ISSN: 2349 - 4891

International
Journal of Recent Research and Applied Studies
(Multidisciplinary Open Access Refereed e-Journal)

Lung Tumor Segmentation and Classification by Using Machine Learning Approach

S.S.Deepthi[1], C.Vinola[2], K. Raja Sundari[2] & K.Siva Kumar[2]

[1]*PG Student, Department of CSE, Francis Xavier Engineering College, Tirunelveli, Tamilnadu, India.*
[2]*Assistant Professor, Department of CSE, Francis Xavier Engineering College, Tirunelveli, Tamilnadu, India.*

Received 12th February 2016, Accepted 1st April 2016

Abstract

The project proposes an improve method of Lung image classification and image segmentation approach. It is automatic support system for stage classification using learning machine and to detect Lung Tumor through Lloyds clustering method for bio medical application. Automated classification and detection of tumours in different medical images is motivated by the necessity of high accuracy when dealing with a human life. The detection of the Lung Tumor is a challenging problem, due to the structure of the Tumor cells. This project presents a segmentation method, Lloyds clustering algorithm, for segmenting Lung images to detect Tumor in its early stages and to analyze anatomical structures. The artificial neural network will be used to classify the stage of Lung Tumor that is benign, malignant and normal. Here DWT decomposition is used to analysis texture of an image. The segmentation results will be used as a base for a Computer Aided Diagnosis (CAD) system for early detection of Lung Tumor which will improves the chances of survival for the patient. Probabilistic Neural Network with radial basis function will be employed to implement an automated Lung Tumor classification. Decision making was performed in two stages: feature extraction using GLCM and the classification using PNN-RBF network. The performance of this classifier was evaluated in terms of training performance and classification accuracies. The simulated results will be shown that classifier and segmentation algorithm provides better accuracy than previous method.

Keywords: Computed Tomography (CT), X-ray, Thresholding, Segmentation, Discrete Wavelet Transformation (DWT), Grey Level Co – Occurrence Matrix (GLCM), Probabilistic Neural Network (PNN), Radial Basics Function (RBF), Lloyds clustering algorithm.

Introduction

Medical imaging plays a vital role in the detection of lung cancer. Lung cancer is the type of cancer that begins as a small tumor in the lung. Among all type cancers, the lung cancer causes the maximum cases of deaths in Men and Women. In US 165,000 people die with lung cancer every year. In a survey, in males more than 70% and in females more than 50% lung cancer is caused by breathing polluted air, Cigarette smoking etc. According to the latest survey in year 2014, total 159,260 people had been died due to lung cancer in US. In INDIA every year 63,000 new lung cancer cases has been reported. The detection of cancer in the early stage is very difficult to analyze. Various CAD systems have been designed for the early detection of lung cancer but, none of them provide better result[9].

Hence, it becomes necessary to detect the lung nodules in earlier stage using chest Computer Tomography (CT) images. To achieve this, Computer Aided Diagnosis (CAD) system is very essential. Radiologists can miss up to 30% of lung nodules in chest radiographs due to the background structure of the lungs which can hide the nodules. Computer aided diagnosis system assists the radiologists to check the lung image in the preprocessing stage and recommends the most possible regions for the occurrence of nodules. Identification of defected portions in the lung regions progresses through various methods. Initially it removes the background regions in lungs such as blood vessels, air volume, white region, ribs and the bronchi. After removing the background region, it provides good chest structure that create better nodule region and also, it shows additional classified information depends on the characteristics like size, contrast and shapes [8].

Here in the studied report, initially there are various processing techniques which are applied to extract the affected region of the lung from the CT images. Then with a clustering algorithm, the segmentation process is performed. Lastly for automatic detection of lung nodules and for better clustering task, Lloyds clustering is used. The experiment is taken out for the presented technique by the CT chest images [16].

