

Automatic Classification of Lung Nodules into Benign or Malignant Using SVM Classifier

B. Sasidhar, G. Geetha, B.I. Khodanpur and D.R. Ramesh Babu

Abstract Carcinoma of lungs is allied to the cancers that are causing the highest number of deaths all over the world. It is very important to improvise the detection methods so that the rate of survival can be increased. In this paper, new algorithm has been proposed to segment the lung regions using Active Contour method. Once the detection of nodules is through and Gray level Co-occurrence Matrix (GLCM) is used to calculate the texture features. HARALICK texture features are calculated and dominant features are extracted. Support Vector Machine (SVM) Classification of the nodules is done using SVM classifier. Satisfactory results have been obtained. Lung CT scan images are taken from LIDC-IDRI database.

Keywords Active contour • HARALICK texture features • GLCM

1 Introduction

Lung cancer is an solitary type of cancer with highest transience rate in the world. According to American cancer society around 1,58,040 people are dying every year, which means around 18 people die every hour. This mortality rate is alarming. If the mortality rate has to be reduced the prior exposure of lung cancer is vital. There have been various techniques researched and tried to detect cancer as early as possible. In India, Mizoram and Manipur have the highest incidence rate. Other

B. Sasidhar (✉) · G. Geetha · B.I. Khodanpur · D.R. Ramesh Babu
Dayananda Sagar College of Engineering, Bangalore, India
e-mail: bolasasi@gmail.com

G. Geetha
e-mail: geetha071992@gmail.com

B.I. Khodanpur
e-mail: bi.khodanpur@gmail.com

D.R. Ramesh Babu
e-mail: bobrammysore@gmail.com

highest incident sites in India are Bangalore, Chennai, kollam, Kolkata, Tripura and Thiruvananthapuram.

There are many screening techniques available but Computed Tomography (CT) screening technique has stood the test of time. In CT scan, images are captured at various angles and they produce several slices of images thereby giving a detailed and accurate view of the internal parts of the body. In CT images artifacts are reduced to a great extent, giving a clear picture of the body part. CT images can be reconstructed to obtain a different volume of data too. In CT images overlapping anatomical structures are completely eliminated, these images will appear as though an actual cross section of the organ is presented for us to view and analyze. CT images gives information about the size, extent, texture, shape and the exact location. This information is very valuable for the analysis of the condition of the disease.

Computer Aided Detection (CAD) systems are computer programs written to detect and classify a disease. CAD systems help the radiologist in making a decision by giving a affirming opinion about the patient. CAD systems assist radiologists in case if any suspicion or confusion arises. CAD systems are developed for better diagnosis and classification of the disease. In simple terms, CAD systems are used to avoid oversights of radiologists and to interpret what is actually seen.

While building a CAD system, classifier plays an important role. Choosing a suitable and appropriate classifier for the CAD system is very crucial to determine the accuracy and efficiency of the system. Support Vector Machine (SVM) is one such binary classifier which is used to classify linearly separable data. Linearly separable means that the data can be separated into two partitions with a clear cut line of difference. Once the data is partitioned into two groups, the data to be tested is evaluated and is assigned to a particular group, thereby classifying it. SVM gives nearly accurate results even if the training samples are less. It classifies very well even when the features selected are more than the samples itself. SVM can handle the problems of over fitting. Training of SVM classifier is easier and faster. Error rate produced by SVM classifier is very less. Compared to other classifiers SVM is always proven to be a good one.

This paper is organized into the following sections. Section 2 addresses the related work. Section 3 depicts about the proposed methodology. Section 4 addresses the results and discussion. Section 5 depicts about conclusion.

2 Related Work

Elizabeth et al. [1] have suggested a new CAD system which is competent of selecting a trivial slice for the scrutiny of each nodule from a set of slices of a CT scan. CT image is pre-processed by segmenting the region of interest (ROI) using greedy snake algorithm, and then radial basis function neural network (RBFNN) for classification of the nodules. Thomas and Kumar [2] have made a assessment among classifiers like SVM, Minimum distance and k-nearest neighbour.

Morphological Operators is used for pre-processing of the images and gray level co occurrence matrix is used for the feature extraction. Narayanan and Jeeva [3] have proposed a CAD system which uses morphological operators and artificial neural networks for classifying the cropped lung regions as benign or malignant.

El-Bazl et al. [4] have proposed a CAD system which uses surface features and K-nearest classification of nodules. Kumar et al. [5] have proposed a system which uses autoencoder features to classify lung nodules. Zhang et al. [6] have proposed an automatic method which uses weighed Clique Percolation Method (CPMw) for classification. Punithavathy et al. [7] have used Contrast Limited Adaptive Histogram Equalization (CLAHE) and Fuzzy c means clustering for classification.

