Journal of Engineering Sciences
Vol 12, Issue 06, JUNE /2021
ISSN NO:0377-9254

LUNG CANCER DETECTION USING IMAGE PROCESSING TECHNIQUES

[1] Mrs. S. PALLAVI, Assistant Professor, Department of ECE, pallavi.s@sreyas.ac.in
[2] B. SREEJA, B.Tech, Department of ECE, sreejayadav22@gmail.com
[3] D. AKANKSHA, B.Tech , Department of ECE, devarapalliakanksha20@gmail.com
[4] K. JESHWANTH YADAV, B.Tech, Department of ECE, keesarajeshwanth@gmail.com
[5] B. LAXMAN GOUD, B.Tech, Department of ECE, lakshmanbandarapu06@gmail.com
Sreyas Institute of Engineering and Technology, Hyderabad, Telangana.

ABSTRACT

From last decade, lung cancer becomes sign of fear among the people all over the world. As a result, many countries generate funds and give invitation to many scholars to overcome on this disease. Many researchers proposed many solutions and challenges of different phases of computer aided system to detect the lung cancer in early stages and give the facts about the lung cancer. Image processing plays vital role to prevent lung cancer. Since image processing is necessary for computer vision, further in medical image processing there are many technical steps which are necessary to improve the performance of medical diagnostic machines. Without such steps programmer is unable to achieve accuracy given by another author using specific algorithm or technique. In this paper we highlight such steps which are used by many author in pre-processing, segmentation and classification methods of lung cancer area detection. If pre-processing and segmentation process have some ambiguity than ultimately it effects on classification process. We discuss such factors briefly so that new researchers can easily understand the situation to work further in which direction.

Keywords: Cancer Detection, Image processing, Feature extraction, Thresholding, Segmentation.

I. INTRODUCTION

Lung cancer is a disease of abnormal cells multiplying and growing into a cancer. Cancer cells can be carried away from the lungs in blood, or lymph fluid that surrounds lung tissue. Lymph flows through lymphatic vessels, which drain into lymph nodes located in the lungs and in the centre of the chest. Lung cancer often spreads toward the centre of the chest because the natural flow of lymph out of the lungs is toward the centre of the chest. Metastasis occurs when a cancer cell leaves the site where it began and moves into a lymph node or to another part of the body through the blood stream [1]. Cancer that starts in the lung is called primary lung cancer. There are several different types of lung cancer, and these are divided into two main groups: Small cell lung cancer and non-small cell lung cancer which has three subtypes: Carcinoma, Adenocarcinoma and Squamous cell carcinomas.

LUNG CANCER TYPES AND STAGES

Lung cancers are broadly classified into two main categories. They're Small Cell Lung Cancer (SCLC) and Non-Small Cell Lung Cancer (NSCLC).NSCLC reports 85-90% of lung cancers, while SCLC accounts for the remaining 10-15%. While the growth and spread of NSCLC is slow, SCLC is a fast-growing cancer and spreads rapidly to other body parts. Smoking is the main cause in all cases of SCLC, NSCLC will be treated with surgery, chemotherapy, radiotherapy, depending on the stages of cancer is diagnosed. SCLC cancer is mostly treated with chemotherapy. Difference between SCLC and NSCLC are listed in Table 1. There are 4 different types of NSCLC, each having different treatment options:

- **Epidermoid / Squamous Cell Carcinoma**

The cancer forms in the lining of the bronchial tubes and is mostcommon in men.

- **Adenocarcinoma**

The cancer forms in the mucus glands of the lungs and is most commonin women and non-smokers.

- **Bronchioalveolar Carcinoma**

The cancer forms near the air sacs of the lungs and is a rare type ofadenocarcinoma.

- **Large-Cell Undifferentiated Carcinoma**

The cancer form near the surface of the lungs or on its outer edges andspreads tremendously.

- **Non-Small Cell Lung Cancer stages are**

Stage 1: Cancer is not in any lymph nodes and is located only in lungs.

Stage 2: Cancer located in the lungs will develop in size and spread nearby lymphnodes.

Stage 3: Cancer cells in lungs will spread in lymph nodes and middle chest described as advances in stage that can be classified into two types.