Correspondence
S.S.Deepthi
E-mail: deepthissnairr@gmail.com

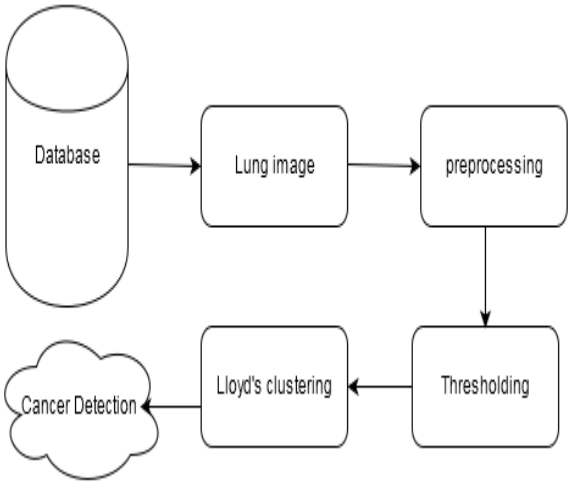

Figure 1. Lloyds Clustering

The database where the images are stored is called Image database. In the preprocessing technique, noise removal, filtering, reshaping of images and the conversion of RGB to gray color takes place. After preprocessing, the lung image is separated from the background using adaptive thresholding method. Then Lloyds clustering is performed is segment the detected part of the lung. Finally, cancer can be detected automatically in short period.

I. Related Works

In [1] Aggarwal P., Vig R., and Sardana H.K. they introduced Semantic and content based medical image retrieval for lung cancer diagnosis with the inclusion of expert knowledge and proven pathology. This paper involves the analysis and experimentation of chest CT scan data for the detection and diagnosis of lung cancer. In lung cancer computer-aided diagnosis (CAD) systems, having an accurate ground truth is critical and time consuming. The contribution of this work include the development of lung nodule database with proven pathology using content based image retrieval (CBIR) and algorithms for detection and classification of nodules. A study and analysis of 246 patients have been carried out for the detection of benign, malignant as well as metastasis nodules. The whole research work has been carried out using Lung Image Database Consortium (LIDC) database by National Cancer Institute (NCI), USA and achieved an average precision of 92.8% and mean average precision of 82% at recall 0.1. Finally, the validations have been carried out with the PGIMER, Chandigarh test cases and achieved an average precision of 88%. Experimental studies show that the proposed parameters and analysis improves the semantic performance while reducing the computational complexity, reading and analysing all slices by physicians and retrieval time.

In [4] Disha Sharma and Gagandeep Jindal they discovered Computer Aided Diagnosis System for Detection of Lung Cancer in CT Scan Images. The automated Computer Aided Diagnosing (CAD) system is proposed for detection of lung cancer form the analysis of computed tomography images. In recent years the image processing mechanisms are used widely in several medical areas for improving earlier detection and treatment stages, in which the time factor is very important to discover the disease in the patient as possible as fast, especially in various cancer types such as the lung cancer, breast cancer. Lung cancer is the second most commonly diagnosed cancer in the United States, and it is the leading cancer related death in the world, with the current fatality rate exceeding that of the next three most common cancers (breast, prostate, and colorectal) combined. In this research, it considered the problem of developing an automated system for detecting the presence of pulmonary nodules in the lung CT. The essence of developing a system like that needed to focus on detecting nodules in their early stages, which are the very small nodules that are likely to be overlooked by the radiologists. This paper involves cancer detection system based on texture features extracted from the slice of DICOM Lung CT images for the identification of cancerous nodules. In developing this system, it passed the available lung CT images and its database in basic three stages to achieve more accuracy in the experimental results. Firstly, a pre-processing stage involving some image enhancement techniques helps to solve the problem. The preprocess images (by contrast enhancement, thresholding, filtering, and blob analysis) obtained after scanning the Lung CT Images and secondly separate the suspected nodule areas (SNA) from the image by a segmentation process by using thresholding segmentation mechanism by Otsu thresholding algorithm and region growing techniques. Finally, relied on Texture features which help us to make a comparison between cancerous and non-cancerous images. For accurate detection of cancerous nodules, it need to differentiate the cancerous nodules from the noncancerous. Thus, it developed an artificial neural network to differentiate them. The neural network is trained by the back propagation algorithm and tested it with different images from a database of the DICOM CT Lung images of NIH/NCI Lung Image Database Consortium (LIDC)

