Lung cancer detection system using lung CT image processing

Moffy Vas[1], Amita Dessai[2]

1Electronics and Telecommunication Dept., Goa College of Engineering, Farmagudi, Ponda, Goa, India- 403401
E-mail id: moffyvas@yahoo.in[1]
2 Electronics and Telecommunication Dept., Goa College of Engineering, Farmagudi, Ponda, Goa, India- 403401
E-mail id: amitachari@gec.ac.in[2]

***Abstract* - Cancer is the root cause for a large number of deaths worldwide, out of which lung cancer is the cause of the highest mortality rates. Computer tomography scan is employed by radiologists to detect cancer in the body and track its growth. Visual interpretation of database can lead to cancer detection at later stages, thus leading to late treatment of cancer which only boosts up the cancer death rates. Therefore, image processing tools can be used for early detection of cancer. In this paper, a lung cancer detection algorithm is proposed using mathematical morphological operations for segmentation of the lung region of interest, from which Haralick features are extracted and used for classification of cancer by artificial neural networks.**

***Keywords**: Computer tomography, Lung cancer, Haralick features, artificial neural networks.*

I. INTRODUCTION

Lung cancer disease is the second largest death threat to the world after heart attack, as this cancer is responsible for the largest number of deaths, compared to the number of deaths caused by any other cancer type. [1]. Lung cancer is the uncontrolled growth of the cells, thus leading to the formation of lung nodules. It is reported that lung cancer is responsible for around 19% deaths globally mostly due to alcohol and tobacco consumption. The rate of survival is assured by only 15% survival chances, for a survival period of 5 years. [2]. The main cause of such high death rate is the detection in later stages, thus leading to delayed treatment. If lung cancer is detected at an earlier stage, chances of survival can increase up to 50-70%. Non small cell lung cancer and small cell lung cancer are the two major groups into which the lung cancer can be classified based on the cell characteristics. [7] Non small cell lung cancer is the most common type of lung cancer contributing to about 85-90% of total lung cancer cases, while the other 10-15% of the cases is diagnosed with small cell lung cancer. [3] The extent of the spread of cancer is the basis for the division of lung cancer into stages. It comprises of four stages namely stage I-The cancer is confined to the lung, stages II and III-the cancer is confined to the chest (with larger and more invasive tumors classified as stage III) and Stage IV-Cancer has spread from the chest to other parts of the body. There are many techniques to diagnose the lung cancer such as X-rays, Computed Tomography (CT), Magnetic Resonance Imaging (MRI scan), and Sputum Cytology. The problem with these techniques is that it can be time consuming and makes detection possible only at later stages. [2]

II. METHODOLOGY AND IMPLEMENTATION

The methodology adopted in this project was carried out in five steps which are shown with the help of a flow chart in Fig.1. Each step of the flow chart is explained below.

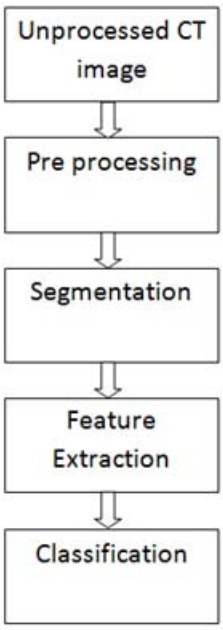


Fig.1. Methodology block diagram

1) Data collection

Images were collected from the V.M.Salgaocar hospital, SMRC and the Manipal hospital both situated in Goa. The CT images of lungs acquired from the hospital database are shown in Fig.2. We will analyze how this algorithm helps us to distinguish between cancerous and non-cancerous images.

978-1-5386-4008-1/17/$31.00 ©2017 IEEE

Fig.2 .a. Non- cancerous image b. Cancerous image

2) Pre-processing

Preprocessing involved the steps shown in Fig.3.