Stage 3A represents that cancer caused cells will spread in lymph nodespresent on same side of chest where it originated.

Stage 3B represents that cancer cells will spread above the collar boneor to lymph nodes on the other chest.

Stage 4: This is an advanced stage of lung cancer, and the disease is in its advanced form. In this stage, the cancer would've spread to both lungs, the fluid area around the lungs and/or to other body parts like liver and other organs.

Literature Review

Thresholding

Based on the fact that healthy lung tissues form darker region compared to other organs like heart and liver, thresholding technique can be used to separate lung tissues from other parts. Selecting the optimum threshold value is the crucial step in thresholding based methods since the intensity values are almost similar for tissues, vessels and lung lobes.

Hu et al[3] proposed an iterative thresholding algorithm for lung segmentation followed by opening and closing operations using morphological operations for refining the segmentation output. Yim et al[4] used region growing method with connected component analysis for extracting lung fields. Pu et al[5] used an initial threshold for segmenting the lung region and for improving the segmentation process a border marching algorithm is used. Gao et el[6] used a threshold based method for separating the pulmonary vessels, lung airways, left and right lungs separately.

Jibi et al.[7] used multiple levels of thresholding to segment the lung from otherparts of the CT image. In their studies they included irregular lung walls. Their work well suited for CT images with solitary nodules.

The major drawback of this thresholding method is that the accuracy depends on the quality of the image which is determined by the acquisition method and scanner type. Also the densities of the different parts are similar and hard to differentiate. Table 2 shows the summary of various studies of lung segmentation using thresholding method.

S. No	Study of	No.of sample images	Method used	Performance Results
1.	Hu et al [3]	24 datasets	Iterative thresholding and morphological operations	RmsD=0.54 (0.8 pixel)
2.	Yim et al. [4]	10 subjects	Region growing, connected component	RmsD = 1.2 pixel
3.	Pu et al. [5]	20 datasets	Thresholding	FP/GT=0.43% FN/GT=1.63%
4.	Gao et al. [6]	8 subjects	Thresholding	DSC = 0.9946
5.	Jibi et al. [7]	60 slices	Multilevel thresholding	-

Table 1 Lung segmentation using thresholding method

The results presented above refer to the lung segmentation where the nodules are not connected to the lung walls.

Deformable Boundaries

The next class of lung segmentation method includes the use of deformable boundary models including snakes, active contours and level sets. Deformable boundary models starts from one initial point and follows the shape based on internal or external assisting factors to fix the shape to any of the objects. Itai et al [8] used the deformable models to fit the lung boundaries. Silveria et al. [9] proposed the Level Set active contour model to separate the left and right lungs. Themajor drawback of this method is that the initial point selection is highly sensitive and failure to adapt the boundaries because of the in homogenities in the lung region.Table 2 summarizes the methods.

The results obtained using deformable boundaries are only qualitative notquantitative.

Table 2 Lung segmentation using Deformable boundaries method

S. No	Study of	No.of sample Images	Method used	Performance Results
1	. Itai et al [8]	-	Snake deformable Model	Qualitative assessment
2.	Silveria et al. [9]	1 slice	Level set Active Contour	Qualitative assessment

II. SYSTEM MODEL

Feature Extraction

The following features are extracted from the segmented nodule: It shows the actual number of pixels in the Region of Interest (ROI).

$$A=\sum_{i,} (i, j) \quad (1)$$

Major and minor axis of ROI:

The equivalent elliptical major axis of irregular region is termed as the largest diameter among any pair of pixels in the boundary region. It is termed as ‘a’ in Figure 1. The equivalent elliptical minor axis of irregular region is defined asthe shortest diameter among any pair of pixels in the boundary region. It is denoted as ‘b’ in Figure 1.

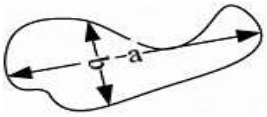

Figure 1: Major axis & Minor axis of a region

Perimeter

It is defined as the distance between any adjacent pair of pixels around theboundary of the region. This will give proper value if the image boundary is continuous. Perimeter represents the count of pixels present in the edge of theregion.

Eccentricity

The eccentricity is defined as the proportion of minor axis and major axisof the ellipse.