In [8] Gomathi M. and Thangaraj P. they discovered An Effective Classification of Benign And Malignant Nodules Using Support Vector Machine. Support Vector Machine (SVM) is a machine learning technique that trains the system with the known data; it analyzes and identifies the patterns. SVM can be used for classifying the medical data because of its simplicity. Real time lung images are taken for the study. Lung images are segmented to retrieve the region of interest and these regions or nodules are used for classification. Proper threshold values are decided for each feature and classification rules are framed. Then these rules are passed to the SVM classifier. In this paper, classifications of benign and malignant nodules are done using different SVM kernels and their performance

measures are compared.

In [15] Yadav N.G. proposed Detection of lung nodule using Content Based Medical Image Retrieval. Lung cancer is the most important cause of cancer death for both men and women. The early detection of cancer can be helpful in curing the disease completely. So the requirement of techniques to detect the occurrence of cancer nodule in early stage is increasing. There are different technique exists but none of those provide better accuracy of detection. This provides content based medical image retrieval Computer Aided Diagnosis System (CAD) for early detection of lung cancer nodules from the Chest Computer Tomography (CT) images. There are different phases involved in the proposed CAD system. They are extraction of lung region from chest computer tomography images, segmentation of lung region, feature extraction from the segmented region, and classification of occurrence and non-occurrence of cancer in the lung. Keywords: CBIR, segmentation algorithm, Gray level Co-occurrence Matrix (GLCM), Support Vector Machine (SVM)

The region based segmentation is performed on the lung region, the features can be obtained from it for determining the diagnosis for detecting the cancer nodules in the lung region perfectly. The feature that are used in this study are texture features using co-occurrence matrix representation. This approach to segmentation examines neighboring pixels of initial seed points and determine whether the pixel neighbors should be added to the region. The process is iterated on, in the same manner as general data clustering algorithm. The first step in region growing is to select a set of seed points. Seed point selection is based on some user criteria for example, pixels in a certain gray scale range, pixels evenly spaced on a grid, etc.). The initial region begins as the exact location of these seeds. It has poor discriminatory power and it does not provide an optimal result for all lighting conditions of images. It does not applicable for multiple images for cancer detection in a short time.

II. Lloyds Clustering For Cancer Detection

In the proposed system Lloyds clustering is introduced to overcome the problems occurred in the existing system. Cluster analysis is one of the effective method for analyzing and discovering useful information from numerous data. It groups the data into classes or clusters, so that the objects within a cluster will have high similarity in comparison with another cluster. The most widely used clustering error criterion is squared-error criterion, it can be defined as,

$$J_c = \sum_{j=1}^{c} \sum_{k=1}^{n_j} \left\| x_k^{(j)} - m_j \right\|^2$$

Where J_c, is the sum of square-error for all objects, $x_k$is the point in space which represents a given object, and m_j is the mean of cluster c_j**.** Adopting the squared-error criterion, Lloyd's clustering works well when the clusters are grouping of clouds that are rather well separated from one another. It does not suitable for discovering clusters with non-convex shapes or clusters of very different size. Lloyd's algorithm is enough to extract mass from the lung cells and thus it produce more accurate result. It can segment the lungs regions from the image accurately which classify the lung cancer images for accurate detection and the lung cancer will be detected in an early stages.