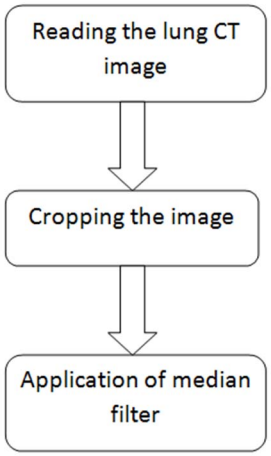
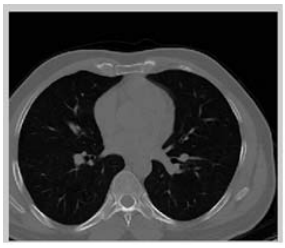


Fig.3. Pre-processing flow diagram

Cropping of the image in first step is done to eliminate the unwanted portions from the image. Next, median filters are applied to the images, which are basically used to get rid of the salt and pepper noise present in the images. A median filter of size 3*3 was used and its contribution towards enhancement of the images is shown in Fig.4.



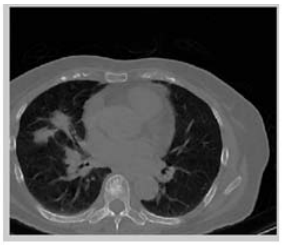


Fig.4. Median filtered images

3) Segmentation

Segmentation steps are depicted in flow diagram, shown in Fig.5. and each step is discussed in detail.

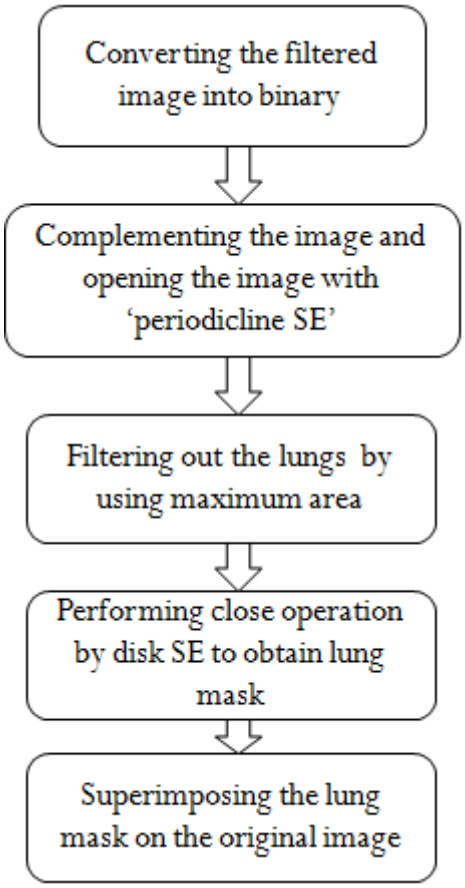


Fig.5. Segmentation block diagram

Converting the images to binary reduces computational complexity and storage issues and also is a pre-requisite for morphological segmentation of lungs. The opening operation using the periodicline structuring element tends to remove some of the foreground pixels from the edges of the region of foreground pixels.

The morphological open operation is expressed as follows

$$A \circ B = (A\theta B) \oplus B \quad (1)$$

Opening is first eroding the image with the structuring element and then dilating the eroded image with the same structuring element as indicated by Eq.1.

The closing operation is given as follows

$$A \bullet B = (A \oplus B)\theta B \quad (2)$$

We perform closing operation wherein, the image is first dilated with the structuring element followed by erosion of the dilated image as given in Eq.2. on the above opened image with a disk structural element of size 15, to obtain the lung masks.

Fig.6. Lung masks

While superimposing the lung mask on the median filtered image, all the areas outside the filtered lung image are multiplied with black areas in the lung mask giving rise to zero intensity value pixels outside the lungs and retaining only the lungs with all its internal features as shown in Fig.7.