Eccentricity = Minor Axis Length / Major Axis Length **(2)**

Equivalent Diameter

It is defined as the largest distance between two points of the ROI.

Steps in the algorithm are as follows

Algorithm 1: segmentation algorithm

1) Input image is considered as a graph with a set of vertices and edges where vertices represent the pixels in the image and edges represent the relationship between pixels
2) Read two user defined seed points, one to mark the nodule and other to represent the lung region.
3) With 2 pre-defined pixels as labels, compute two probabilities P_n andP_l for each pixel the probability of walking to reach the nodule and the lung seeded pixels using (3.2).
4) A couple of probabilities for each pixel is defined with two values representing one for the nodule region and the other for the lung region.
5) From the vector of probabilities, select the pixels with maximum probability ($p_n > p_l$) for nodule pixel and label that pixel as a nodule.
6) Segment the nodule labeled pixels as separate region representing the nodule.
7) From the segmented nodule, features including major axis, area,minor axis, diameter and accentricity are obtained.

Similarity Measure

The segmentation precision is calculated with Dice Similarity Coefficient. DSC is a statistical confirmation measure to evaluate the dissimilarity among physical segmentation and automatic segmentation of images. The Dice Similarity Coefficient is calculated as

$$DSC= 2 (MnA)/(M+A) \quad (3)$$

With M as segmenting the nodule manually and A represents nodule segmentation using random walk . Images with high irregularity have the greatest of DSC values using random walk is 0.92.

The following are the rewards of proposed method

- Nodules with irregular shapes are segmented precisely.
- Free parameter is replaced by a constant β.
- With small amount of seed points accuracy is improved.

III. PROPOSED SCHEME

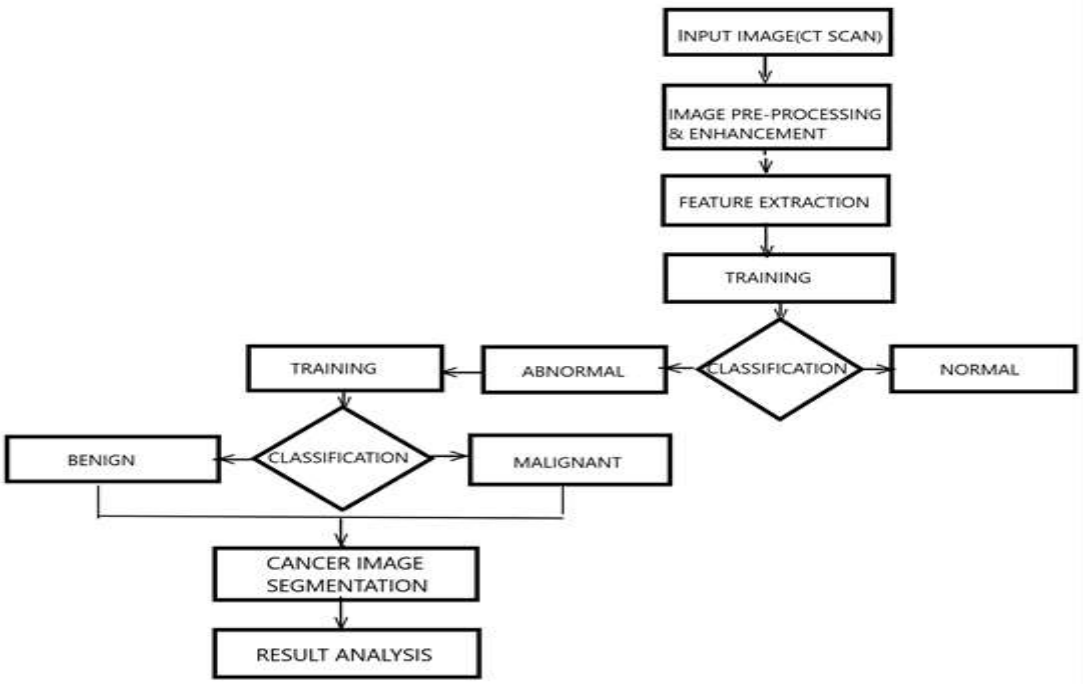

Fig 2: FLOW CHART FOR THE PROPOSED LUNG CANCER DETECTION AND CLASSIFICATION

NEURAL NETWORKS

A neural network is a network or circuit of Neurons or in a modern sense, an Convolutional Neural Network, composed of Neurons or nodes. Thus a neural network is either a Biological Neural Network, made up of real

biological neurons, or an artificial neural network, for solving Artificial Intelligence (AI) problems. The connections of the biological neuron are modeled as weights. A positive weight reflects an excitatory connection, while negative values mean inhibitory connections. All inputs are modified by a weight and summed. This activity is referred as a linear combination. Finally, an activation function controls the Amplitude of the output. For example, an acceptable range of output is usually between 0 and 1, or it could be −1 and 1.

