Classification of Lung Tumor Using SVM

[1]Ms. Swati P. Tidke, [2]Prof. Vrishali A. [3]Chakkarwar

Department of computer science and engineering, Government College of engineering,
Aurangabad, Maharashtra

Abstract—Lung cancer is the most important cause of cancer death for both men and women. Early detection is very important to enhance a patient's chance for survival of lung cancer. This paper provides a Computer Aided Diagnosis System (CAD) for early detection of lung cancer nodules from the Chest Computer Tomography (CT) images. There are five main phases involved in the proposed CAD system. They are image pre-processing, extraction of lung region from chest computer tomography images, segmentation of lung region, feature extraction from the segmented region, classification of lung cancer as benign or malignant. Initially total variation based denoising is used for image denoising, and then segmentation is performed using optimal thresholding and morphological operations. Textural features extracted from the lung nodules using gray level co-occurrence matrix (GLCM). For classification, SVM classifier is used. The main aim of the method is to develop a CAD (Computer Aided Diagnosis) system for finding the lung tumor using the lung CT images and classify the tumor as Benign or Malignant.

Keywords—Computer Aided Diagnosis System, optimal thresholding, gray level co-occurrence matrix (GLCM), Support vector machine (SVM).

I. Introduction

Lung cancer is the leading cause of tumor-related deaths in the world [1]. At the same time, it appears that the rate has been steadily increasing. Lung cancer is caused by the uncontrolled growth of tissues in the lung. The American cancer society estimates that 213, 380 new cases of lung cancer in the U.S will be diagnosed and 160, 390 deaths due to lung cancer will occur in 2007.The majority of all cases are caused by tobacco smoking. Exposure to asbestos, radon, uranium and arsenic are other factors for lung cancer. Lung cancer is a deadly disease and has chances to spread to other parts of the body, e.g. the brain, liver, bone and bone marrow. The early detection and diagnosis of nodules in CT image are among the most challenging clinical tasks performed by radiologists. Radiologists can miss up to 25% of lung nodules in chest radiographs due to the background anatomy of the lungs which can hide the nodules. Computer aided diagnosis system helps the radiologists by doing preprocessing of the images and recommending the most possible regions for nodules. The complexity for finding the lung nodules in radiographs are as follows:

1. A nodule diameter may be differed from a few millimeters
2. Nodules vary widely in density.
3. As nodules can be found anywhere in the lung region, they can be hidden by ribs and structures below the diaphragm, resulting in a large variation of contrast to the background.
4. To overcome these difficulties, the author proposed a Computer Aided Diagnosis (CAD) [2] system for detection of lung nodules [3]. The lung tumor prediction system is shown in Figure 1:

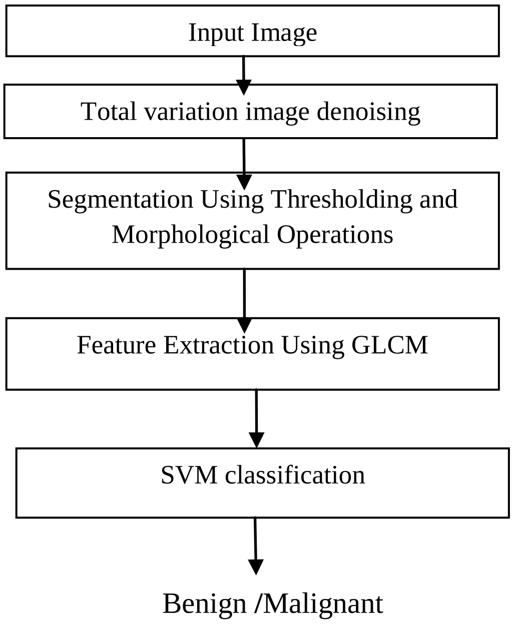
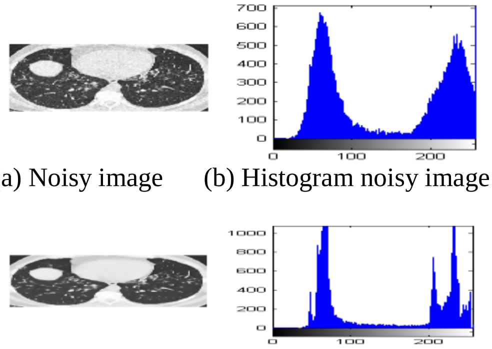

Input Image
Total variation image denoising
Segmentation Using Thresholding and Morphological Operations
Feature Extraction Using GLCM
SVM classification
Benign /Malignant

Fig. 1 lung tumor prediction system

This paper initially apply the different image processing techniques such as image denoising for removing noise in the image, optimal thresholding for converting gray level image into binary image, morphological operations for segmentation, feature extraction using GLCM and Support vector machine is used for classification.