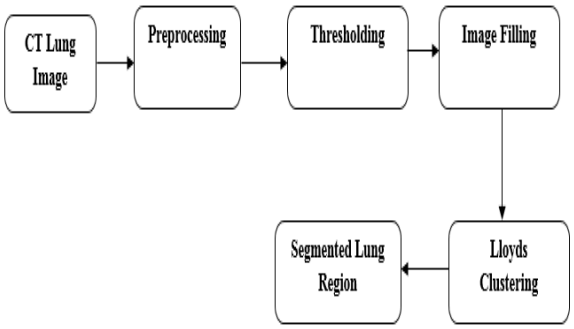

Figure 2. System Architecture

In Fig 2 describes the task of lung cancer detection. The small changes in the pixel leads to false detection. So when the input image is given, it must remove the noise present in the image. Therefore, there is a fundamental need of noise reduction from medical images. Preprocessing performs filtering the noise in the image and other artifacts in the image such as sharpening the edges. The conversion of RGB to grey color and reshaping also takes place in the preprocessing stage. It includes median filter for noise removal.

The simplest method of image segmentation is called the thresholding method. First a gray-level 'T' between two dominant levels has to be selected, which will serve as a threshold is to distinguish the two classes (foreground and background), Where the value marked as T is a value of actual threshold. Using this threshold, a new binary value can be produced. From this value, the outer region of the objects are painted completely black, and the remaining pixels values be white. After thresholding, image filling is done. The `imfill` function performs a flood-fill or seed fill operation on binary and grayscale images. The conversion of background pixels to foreground pixels will done in the binary operation (i.e.,),the pixel will convert from 0s to 1s and get stops when it reaches object boundaries.

The Lloyd's clustering is an iterative technique that is used to partition an image into K clusters. This algorithm is guaranteed to converge, the quality of the solution depends on the initial set of clusters and the value of K.

A. Image Preprocessing

Image preprocessing is the main task in Lung cancer detection. The small changes in the pixel lead to

false detection. Noise can be added in the lung due to various reasons. Due to the noise the pixel values might be changed. So image preprocessing is very essential.

Noise Removal

In medical image processing, it is very important to obtain precise images to facilitate rate observations. For the given application, low image quality is an obstacle for effective feature extraction, analysis and recognition. Therefore, there is a fundamental need for noise reduction in the medical images. There are currently a number of imaging modalities which are used for study of medical image processing. Among the newly developed medical imaging modalities, CT and Ultrasound imaging are believed to be very potential for accurate measurement of organ anatomy in a minimally invasive way. MRI is a powerful diagnostic technique.

Preprocessing performs noise filtering and other artifacts in the images. It also sharpening the edges in the image. RGB to grey conversion and Reshaping also takes place here. Images taken with both digital cameras and conventional film cameras will pick up noise from variety of sources. Further use of these images will often require that the noise be removed for aesthetic purpose such as computer vision.

In salt and pepper noise, the pixels in the images are very different in color intensity from the surrounding pixels. In Gaussian noise, each pixel in the image will be changed from original value by a small amount. In either case, the noise at different pixel can be either correlated or uncorrelated, in many cases, noise values at different pixels are modeled as being independent and identically distributed, and hence correlated

Median Filtering

Median filtering performs some kind of noise reduction in image or signal. It is a nonlinear filtering technique. The main idea of this method is run through the signal entry by entry with median of neighbouring entries. It is more effective in removing 'salt and pepper' type of noise. It works by moving through the image pixel by pixel, and replaces each value with the median value of neighbouring pixels. The pattern of neighbours is called the "window", which slides, pixel by pixel over the entire image pixel, over the entire image. It is a kind of smoothing technique but which affect the edges. Edges are of critical importance of the visual appearance of images.

B. Thresholding

The simplest method segmentation is called thresholding. It replaces each pixel in the image with black pixel if the image intensity is less than some fixed constant value. That is, it will a gray-scale image into a binary image. The key method is to select the threshold value (or select values when multiple-levels are selected). To make thresholding completely automated, it is necessary for the computer to automatically select the select the threshold T. clustering based methods, where the grey level samples are clustered in two parts as background and foreground (object) or alternately are modeled as a mixture of two Gaussians. Color images can also be threshold.