These artificial networks may be used for Predictive Modeling, adaptive control and applications where they can be trained via a dataset. Self-learning resulting from experience can occur within networks, which can derive conclusions from a complex and seemingly unrelated set of information.

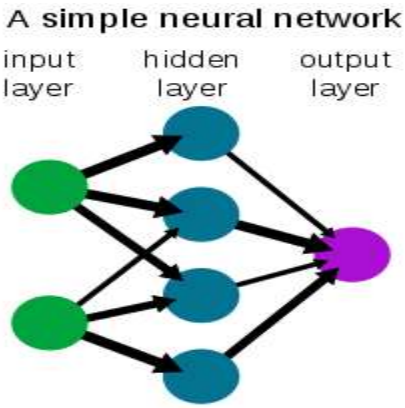

Figure No: 3. The idea of how neural networks work:

Recently there has been a great buzz around the words "neural network" in the field of computer science and it has attracted a great deal of attention from many people. Essentially, neural networks are composed of layers of computational units called neurons, with connections in different layers. These networks transform data until they can classify it as an output. Each neuron multiplies an initial value by some weight, sums results with other values coming into the same neuron, adjusts the resulting number by the neuron's bias, and then normalizes the output with an activation function.

How CNNs Work

A convolutional neural network can have tens or hundreds of layers that each learn to detect different features of an image. Filters are applied to each training image at different resolutions, and the output of each convolved image is used as the input to the next layer. The filters can start as very simple features, such as brightness and edges, and increase in complexity to features that uniquely define the object.

CNNs perform feature identification and classification of images, text, sound, and video.

Feature Learning, Layers, and Classification

Like other neural networks, a CNN is composed of an input layer, an output layer, and many hidden layers in between.

These layers perform operations that alter the data with the intent of learning features specific to the data. Three of the most common layers are: convolution, activation or ReLU, and pooling.

Convolution puts the input images through a set of convolutional filters, each of which activates certain features from the images.

Rectified linear unit (ReLU) allows for faster and more effective training by mapping negative values to zero and maintaining positive values. This is sometimes referred to as *activation*, because only the activated features are carried forward into the next layer.

Pooling simplifies the output by performing nonlinear down sampling, reducing the number of parameters that the network needs to learn.

These operations are repeated over tens or hundreds of layers, with each layer learning to identify different features.

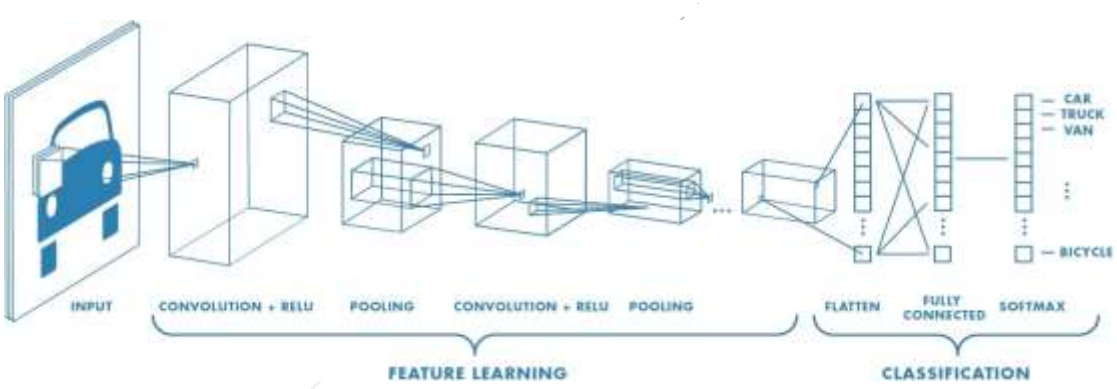

Figure 4. Convolutional Neural Network

Example of a network with many convolutional layers. Filters are applied to each training image at different resolutions, and the output of each convolved image is used as the input to the next layer.