II. IMAGE PREPROCESSING

CT imaging process may contain various noises. These images are not clean so as to use directlyforprocessing,thuswe need to denoise these images. The segmenting scheme introduced in this paper performs an image preprocessing task to remove noise in a lung CT image at first. Total variation denoising is a preprocessing step to reduce the effect of undesired perturbations. Total variation denoising is very effective at simultaneously preserving edges while smoothing away noise in flat regions, even at low signal-to-noise ratios. total variation denoising results are shown in figure 2.



(a) Noisy image (b) Histogram noisy image

(c) Denoised image (d) Histogram denoised image

Fig. 2 Histogram of lung CT before and after denoising

III. Segmentation Of Lung Region

In chest CT images, there are two types of pixels with different density distribution:

1) Pixels within the very dense body and chest wall structures (the body pixels) and 2) low-density pixels. Optimal thresholding proposed by Shiying et al. [3], is applied on the pre-processed lung image to select a segmentation threshold to separate the body and non-body pixels through an iterative procedure. Let Ti be the segmentation threshold after the ith step. To choose a new segmentation threshold, we apply Ti to the image to separate the pixels into body and nobody pixels. Let μ_b and μ_n be the mean grey level of body and non-body pixels. Then new threshold will be:

$$T^{i+1} = \frac{1}{2}(\mu_b + \mu_n)$$

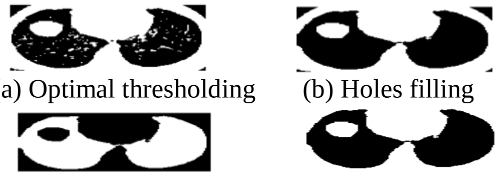
The pixels with a density lower than the threshold value are recognized assigned a value 1 and appear white, whereas other pixels are assigned the value of 0 and appear black. The lung image after segmentation with optimal threshold value contains non-body pixels such as the air surrounding the lungs, body and other low-density regions within the image and is removed through morphological operations such as erosion, dilation and labeling.

IV. Roi Extraction

Lung nodules are the small masses of tissue in the lung. They appear approximately round and white in CT scan or X-ray images. In the proposed method our region of interest is lung nodule and labeling algorithm is applied for region extraction. Connected component labeling is method of addressing different group of pixels based on their characteristics, in this case intensity values of pixels and their neighborhood. There are number of locations where pixels have same intensity values they are gathered as one group and uniquely labelled.Labeling is usually used to detect connected regions in binary images. Color images and data with higher-dimensionality can also be processed. In proposed method 4-connected labeling algorithm is used. Overview of 4-connected labeling algorithm is as follows:

- Given a binary image.
- Negate the image.
- For every pixel check the north and west pixel.
- If the pixel to the west or north has the same intensity value, the pixel belongs to same region. Assign the same label to the current pixel.
- If the pixel to the west has a different value and the pixel to the north has the same value, assign the north pixel's label to current pixel.
- If the north and west neighbors of pixel have different pixel values, create a new label and assign that label to the current pixel.
- Do this recursively for all pixels that have a 4-connectivity.

All pixels above the threshold that have a 4-connectivity will get the same number and thereby each region a unique label. Find the centroid of each label, if centroid of label is not present at significant height and width considering our region of interest eliminates that label. In this way we will get desired lung region. The extracted ROIs are then subject to feature extraction for analysis.



(a) Optimal thresholding (b) Holes filling

(c) Labeling (d) Lung region

Fig. 3 Optimal thresholding and morphological operations

V. Feature Extraction

The extracted ROIs can be distinguished as either cancerous or not using their texture properties. Gray Level Co-occurrence Matrix (GLCM) is one of the most popular ways to describe the texture of an image. A GLCM denote the second order conditional joint probability densities of each of the pixels, which is the probability of occurrence of grey level i and grey level j within a given distance 'd' and along the direction 'θ'.7 features are considered for proposed method.