Adaptive Threshold

Adaptive thresholding is a form of thresholding. The technique of real time adaptive thresholding uses the input of integral image. First a gray-level T between those two levels must be chosen, which will serve as a threshold to the two classes (objects and background). Where the value marked as T is initial choice for a threshold. Using this threshold value, a new binary image can then be produced, where the objects are painted completely black, and the remaining pixels are white. Let the original image be f (x, y), then the threshold product is achieved by scanning the original image. It scans the image pixel by pixel, and finally testing each pixel across the selected threshold: if f (x, y) > T , then the pixel is classified as being a background pixel, otherwise the pixel is classified as an object pixel.

In the general case, a threshold is produced for each pixel in the original image; this threshold is then used to test the pixel and produce the valuable result (in this case, a binary image). According to this, the definition of the threshold value can be written as,

$$T= T [x, y, p(x, y) , f (x, y)]$$

Where, f (x, y) is the function of gray level point (x, y) in original image and p(x , y) is the local property of this particular point. When T depends only on the gray-level. Then, it degenerates to form a simple global threshold. Special attention needs to be given to the factor p(x, y).

Design Steps

(1) Set the initial threshold T= (the maximum value of the image brightness + the minimum value of the image brightness)/2.
(2) Using the threshold value T, segment the image to get two sets of pixels say B (all the pixel values are less than T) and say N (all the pixel values are greater than T);
(3) Then again calculate the average value of B and N separately, it forms the mean ub and un.
(4) Now, the new threshold value can be calculated as, T= (ub+un)/2
(5) Repeat Second steps to fourth steps up to iterative conditions are met and get necessary region from the lung image.

C. Image Filling

The imfill function performs a flood-fill (seed fill) algorithm on binary and grayscale images. For binary images, imfill convert the background pixels (0s) to foreground pixels (1s) and this process gets stops

when the filling operation reaches object boundary. But, in case of grayscale images, imfill brings the intensity values in the surrounding dark areas by lighter areas up to the same intensity level as surrounding pixels.

This algorithm also gives the information about,

- Specify connectivity in the flood-fill operation
- Specify the starting point of binary image fill operations
- Specify holes filling operations in binary or grayscale images

D. Lloyd's Clustering

The Lloyd's clustering is an iterative technique that is used to partition an image into K clusters. The basic algorithm is:

1. Pick K cluster centers, either randomly or based on some heuristic
2. Assign each pixel in the image to the cluster that minimizes the distance between the pixel and the cluster center.
3. Re-compute the cluster centers.
4. Repeat the steps above until convergence is attained (e.g. no pixels change clusters)

It is the case where the distance is squared or absolute difference between a pixel and a cluster center are chosen. The difference is typically based on pixel color, intensity, texture, and location, the difference is made. It also differs based on the weighted combination of these factors. It is closely related to the K means clustering, where the K value can be selected manually or randomly. This K means algorithm is guaranteed to converge, but it may not return the desired solution. The quality of the solution will be depends on the value of K.

In statistics and machine learning algorithm, which is used for scalar and vector quantization. It also used in data compression technique in information theory. It is used for drawing dot product to match an input image.

In the finite element method, an input domain with a complex geometry is partitioned into elements with simpler shapes, for instance, two dimensional domains (either subset of the Euclidean plane or surface in three dimensions) are often partitioned into triangles. Lloyds clustering algorithm can change the topology of the mesh.

III. Results

The input may contain noise and it must be removed before segmentation process.