Classification Layers

After learning features in many layers, the architecture of a CNN shifts to classification .The next-to-last layer is a fully connected layer that outputs a vector of K dimensions where K is the number of classes that the network will be able to predict. This vector contains the probabilities for each class of any image being classified.

The final layer of the CNN architecture uses a classification layer such as softmax to provide the classification output.

CONVOLUTIONALNEURALNETWORK

An image convolution is a transformation pixel by pixel, done by applying to an image some transformation defined by a set of weights, also known as a • lter. Let s be a set of source pixels, and w a set of weights, a pixel y is transformed as follows.

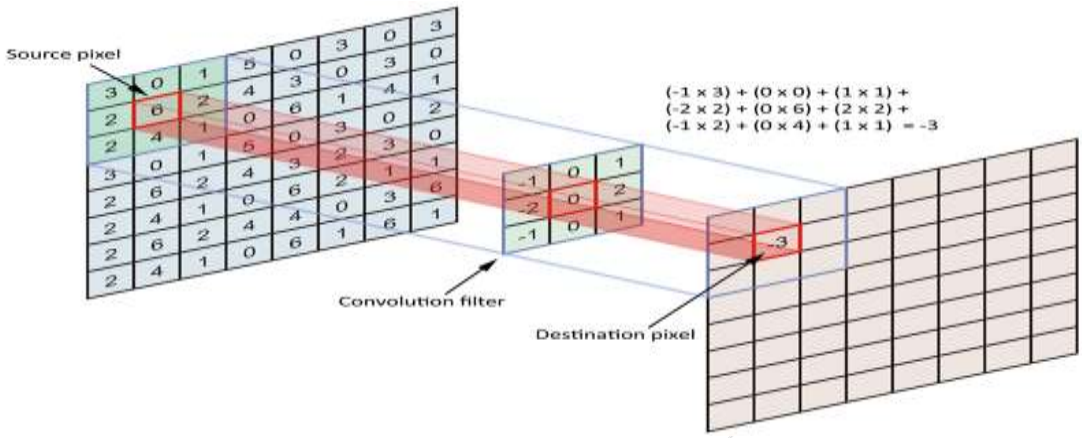

Figure 5. Architecture of CNN

GLCM FEATURE EXTRACTION

GLCM is a statistical method which was modified by Hasan and Meziane , and was used to extract the second order texture features by inspecting the combined frequencies of all grey levels of pixel configuration of each pixel in the left hemisphere (reference pixel) with one of nine opposite pixels that exist in the right hemisphere. These features measure statistically the degree of symmetry between both sides of the lung. Symmetry is an important parameter that is used within the diagnosing process to detect the normality and abnormality of the human lung. Consequently, nine co-occurrence matrices are extracted for each MRI slice under nine offsets D (45,45), (0,45), (315,45), (45,0), (0,0), (315,0), (45,315),(0,315), (315,315), and one distance as shown in Fig. 3.The co-occurrence relative frequencies between joint pixels are calculated after normalization by the total sum of all its elements, equation

$$P(i,j)_{(\theta_1,\beta_2)} = \frac{1}{256^2}\sum_{x=1}^{M}\sum_{y=1}^{N} \times \begin{cases} 1, & \text{if } L(x,y) = i \\ & \text{and } R(x+\Delta x, y+\Delta y) = j \quad (1) \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where L and R are the left and right parts of the lung's hemispheres respectively, M and N are the width and height of MRI slice respectively, i and j are the co-occurrence matrix's coordinates, 1x and 1y values are subject to the directions of measured matrix and undergo to a set of rules that are demonstrated clearly in, and P is the resulting

concurrence matrix. There are twenty-one texture measures extracted from each co-occurrence matrix and these measures represent the most common and widely-used texture features. Hasan and Meziane rained these texture measures by ignoring the irrelevant features using analysis of variance method (ANOVA) and reduced to eleven texture measures for each co-occurrence matrix, namely, the contrast, the dissimilarity, the correlation, the sum of square variance, the sum variance, FIGURE 5. Convolution of a image with a kernel the sum average, the difference entropy, the inverse difference normalized (IDN), the information measure of correlation I(IMC1), the inverse difference moment normalized (IDMN)and the weighted distance in addition to the cross correlation. The total number of texture measures was reduced from190 to 100 feature measures after using ANOVA.