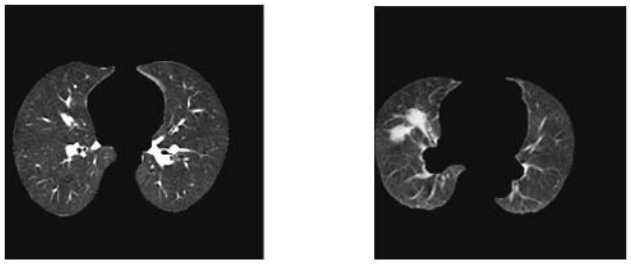


Fig.7.Segmented lungs

4) Feature extraction

The flow diagram for feature extraction is shown in Fig.8.

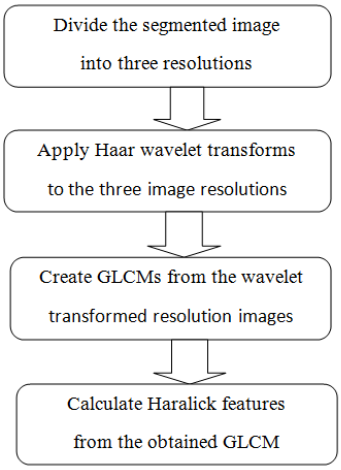


Fig.8: Feature extraction flow diagram

Feature extraction helps in extracting out significant items of data which serve as an input to the classifier. The first step is to resize the image into three different resolutions followed by applying Haar wavelet transforms to these images. The Haar wavelet is the most conventional and basic orthonomal wavelet. The Haar wavelet is memory efficient and exactly reversible without the edge effect characteristics like the other wavelets. This wavelet reflects only changes between adjacent pixel pairs and thus calculates pairwise average and differences. Next, the GLCMs are calculated in different directions four (0^o, 45^o, 90^o, 135^0) and then seven features are extracted from each them. The GLCM function distinguishes the texture of an image by calculating how often pixel pairs with given pixel value and spatial relationship occur. Once the GLCM is calculated, the next step is to extract the second order statistical features or the Haralick features.

The seven Haralick features extracted are as follows, where Ng is the number of gray levels, P is the normalized symmetric GLCM and p(i, j) is the (i, j)th element of the normalized GLCM.

$$\text{Energy} = \sum_i \sum_j p(i,j)^2 \qquad (3)$$

Energy calculates the local uniformity of the gray levels in an image. Higher the similarity in pixels, higher is the energy value.

$$Correlation = \sum_i \sum_j \frac{(i-\mu_x)(j-\mu_y)}{\sigma_x \sigma_y} \qquad (4)$$

Where μ_x, μ_y, σ_x , σ_y are the means and standard deviations of the GLCM.

Correlation is a measure of linear dependency of gray intensity values in the co-occurrence matrix.

1) $$Variance = \sum_i \sum_j (i-u)^2 p(i,j) \qquad (5)$$

Variance feature measures the spread of intensity values of GLCM pixels about the mean. It is similar to entropy.

2) $$IDM = \sum_i \sum_j \frac{1}{1+(i-j)^2} p(i,j) \qquad (6)$$

Inverse difference moment (IDM) gives an account of the local homogeneity in the image. When the local gray level in an image is uniform, IDM is high.

$$Difference\ Entropy = -\sum_{i=0}^{N_g-1} P_{(x-y)}(i) \log(P_{(x-y)}(i)) \qquad (7)$$

$$IMC1 = \sum \frac{HXY-HXY1}{\max\{HX,HY\}} \qquad (8)$$

IMC1 is the information coefficient of correlation I, where

$$HXY = -\sum_i \sum_j p(i,j) \log(p(i,j)) \qquad (9)$$

$$HXY1 = -\sum_i \sum_j p(i,j) \log\{p_x(i)p_y(j)\} \qquad (10)$$

$$HXY1 = -\sum_i \sum_j p_x(i)p_y(j) \log\{p_x(i)p_y(j)\} \qquad (11)$$

$$Contrast = \sum_i \sum_j (i-j)^2 P(i,j) \qquad (12)$$

Contrast indicates the intensity variations between the pixel under consideration and its neighboring pixel. Larger contrast means larger variation.