El-Baz et al. [8] have proposed a method which uses visual appearance features and rotation invariant second order Markov-Gibbs random field for classification than the conventional growth factor. Kaya and Can [9] have proposed a CAD system which uses radiographic descriptors for classification. Mukherjee et al. [10] have proposed a CAD system which uses bilateral filtering for noise removal, thresholding for segmentation and feature extraction for classifying the lung nodules into benign or malignant.

Kaur et al. [11] have proposed a CAD system which uses thresholding and textural and statistical features for classification using artificial neural networks.

Vivekanandan and Raj [12] have proposed a CAD system which uses Snake algorithm for preprocessing, Grey Level Co-occurrence Matrix (GLCM) for extraction of ROI and Nearest Neighbour (NN) for classification. Han et al. [13] have used 2D and 3D Haralick texture feature model for texture feature analysis and SVM for classification. Farag et al. [14] they have used active appearance models (AAM) to detect nodules and K-NN classifier for classification. There are various method used by researchers across the globe to identify a better technique for earlier lung carcinoma detection. However there are few challenges that are being faced by them always.

3 Methodology

The proposed methodology tells about the techniques used to segment and detect the lung nodules (Fig. 1):

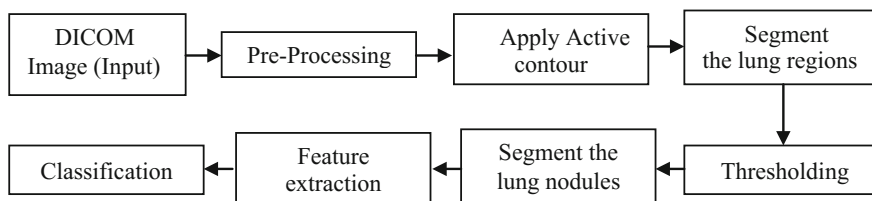


Fig. 1 Flow diagram of the proposed methodology

DICOM image is taken as input. This image is pre-processed to remove the noise and unwanted anatomical structures. To segment the lung nodules Active Contour algorithm is used. After pre-processing the image Active contour is applied to obtain an initial boundary of the lung regions. Then area is calculated to identify the largest connected components. Then the connected components with very large area only is retained and other connected components with lesser area is discarded. Once this is obtained morphological operators are used to fill the holes in the lung region. Now this acts like a mask. This mask is superimposed on the original DICOM image to segment out the lung regions. Now the original lung regions with the nodules are segmented accurately.

Nodules are extracted by calculating eccentricity. Blood vessels and other anatomical structures other than the nodules are discarded. Now Thresholding is done to discard the non nodules. Finally we obtain a mask for nodules, This mask for the nodules is superimposed on the segmented original lung regions. After this we obtain the original lung nodule.

For the nodule obtained, Haralick features are calculated and the texture features are extracted. Dominant features are suitably selected and the SVM classifier is feeded with the values of the dominant features. After the SVM classifier is trained the SVM is subjected to classify the nodules into benign or malignant

Algorithm 1-To segment lung region using Active contour and morphological operators:

Input: DICOM image of a lung.

Output: Segmented lung region.

Step 1: Input a DICOM format lung CT scan image.

Step 2: Perform edge detection using canny edge detector

Step 3: Use Active Contour to obtain a initial boundary of the lungs.

Step 4: Clear image borders and find the connected components.

Step 5: Identify the connected components with large area and fill the holes using morphological operators

Step 6: Superimpose the obtained mask over the original image to obtain the original lung regions with the nodules.

Algorithm 2-To segment the nodules from the lung regions:

Input: Segmented lung region

Output: Segmented lung nodules

Step 1: Calculate eccentricity for the components inside the lung region.

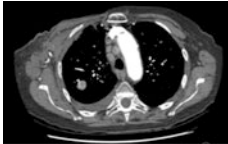

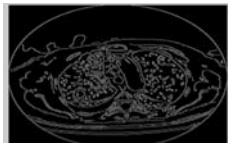


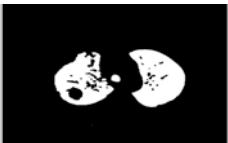
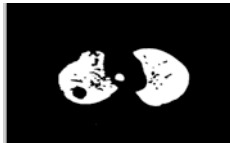
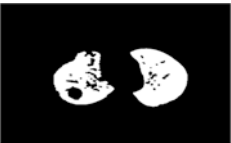



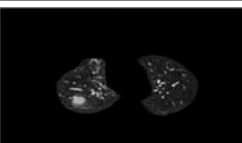
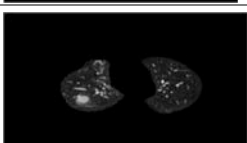

Step 2: Remove the components whose eccentricity is greater than and lesser than that of the nodules.