Table 3 Sample Texture Based Features for a benign and malignant nodule

Measures	Benign	Malignant
Autocorrelation	1.065398	2.196193
Contrast	0.51041	0.666249
Correlation	0.897825	0.964503
Energy	0.995765	0.958517
Entropy	3.019124	4.123949
Homogeneity	0.999397	0.597462
Dissimilarity	0.00257	0.102755
Cluster_promin	20.87711	652.8795
cluster_shade	1.984402	48.23242
sumOfSq	1.039291	2.193984
sumVar	4.197652	8.315266
diffVariance	0.01041	0.066249
infmeas_corr	0.15043	0.409441
sumAvg	2.019566	2.27698
sumEnt	0.018492	0.120853
diffEntropy	0.008238	0.030222
invdiffN	0.999783	0.999005
invdiffMN	0.999878	0.999333

Entropy, contrast, energy, homogeneity, and correlation are the major attributes that prominently discriminate the benign and malignant nodules. According to Qian Zhao et al [96] the values of entropy, contrast, energy, homogeneity, and correlation were approximately 3.5597 ± 0.6470, 0.5384 ± 0.2561, 0.9521 ± 0.1256,

0.8281±0.0604, and 0.8748 ± 0.0740 in the benign nodules and 3.8007±0.6235, 0.9988±0.2961, 0.1673±0.1070, 0.7980±0.0555, and 0.8550±0.0869 in the malignant nodules where the size of the nodule is less than 20mm.

Support Vector Machine (SVM)

Hiram Madero Orozco [91] stated that Support Vector Machine (SVM) isa commonly used classification method for medical images [92]. SVM was proposed by Vapnik [93]. For investigating the samples used for classification supervised learning algorithms like SVM is used. . Hiram Madero Orozco et al[91] delivered an professional method for nodule classification without segmenting the nodules. 8 texture features are extorted from the histogram and the GLCM matrix (using four different angles) is constructed from the image. Support vector machine (SVM) is applied to group the lung tissues into two classes: with cancer and without cancer. They obtained the good consistency outputs with 90◦ and 135◦ of the GLCM in their work.

The basic SVM is a non-probabilistic binary classifier that reads a sample of data and for each given sample it forecasts one of the two output classes. With the given set of examples, each belong to one of the output categories, an SVMtraining algorithm constructs a model that categorizes the given input to one of the output classes.

IV. NUMERICAL RESULTS

The input to this subsystem is a DICOM image of a chest CT scan of size256 × 256 pixels. The major cause of noise in CT image is random noise, and then Gaussian noise is also present in CT images. Median filtering is applied in the proposed method. The objective of median filtering is to remove the noise that has distorted image. It is based on a statistical approach. Median filtering is a nonlinear procedure often applied in image processing to reduce “salt and pepper” noise. A median filter is better than convolution after the goal is to concurrently remove noiseand retain boundaries.

Segmentation of lungs can be done with thresholding, deformable boundary models, Region Growing and various atlas based and edge based techniques. The proposed method used thresholding and morphological operationsto extract the lung area. With the use of morphological procedures the lung region is extracted. The grayscale image is initially transformed to binary image. Entirely the pixels in the input image with an intensity larger than a threshold level is substituted with value ‘1’ and all pixel values with an intensity fewer than threshold level is replaced with value ‘0’. The threshold level is computed by choosing the value to minimize the intra-class variance of the black and white pixels.

Confusion Matrix

The results of any classification can be summarized with the help of a confusion matrix. A confusion matrix shows the details of the actual and expected results given by the classifier. Based on the values present in the

confusion matrix the performance of the classifier is evaluated [22]. Table 4 shows the confusionmatrix for a binary classifier.