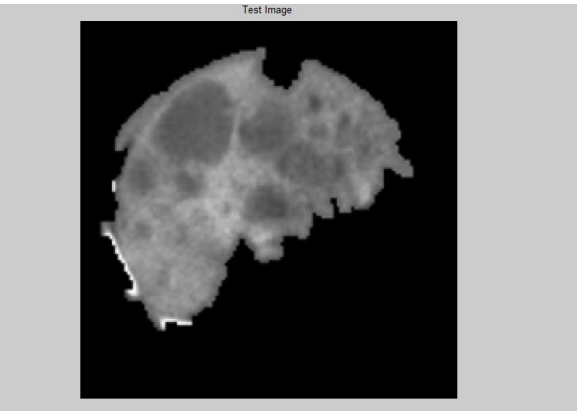

Figure 3. Noisy Image

The Figure 3 describes that the input Image contains noise which leads to degrade the performance of output. The noise is clearly displayed in the image as black and white dots. In order to obtain the accuracy of lung cancer, the noise must be removed from the original input image in the pre-processing stage.

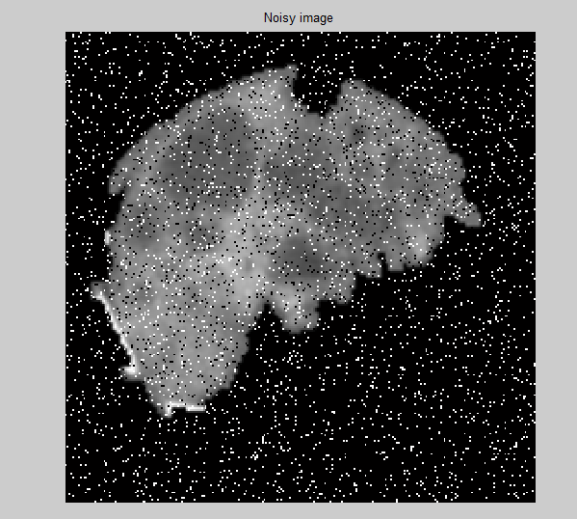

Figure 4. Denoised Image

The Figure 4 is a Denoised Image which remove noise from the image and it is done by using median filter. The median filter works by moving through the image pixel by pixel, replacing each value with the median value of neighboring pixels. It is an efficient method to remove salt and pepper type noise. This filter will improve the performance of an image.

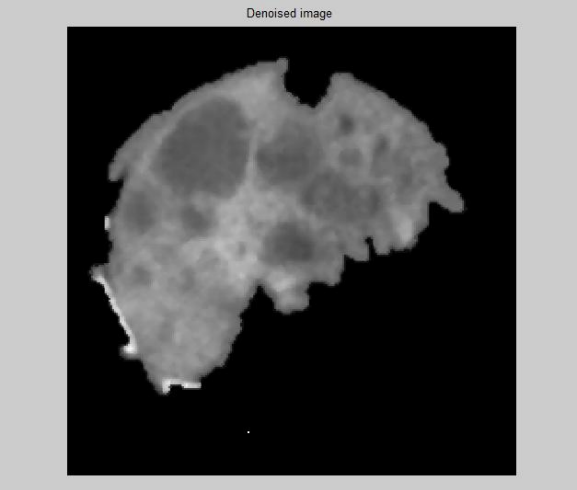

Figure 5. Background removal

The fig 5 shows the background removal process, which separate the lung image from the background. Thus the color of outer area will convert from white to black (i.e.,) the pixel changes from 1's to 0's. This process will stops when it reaches the outer edges of the image.

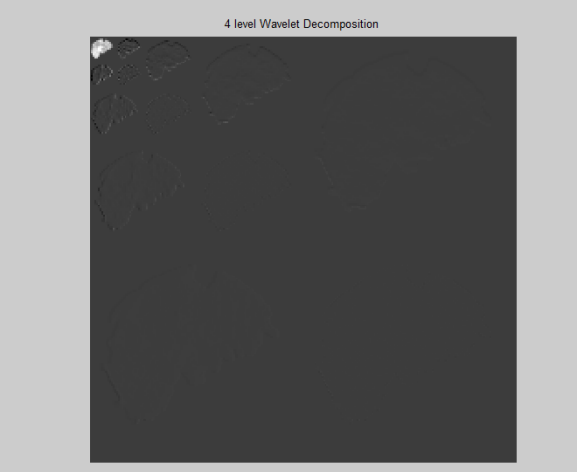

Figure 6. DWT

The (DWT) Discrete Wavelet Transformations Result is shown in the Figure 4. Usually in the numerical analysis and functional analysis, a discrete wavelet transform (DWT) is any wavelet transform for which the wavelets are discretely sampled. As with other wavelet transforms, a key advantage it has over Fourier transforms is temporal resolution: it captures both frequency and location information (location in time).