1. Area: It gives the actual number of pixels in the ROI.

2. Convex *Area*: It gives the number of pixels in convex image of the ROI.

3. Equivalent Diameter: It is the diameter of a circle with the same area as the ROI.

$$\text{Equivalent Diameter} = \frac{\sqrt{4*\text{Area}}}{\sqrt{\pi}} \qquad (3)$$

4. Solidity: It is the proportion of the pixels in the convex hull that are also in the ROI.

$$\text{Solidity} = \frac{\text{Area}}{\text{ConvexArea}} \qquad (4)$$

5. Energy: It is the summation of squared elements in the GLCM and its value ranges between 0 and 1.

$$\text{Energy} = \sum_{k=0}^{n} p^2(i,j) \qquad (5)$$

6. Contrast: It is the measure of contrast between an intensity of pixel and its neighboring pixels over the whole ROI.where N is the number of different gray levels.

$$\text{Contrast} = \sum_i^N \sum_j^N (i-j)^2\, p(i,j) \qquad (6)$$

7. Homogeneity: It is the measure of closeness of the distribution of elements in the GLCM to the GLCM of each ROI and its Value ranges between 0 and 1.

$$\text{Homogeneity} = \sum_{i,j} \frac{p(i,j)}{1+|i-j|} \qquad (7)$$

8. Correlation: It is the measure correlation of pixel to its neighbor over the ROI.

$$\text{Correlation} = \sum_i^N \sum_j^N \frac{p(i,j)-\mu r\ \mu c}{\sigma r\ \sigma c} \qquad (8)$$

9. Eccentricity: The eccentricity is the ratio of the distance between the foci of the ellipse and its major axis length.

II. Svm Classification

Support vector machines are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification The basic SVM takes a set of input data and for each given input, predicts, which of two classes forms the input, making it a non-probabilistic binary linear classifier. From given set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other. In the proposed method we are using linear classifier. Best hyper plane is the one that represents the largest separation or margin between the two classes. So we choose the hyperplane so that the distance from it to the nearest data point on each side is maximized. If such a hyper plane exists, it is known as the maximum margin hyperplane and the linear classifier it defines is known as a maximum classifier, which is shown in fig.4

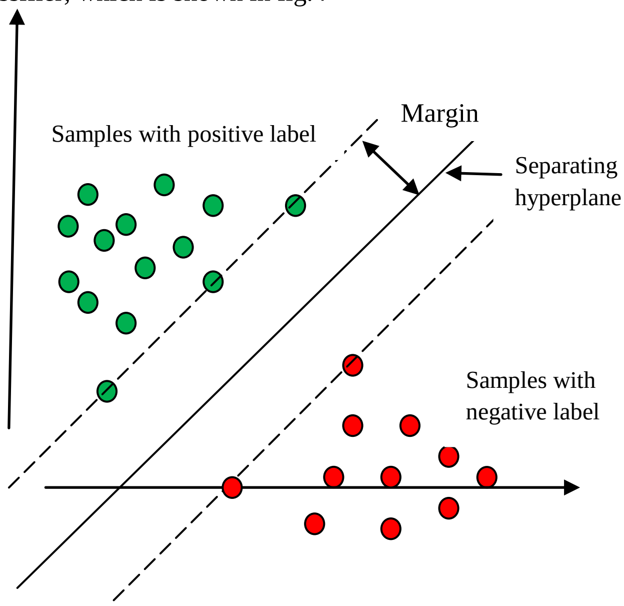
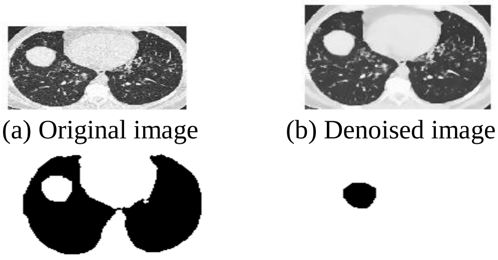


Fig. 4 Maximum margin classifier

III. Experimental Results And Analysis

The data set used for study of the proposed system consists of 25 diseased lung CT JPEG images of size196x257. A total of 40 ROIs were extracted. The system was tested with 15 diseased lung images. The results obtained for a diseased lung image is shown in Fig. 5


(a) Original image (b) Denoised image

(c) segmented image (d) tumor region

Fig. 5 Results obtained for a diseased lung image In case of SVM classifier out of 9 features at a time only twofeatures are selected for classification, which produces result as either benign or malignant. In the proposed method first two features provides the best results. A few classification results using Support vector machine is listed in table I

TABLE I
TUMOR CLASSIFICATION USING SVM

Image	Output Value	Classification
Image_1	0	Benign
Image_2	1	Malignant
Image_3	0	Benign
Image_4	1	Malignant
Image_5	1	Malignant
Image_6	1	Malignant
Image_7	1	Malignant
Image_8	1	Malignant
Image_9	0	Benign
Image_10	1	Malignant

Conclusion

Proposed system helps physician to extract he tumor region and evaluate whether the tumor is benign or malignant. The computer tomography image is used in this paper. SVM provides the accuracy of 92.5%.The accuracy of system can be improved if training is performed by using a very large image database. The different basic image processing techniques are used for prediction purpose. In the first phase the image is denoised. In second phase lung region is separated from surrounding anatomy. In next phase ROI is extracted, after ROI extraction features extraction is performed by GLCM. Finally with the obtained texture features classification is performed to detect the occurrence of cancer nodules.