Table 4: confusion matrix

Known Label	Predicted Label	
	Positive	**Negative**
Positive	True Positive (TP)	False Negative (FN)
Negative	False Negative (FN)	True Negative (TN)

Given a set of known labels for positive and negative classes, any input data may be classified into positive or negative class. According to the results the classification may be grouped as True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN).

We calculated the Accuracy, Precision, Recall, and F1-Score of proposed CNN and other pre-trained models. Accuracy is the measurement of actual true classifications.

$$\textbf{Accuracy} = \textbf{TruePositive(TP)} + \textbf{TrueNegative(TN)} / \textbf{(Total No of Samples)}$$

Precision estimates how many positive labels we had predicted.

$$\textbf{Precision} = \textbf{TruePositive(TP)} / \textbf{(TruePositive(TP)} + \textbf{FalsePositive(FP))}$$

Recall evaluates how many positive labels we had correctly predicted from our data.

$$\textbf{Recall} = \textbf{TruePositive(TP)} / \textbf{ (TruePositive(TP)} + \textbf{FalseNegative(FN))}$$

Journal of Engineering Sciences
Vol 12, Issue 06, JUNE /2021
ISSN NO:0377-9254

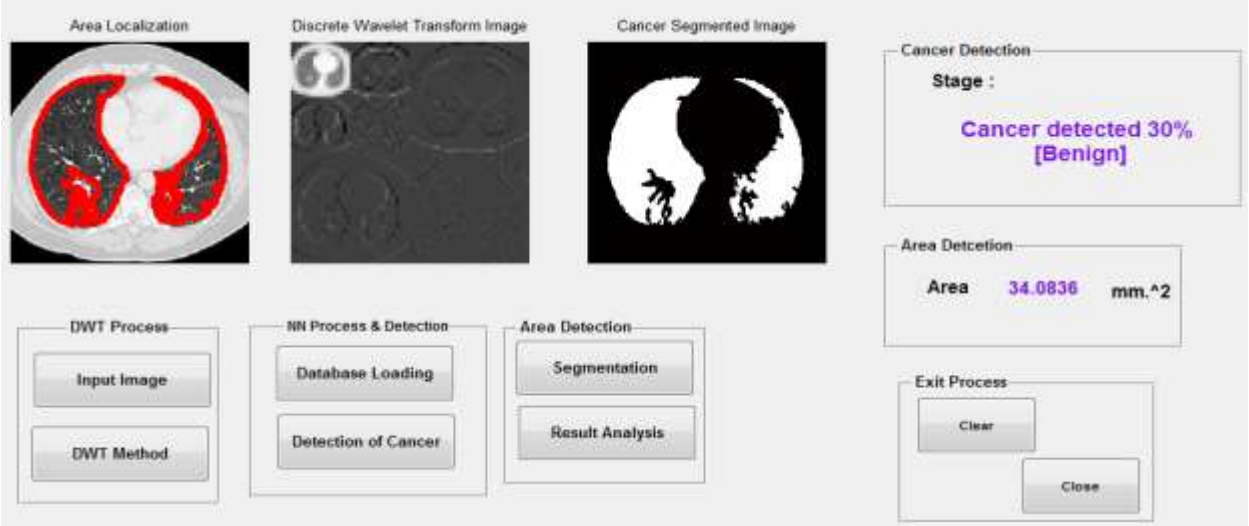

Fig: 6. Benign cancer analysis for input image

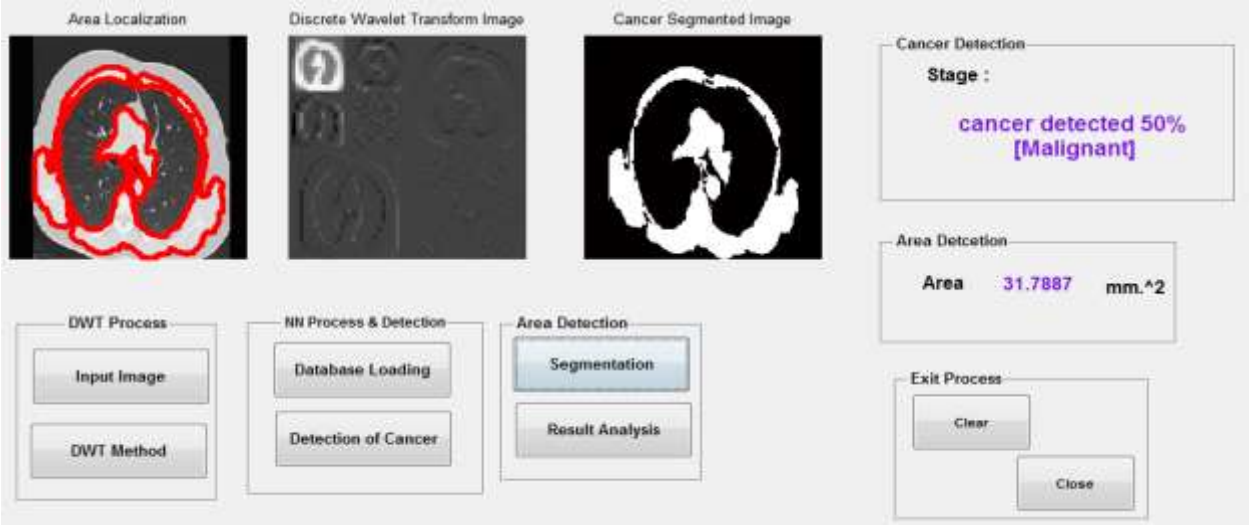

Fig 7: Malignant cancer analysis for input images

CONCLUSION

In this paper, a new approach was presented to classify lung cancer. First, using the pre-processing techniques we improve the quality of the image , we find the region of interest in CT images and enhance that region and apply feature extraction . Second, we provide an efficient methodology for lung cancer classification by proposing a simple CNN network. For sophisticated and accurate results neural network requires a large amount of data to train on, but our experimental result shows that even on such a small dataset, we can attain full accuracy and our accuracy rate is very fine. So, our model needs less computational specifications as it takes less execution time.