Step 3: Superimpose this nodule mask on the segmented lung region to get the original nodules.

Once the nodules are obtained using Gray level Co-occurrence Matrix (GLCM), calculate the HARALICK texture features. Dominant feature values are selected and then SVM classifier is trained to classify the nodules into benign or malignant.

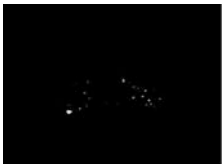


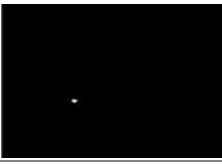
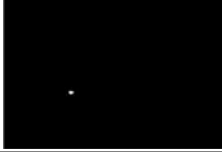
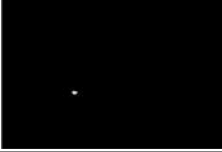
4 Results and Discussion

The CT scan images are taken from LIDC-IDRI database. The obtained results are shown below:

No	Input image	Output image	Details
1			Output obtained is an edge detected image
2			Output obtained is initial boundary of the lung regions
3			Output obtained is connected components in the lung region
4			Output obtained is the components with large area
5			Output obtained is a holes filled image
6			Output obtained is original lung regions
7			Output obtained is anatomical components inside lungs

(continued)

(continued)

No	Input image	Output image	Details
8			Output obtained is nodules and non nodules
9			Output obtained is just nodule's mask
10			Output obtained is original lung nodule

4.1 Feature Selection

HARALICK texture features gives us information about the texture of the cancerous lung regions. GLCM is used for the calculation of texture features. Features are calculated and dominant 5 features have been selected. The graphs for the selected features are as shown below (Graphs 1, 2, 3, 4 and 5):

The below formula is used to calculate the accuracy of the system:

$$accuracy (acc.) = \frac{TP + TN}{TP + TN + FP + FN}$$

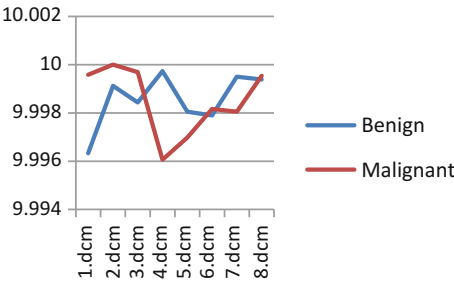
(1)