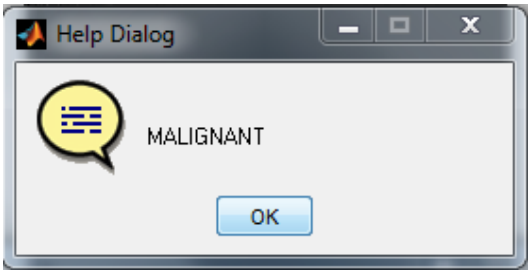

Figure 7. Classification

The Figure 7 Shows the Classification of the input image. This Classification is done by using Probabilistic Neural Network (PNN). It classifies the image in to three parts Normal, Beginant, Malignant stage according to the defected portion. It improves the accuracy of the image.

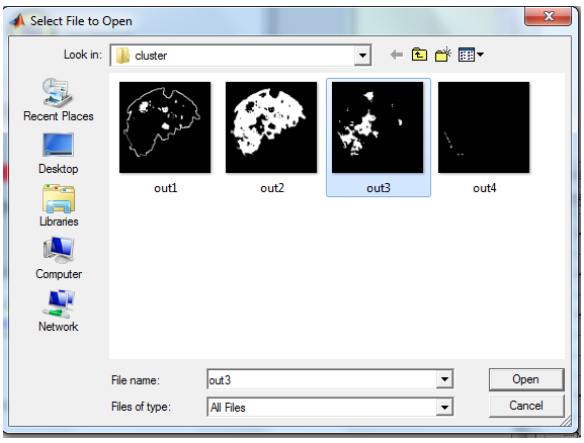

Figure 8. Lloyds Clustering

Lloyd's clustering is used for clustering the image. It is an iterative technique that partition the input image into K number of clusters. The final result can be obtained from this clustering processing and thus the accuracy can be improved.

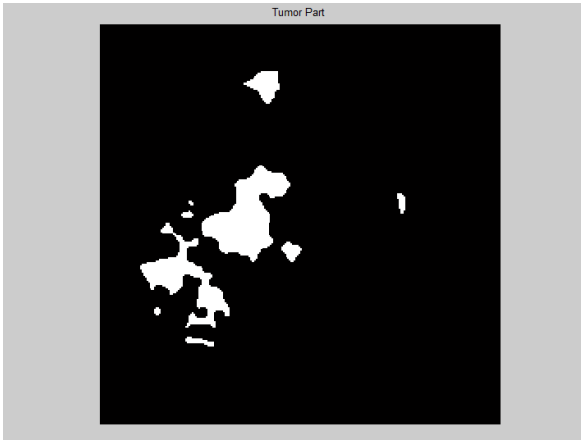

Figure 9. Segmentation

This figure 9 shows the segmented region of the defected lung portion. This is followed by Lloyds clustering and uses morphological operations. It will sharpen the edges of the defected portion and enhances the cancer area.

V. Conclusion and Future Work

Cancer is now the biggest cause of death, due to lung cancer that increasing gradually. This paper focused on detection of lung cancer from CT image of lung by using image processing. Also, it aims to find a better segmentation algorithm which produces a good result. There may be as mass in lung or malignant over the lung. Suppose if it is a mass, then Lloyds algorithm is enough to extract it from the lung cells. If there is any noise present in the CT image, it must be removed before the segmentation process. Reshaping and edge shaping is in the image preprocessing itself. Afterwards, the noise free image is given as an input to the Lloyds cluster, which segment the image and cancer affected area is extracted from the CT image. The proposed method gives more accurate result automatically in short amount of time.