Our proposed system can play a prognostic significance in the detection of cancers in lung cancer patients. To further boost the model efficiency, comprehensive hyper-parameter tuning and a better pre-processing technique can be conceived. Our proposed system is for binary classification problems, however, in future work, the proposed method can be extended for categorical classification problems such as identification of lung cancer types such as Glioma, Meningioma, and Pituitary or may be used to detect other lung abnormalities. Also, our proposed system can play an effective role in the early diagnosis of dangerous disease in other clinical domains related to medical imaging, particularly lung cancer and breast cancer whose mortality rate is very high globally. We can further develop this approach in other scientific areas as well where there is a problem in the availability of large data or we can use the different transfer learning methods with the same proposed technique.

REFERENCES

1.Non-Small Cell Lung Cancer, Available at: http://www.katemacintyrefoundation.org/ pdf/non-small-cell.pdf, Adapted from National Cancer Institute (NCI) and Patients Living with Cancer (PLWC), 2007, (accessed July 2011).

2.Tarawneh M., Nimri O., Arqoub K., Zaghal M., Cancer Incidence in Jordan 2008, Available at: http://www.moh.gov.jo/MOH/Files/Publication/Jordan%20Cancer%2 0Registry_2008%20Report_1.pdf, 2008, (accessed July 2011).

3.Lung Cancer Database, Available at: https://eddie.via.cornell.edu/cgi-bin/datac/signon.cgi, (accessed July 2011).

4.Gonzalez R.C., Woods R.E., Digital Image Processing, Upper Saddle River, NJ Prentice Hall, 2008.

5.Cristobal G., Navarro. R., Space and frequency varient image enhancment based in Gabor representation, Pattern Recognition Letters, Elsevier, 1994, 15, p. 273-277.

6.Krishan A., Evaluation of Gabor filter parameters for image enhancement and segmentation, in Electronic Instrumentation and Control Engineering, Master. Punjab: Thapar University, 2009, p. 126.

7.Nunes É.D.O., Pérez M.G., Medical Image Segmentation by Multilevel Thresholding Based on Histogram Difference, presented at 17th International Conference on Systems, Signals and Image Processing, 2010.

8.Venkateshwarlu K., Image Enhancement using Fuzzy Inference System, in Computer Science & Engineering, Master thesis, 2010.

9.Shapiro L.G., Stockman G.C., Computer Vision: Theory and Applications, Prentice Hall, 2001.