where

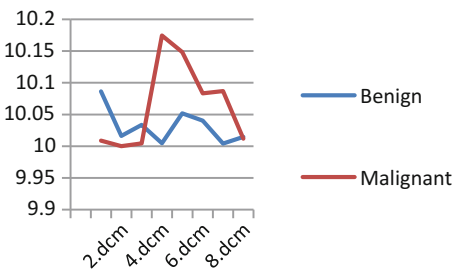
TP = True positive

TN = True negative

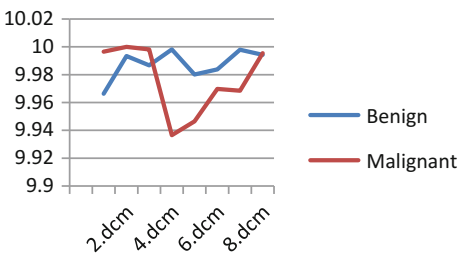
Graph 1 Homogeneity values



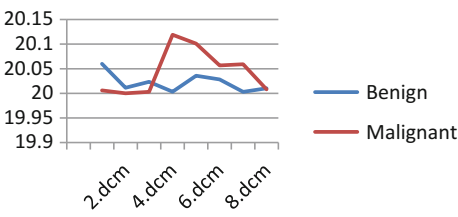
Graph 2 Correlation values



Graph 3 Maximum probability values



Graph 4 Sum average values



Graph 5 Variance values

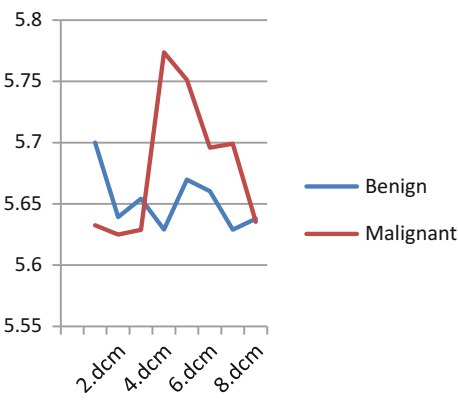


Fig. 2 ROC curve of existing system

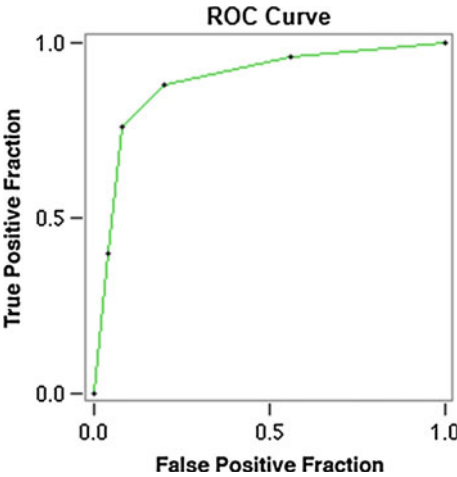
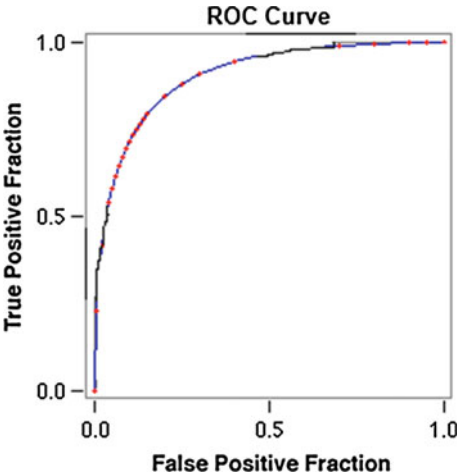


Fig. 3 ROC curve of the proposed system



FP = False positive
FN = False negative

ROC curve for the proposed and existing system is as given below (Figs. 2 and 3):
The ROC curve of the existing system does not have a smooth curve; this is one of the indications that the system provides lesser accuracy. The ROC curve of the proposed system is more inclined towards the upper left corner; this means that the proposed system has better accuracy than the existing system.

5 Conclusion

This paper uses Active contour method to segment the lung regions and HAR-ALICK texture features. Support Vector Machine is used for classifying the nodules into benign or malignant. The proposed algorithm gives better classification comparatively. Segmentation produces results accurately and detection of nodules is very accurate. SVM classification gives nearly accurate results. The algorithm can be further improvised to get improved results. The results obtained are comparatively good and the accuracy obtained is around 92 %.

References

1. D.S. Elizabeth, H.K. Nehemiah, C.S. Retmin Raj, A. Kannan, Computer-aided diagnosis of lung cancer based on analysis of the significant slice of chest computed tomography image. IET Image Process. (2011)
2. R.A. Thomas, S.S. Kumar, *Automatic Detection of Lung Nodules Using Classifiers* (IEEE, 2014)
3. A.L. Narayanan, J.B. Jeeva, *A Computer Aided Diagnosis for Detection and Classification of Lung Nodules* (IEEE, 2015)
4. A. El-Bazl, M. Nitzkenl, E. Vanbogaertl, G. Gimel'jarb, R. Falfi, M. Abo El-Ghar, *A Novel Shape-Based Diagnostic Approach for Early Diagnosis of Lung Nodules* (IEEE, 2011)
5. D. Kumar, A. Wong, D.A. Clausi, *Lung Nodule Classification Using Deep Features in CT Images* (IEEE, 2015)
6. F. Zhang, W. Cai, Y. Song, M.-Z. Lee, S. Shan, D.D. Feng, *Overlapping Node Discovery for Improving Classification of Lung Nodules* (IEEE, 2013)
7. K. Punithavathy, M.M. Ramya, S. Poobal, Analysis of statistical texture features for automatic lung cancer detection in PET/CT images. Int. Conf. Robot. Autom. Control Embed Syst.—RACE (2015)
8. A. El-Baz, G. Gimel'farb, R. Falk, M. El-Ghar, *Appearance Analysis for Diagnosing Malignant Lung Nodules* (IEEE, 2010)
9. A. Kaya, A.B. Can, in *Characterization of Lung Nodules* (IEEE, 2013)
10. J. Mukherjee, A. Chakrabarti, S.H. Skaikh, M. Kar, *Automatic Detection and Classification of Solitary Pulmonary Nodules from Lung CT Images* (IEEE, 2014)
11. J. Kaur, N. Garg, D. Kaur, *An Automatic CAD System for Early Detection of Lung Tumor Using Back Propagation Network* (IEEE, 2014)
12. D. Vivekanandan, S.R. Raj, *A Feature Extraction Model for Assessing the Growth of Lung Cancer in Computer Aided Diagnosis* (IEEE, 2011)
13. F. Han, G. Zhang, H. Wang, B. Song, H. Lu, D. Zhao, H. Zhao, Z. Liang, *A Texture Feature Analysis for Diagnosis of Pulmonary Nodules Using LIDC-IDRI Database* (IEEE, 2013)
14. A. Farag, A. Ali, J. Graham, A. Farag, S. Elshazly, R. Falk, *Evaluation of Geometric Feature Descriptors for Detection and Classification of Lung Nodules in Low Dose CT Scans of the Chest* (IEEE, 2011)