The total number of features from each image was 252 (7 features * 4 directions * 3 haar approximations (horizontal, vertical and diagonal)* 3 different resolutions of image = 252 features)

The feature plot is shown in Fig.9. The seven extracted features which are numbered 1 to 7 are energy, correlation, variance, homogeneity, difference entropy, information measure of correlation I and contrast respectively.

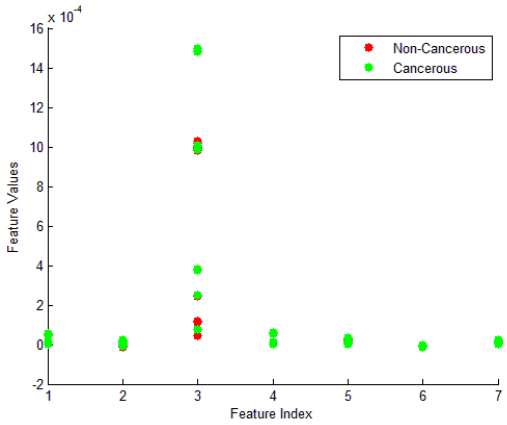


Fig.9. Features plot

5) Classification

Artificial neural networks are reckoning systems made up of numerous simple and highly interconnected processing elements, which process information by their dynamic state response to external inputs. In this paper, feed forward neural network with back propagation algorithm was used. The back propagation looks for the least of the error function in the weight space using the method of gradient descent. The weights are altered such that, the error function has the minimum value. The algorithm has 252 input nodes, 20 hidden nodes and a couple of output nodes. The block diagram of the feed forward network is shown in Fig.10

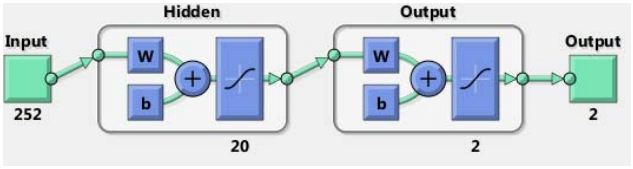


Fig.10. Feed forward ANN schematic diagram

III. RESULTS

A total number of 216 images were used, out of which 128 images were used for training and 88 images were used for testing. The confusion matrix for the ANN classification is given in Fig.11. As can be seen in Fig.11 seven images are misclassified.

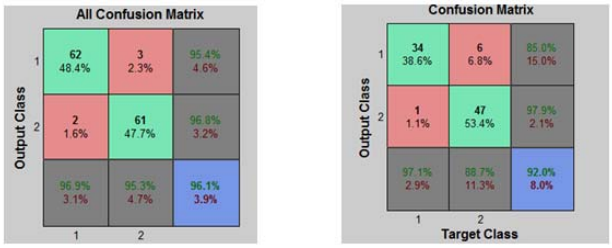


Fig.11. Confusion matrix for training and testing

The training accuracy was 96% and testing accuracy was 92%. The sensitivity was 88.7% and specificity was calculated to be 97.1%.

Reasons for misclassifications

As can be seen from Fig.12 the lungs in the non-cancerous image are separated by the white region, which is the aortic region.

The problem faced by the image shown in Fig.13.is improper segmentation of the lungs due to the presence of this white region in between them which is depicted in Fig.12

Fig.12. Non-cancerous image

Fig.13. Segmentation issue

The segmentation issue is that, the white region is also segmented as the region of interest leading image misclassification as cancerous.

The second reason for misclassification of images is that the cancer nodules lie on the border of the lungs as shown in Fig.14.

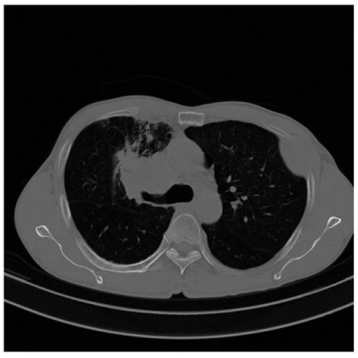


Fig.14. Cancerous image

The morphological closing operation with a disk structuring element of size 15 results in the loss of cancer nodules.