REFERENCES

[1] Weir HK et. al.,"Annual report to the nation on the status of cancer, 1975-2007"*Journal National Cancer Institute*, vol. 95 , No. 17.

[2] Ayman El-Baz, Aly A. Farag, Robert Falk, Renato La Rocca, “Detection, Visualization and identification of Lung Abnormalities in Chest Spiral CT Scan: Phase-I”, *International Conference on Biomedical Engineering*, Cairo, Egypt, 12-01-2002.

[3] Shiying Hu, Eric A. Huffman, and Joseph M. Reinhardt," Automatic lung segmentation for accurate quantification of volumetric X-Ray CT images", *IEEE Transactions on Medical Imaging*, vol. 20, no. 6 ,pp. 490-498, June 2001

[4] Gurcan, M.N, Sahiner, B., Patrick, N., Chan, H., Kazerooni, E.A., Cascade, P.N. and Hadjiiski,L., “Lung Nodule Detection on Thoracic Computed Tomography Images: Preliminary Evaluation of a Computer-Aided Diagnosis System”, *Medical Physics*, Vol. 29, No. 11, Pp. 2552- 2558, 2002.

[5] J.Padmavathi, ”A comparative study on breast cancer prediction using RBF and MLP”, *International Journal of Scientific and Engineering Research*”, Vol.2, Issue.1,2011.

[6] R.M. Haralick, K. Shanmugam and I.H. Dinstein," Textural features for image classification," *IEEE Transactions on Systems, Man and Cybenatics*, vol.3, no. 6, pp. 610-621, Nov. 1973.

[7] R.C. Gonzales and R.E. Woods, *Digital Image Processing*: Pearson Education, 2003.

[8] Kanazawa, K., Kawata, Y., Niki, N., Satoh, H., Ohmatsu, H., Kakinuma, R., Kaneko, M., Moriyama, N. and Eguchi, K., “Computer-Aided Diagnosis for Pulmonary Nodules Based on Helical CT Images”, *Compute Med. Image Graph*, Vol. 22, No. 2, Pp. 157-167, 1998.

[9] Ayman El-Baz, Aly A. Farag, Robert Falk, Renato La Rocca, “A Unified Approach for Detection, Visualization and Identification of Lung Abnormalities in Chest Spiral CT Scan”, *Proceedings of Computer Assisted Radiology and Surgery*, London 2003.

[10] Samuel G. Armato III, Maryellen L. Giger and Catherine J. Moran, “Computerized Detection of Pulmonary Nodules on CT Scans”, *Radio Graphics*, vol. 19, pp. 1303-1311, and 1999.

[11] S.K.Vijai Anand “Segmentation coupled Textural Feature Classification for Lung Tumor Prediction”, *International Conference on Communication, Control and Computing Technologies 2010.*

[12] M. Gomathi, P.Thangaraj “A Computer Aided Diagnosis System for Lung Cancer Detection using Machine Learning Technique”, *European Journal of scientific research*, Vol.51 No.2, pp.260-275, 2011.

[13] B.Magesh, P.Vijayalakshmi, M. Abirami “computer aided diagnosis system for identification and classification of lesions in lungs”, *International Journal of Computer Trends and Technology*- May to June Issue 2011.

[14] Nisar Ahmed Memon, Anwar Majid Mirza, S.A.M. Gilani “Deficiencies Of Lung Segmentation Techniques using CT Scan Images for CAD”, *world Academy of Science, Engineering and Technology 2006.*

[15] Laura Auria1 and Rouslan A. Moro, “Support Vector Machines (SVM) as a Technique for Solvency Analysis”, Berlin August 2008

[16] H. Samet and M. Tamminen (1988). "Efficient Component Labeling of Images of Arbitrary Dimension Represented by Linear Bintrees". *IEEE Transactions on Pattern Analysis and Machine Intelligence.*

[17] Michael B. Dillencourt and Hannan Samet and Markku Tamminen (1992).”A general approach to connected-component labeling for arbitrary image representations”.

[18] Kenji Suzuki and Isao Horiba and Noboru Sugie. "Linear-time connected-component labeling based on sequential local operations". *Computer Vision and Image Understanding, 2003.*

[19] Boser, Bernhard E.,Guyon, Isabelle M., and Vapnik, Vladimir N.“A training algorithm for optimal margin classifiers”.