary nodules in computed tomography using a regression neural network approach and its application to the lung image database consortium and image database resource initiative dataset', Med. Image Anal., 2015, 22, pp. 48–62.
11. Diciotti, S., Lombardo, S., Falchini, M., et al.: 'Automated segmentation refinement of small lung nodules in CT scans by local shape analysis', IEEE Trans. Biomed. Eng., 2011, 58, pp. 3418–3428.
12. Zhang, F., Song, Y., Cai, W., et al.: 'Lung nodule classification with multilevel patch-based context analysis', IEEE Trans. Biomed. Eng., 2014, 61, pp. 1155–1166.
13. Muramatsu, C., Hara, T., Endo, T., et al.: 'Breast mass classification on mammograms using radial local ternary patterns', Comput. Biol. Med., 2016, 72, pp. 43–53
14. Beura, S., Majhi, B., Dash, R.: 'Mammogram classification using two dimensional discrete wavelet transform and gray-level co-occurrence matrix for detection of breast cancer', Neurocomputing, 2015, 154, pp. 1–14
15. Sethi, G., Saini, B.S.: 'Computer aided diagnosis system for abdomen diseases in computed tomography images', Biocybern. Biomed. Eng., 2015, 36, pp. 42–55
16. Torheim, T., Malinen, E., Kvaal, K., et al.: 'Classification of dynamic contrast enhanced MR images of cervical cancers using texture analysis and support vector machines', IEEE Trans. Med. Imaging, 2014, 33, pp. 1648–1656.
17. Bianconi, F., Fernandez, A.: 'Evaluation of the effects of Gabor filter parameters on texture classification', Pattern Recognit., 2007, 40, pp. 3325-3335.
18. Liu, L., Lao, S., Fieguth, P.W., et al.: 'Median robust extended local binary pattern for texture classification', IEEE Trans. Image Process., 2016, 25, pp.1368–1381.
19. Rastghalam, R., Pourghassem, H.: 'Breast cancer detection using MRF-based probable texture feature and decision-level fusion-based classification using HMM on thermography images', Pattern Recognit., 2016, 51, pp. 176–186.
20. Guo, Z., Wang, X., Zhou, J., et al.: 'Robust texture image representation by scale selective local binary patterns', IEEE Trans. Image Process., 2016, 25, pp. 687–699.
21. Rastghalam, R., Pourghassem, H.: 'Breast cancer detection using MRF-based probable texture feature and decision-level fusion-based classification using HMM on thermography images', Pattern Recognit., 2016, 51, pp. 176–186.

AUTHOR PROFILE

Corresponding Author, Dr. M.Jasmine Pemeena Priyadarsini obtained B.E. degree from Madras University in 1992 and M.E. degrees from Madurai Kamaraj University, Madurai in 1995.. She earned his Ph.D from Vellore Institute of Technology, Vellore, INDIA in 2014. She has published more than 45 research papers in National and International journals and reputed conferenes. He has a teaching experience of about 23 years in Vellore Institute of Technology, Vellore in India. Ppresently, he is serving as Professor at Vellore Institute of Technology, India. She is a life member of Indian Society for Technical Education, IEEE society Membership, Fellow of Institution of Engineers, Fellow of Institution Electronics and Telecommunication Engineers. She has authored about four technical books. His research areas include Digital Image Processing, Digital signal processing,, Optical Signal Processing, Lightwave Communication Systems, Optical Coding Theory and Biometric Image Processing. She is a reviewer of several international conferences and journals.