In future work, the classification and performance will be calculated using content based image retrieval method. The discrete wavelet decomposition is used to decompose the image for representing counter edges. The performance will be evaluated in terms of classification accuracy. Decision making will be produced by using haralick technique. The CT lung image will be classified automatically by PNN (Probabilistic Neural Network).

References

1. Aggarwal P., Vig R., and Sardana H.K. (2013), 'Semantic and content based medical image retrieval for lung cancer diagnosis with the inclusion of expert knowledge and proven pathology', Image Information Processing (ICIIP), IEEE Second International Conference on , Vol., no., pp.346,351, 9-11.
2. Daniele Z., Andrew H. and Nickerson J. (2009), 'Nuclear Structure in Cancer Cells', Nature Reviews Cancer, Medical School, Vol. 4, no. 9, pp. 677-87.
3. Dignam J.J., Huang L., Ries L., Reichman M., Mariotto A. and Feuer E. (2009), 'Estimating cancer statistic and other-cause mortality in clinical trial and population-based cancer registry cohorts', Cancer 10.
4. Disha Sharma and Gagandeep Jindal (2011), 'Computer Aided Diagnosis System for Detection of Lung Cancer in CT Scan Images', International Journal of Computer and Electrical Engineering, Vol. 3, No. 5.
5. Gangotri Nathaney (2015), 'Lung Cancer Detection System on Thoracic CT Images Based on ROI Processing', International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 4
6. Georgiadis. Et all (2008), 'Improving Lung Cancer characterization on CT by probabilistic neural networks and non-linear transformation of textural features', Computer Methods and program in biomedicine, Vol 89, pp24-32.
7. Gile N., Wang Ning, Nathalie C.K., Siewe, F., Lin Xudong and Xu De(2008), 'A case study of image retrieval on lung cancer chest Xray pictures', Signal Processing, 9th International Conference on, Vol., no., pp.924, 927, 26-29.
8. Gomathi M. and Thangaraj P. (2012), 'An Effective Classification of Benign And Malignant Nodules Using Support Vector Machine', Journal of Global Research in Computer Science Volume 3, No. 7.
9. JaspinderKaur, NidhiGarg and DaljeetKaur (2014), 'A survey of Lung Cancer Detection Techniques on CT scan Images', International Journal of Scientific & Engineering Research, Volume 5, Issue 6.
10. Kennedy T.C., Miller Y. and Prindiville S. (2005), 'Screening for Lung Cancer Revisited and the Role of Sputum Cytology and Fluorescence Bronchoscopy in a High-Risk Group', Chest Journal, Vol. 10, pp. 72-79.
11. Kumar A., Jinman Kim, Lingfeng Wen and Dagan Feng (2012), 'A Graph-based approach to the retrieval of volumetric PET-CT lung images', Annual International Conference of the IEEE , Vol., no., pp.5408,5411.
12. Sheila A. and Red T. (2010), 'Interphase Cytogenetic of Sputum Cells for the Early Detection of Lung arcinogenesis', Coordinating Center for Clinical Trials, National Cancer Institute, 6120 Executive Boulevard, Bethesda, MD 20852-4910.
13. Specht D.F. (1990), 'Probabilistic Neural Networks', Neural Networks, Vol. 3, No.1, pp. 109-118.
14. Paulus P. and Gaol F.L. (2010), 'Lung Cancer Diseases Diagnostic Asistance Using Gray Color Analysis', Computational Intelligence, Modelling and Simulation (CIMSiM), 2010 Second International Conference on, Vol., no, pp.355, 359, 28-30.
15. Yadav N.G.(2013), 'Detection of lung nodule using Content Based Medical Image Retrieval', International Journal of Electrical, Electronics and Data Communication, ISSN pp. 2320-2084, Volume-1, Issue-2.