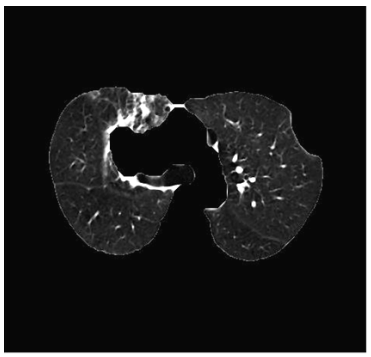


Fig.15.Segmentation issue

The loss of the cancer nodules as shown in Fig.15 leads to the misclassification of the cancerous image as a non-cancerous one. The size of structuring element creates an issue as the lungs vary in size and a fixed size structuring element cannot be used to segment all the images.

IV. CONCLUSION

The methodology adopted in this project aims to develop an automated system for lung cancer detection. Application of median Filter to eliminate impulse noise in the images proved to be a success. The morphological operations also contributed towards satisfactory results in the process of segmentation. Artificial neural networks proved to be a good classifier with acceptable accuracy. The methodology adopted in this project resulted in an accuracy of 92% for the hospital database. This system aims at increasing the accuracy and speed of the lung cancer detection system. It also helps in detecting the cancer at earlier stages.

V. FUTURE SCOPE

The accuracy of the cancer detection system can be improved by using a different segmentation technique like p-tile thresholding and watershed segmentation followed by binary morphology. Using different feature set like curvelet transformation features together with morphological features other than Haralick features, may have a positive impact on the accuracy of the system.

VI. ACKNOWLEDGEMENT

I am deeply indebted to my guide Prof. Amita Dessai for allowing me to carry out the project under her supervision. She has given me the confidence to take up this project and is my pillar of support that I have banked on in times of difficulty. She has always guided me and motivated me, when this project gave me unsatisfactory results at times. I sincerely appreciate the encouragement extended to me by our HOD, Dr. H.G.Virani who was always on his toes when his signatures were required for getting permission to access the hospital database. I also express gratitude to our Principal, Dr.V.N.Shet who always extended his help whenever required. I also thank SMRC hospital and Manipal hospital, Goa for providing me with the lung CT images. I also need to thank God, my parents, family members, and all my well wishers without whom this work would not have taken shape.

REFERENCES

[1] K.Punithavathy, M.M.Ramya, Sumathi Poobal, "Analysis of statistical texture features for automatic lung cancer detection in PET/CT images", International Conference on Robotics, Automation, Control and Embedded systems(RACE),IEEE ,18-20 February 2015.

[2] Badrul Alam Mia, Mohammad Abu Yusuf , "Detection of lung cancer from CT image using image processing and neural network", International conference on Electrical Engineering and Information Communication Technology (ICEEICT) ,IEEE ,May, 2015.

[3] Anita Chaudhary, Sonit Sukhraj Singh "Lung cancer detection on CT images using image processing", computing sciences 2012 international conference, IEEE, 2012.

[4] Nooshin Hadavi, Md Jan Nordin, Ali Shojaeipour , "Lung cancer diagnosis using CT-scan images based on cellular learning automata", International conference on Computer and Information Sciences(ICCOINS), IEEE , 2014.

[5] Muhammed Anshad, S.S Kumar , Recent methods for the detection of tumor using computer aided diagnosis", International Conference on Control, Instrumentation, Communication and Computational technologies(ICCICCT), IEEE transactions,2014.

[6] Mohsen Keshani, Zohreh Azimifar and Reza Boostani, "Lung nodule segmentation using activecontour modeling" MVIP, IEEE, 2016

[7] Gawade Prathamesh Pratap, R.P Chauhan, "Detection of lung cancer cells using image processing techniques ", 1st IEEE International Conference on Power Electronics, Intelligent control and energy systems, IEEE, 2016

[8] Robert M Haralick, K.Shanmugam, Itshak Dinstein, "Textural Features for Image Classification" IEEE Transactions on Systems, Man and Cybernetics, 3(6), pp. 610 - 621, 1973.