Lung Cancer Nodule Detection

Methodology

The whole system of lung cancer detection divided into following steps: Image Acquisition, Image Preprocessing, Segmentation, Neural Network for Healthy/Infectious Lungs followed by Image Preprocessing with Discrete Wavelet Transform and Deep Neural Network for Identification of Cancer Nodules.

A. Image Acquisition

Normally a special type of digital X-Ray machine is used to acquire detailed pictures or scans of areas inside the body called computerized tomography (CT). Computed tomography is an imaging procedure. The system has collected total **891** Lung CT images that are cancer (514) and normal image (377) of lung from kaggle.com website. The system used Lung CT images that are dcm file format. The Dataset is divided into 'healthy' and 'unhealthy' lungs.

B. Pre-Processing of Images

After Image Acquisition, images are passed through the image preprocessing steps. Fig. 1 shows the block diagram of image preprocessing steps.

i. Normalization

Normalize the acquired image by using the Matlab function imresize. The system uses imresize function with the value of 512 x 512 pixels. This size gives enough information of the image when the processing time is low.

ii. Gray Scale Conversion

RGB image converted into gray scale image by using the Matlab function rgb2gray. It converts RGB image or color image to grayscale by eliminating the hue and saturation information while retaining the luminance.

iii. Noise Reduction

To remove the noise the system used median filter i.e. medfilt2. Medfilt2 is 2-D median filter. Median filtering is a nonlinear operation often used in image processing to reduce "salt and pepper" noise. A median filter is more effective than convolution when the goal is to simultaneously reduce noise and preserve edges.

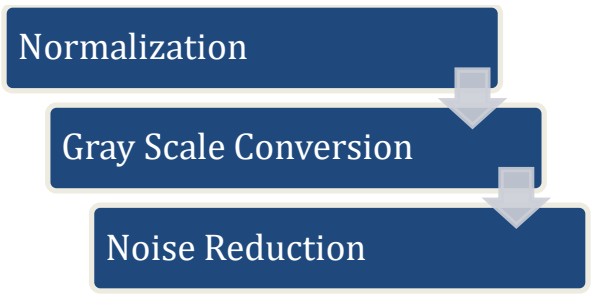


Figure 1: Block Diagram of Image Pre-processing

C. Segmentation

Image Segmentation in computer vision system, is the process of partitioning a digital image into multiple segments. The goal of segmentation is to simplify and/or change the representation of an image into more meaningful and easier to analyze. Image segmentation is typically used to locate objects and boundaries (lines, curves, etc.) in images. More precisely, image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain characteristics. In the proposed system, segmentation processes consists of region growing segmentation as shown in figure 2.

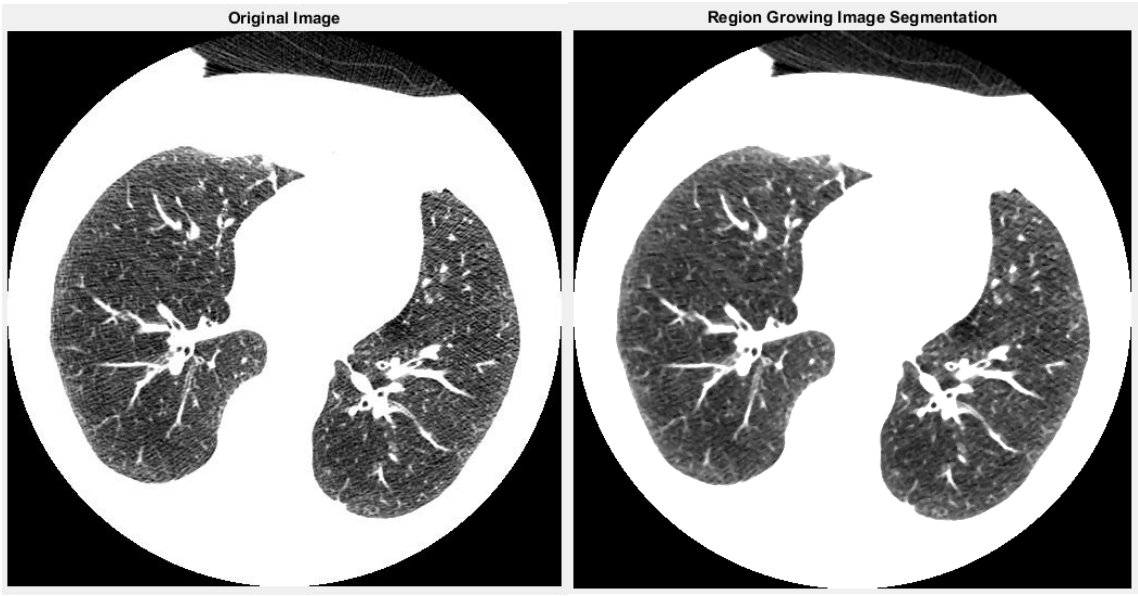


Figure 2: *(a) Original Image* *(b) Segmented Image*

D. Neural Network Detection

After the Segmentation method, first part of the Lung Cancer Detection System uses neural network which is very efficient and reliable. After the segmentation process, these features are passed through the neural network to train up the system for classification purpose or detection purpose. The whole proposed training system of lung cancer detection consist of the following steps- Image Acquisition, Image Preprocessing, Segmentation, Neural Network Classification for the first part of the Lung Cancer Detection.

A neural network is employed for lung cancer detection. A multilayer feed forward neural network with supervised learning method is more reliable and efficient for this purpose. Neural Network design of the proposed system is shown in Fig. 3 which had **17 hidden layers** for the classification.

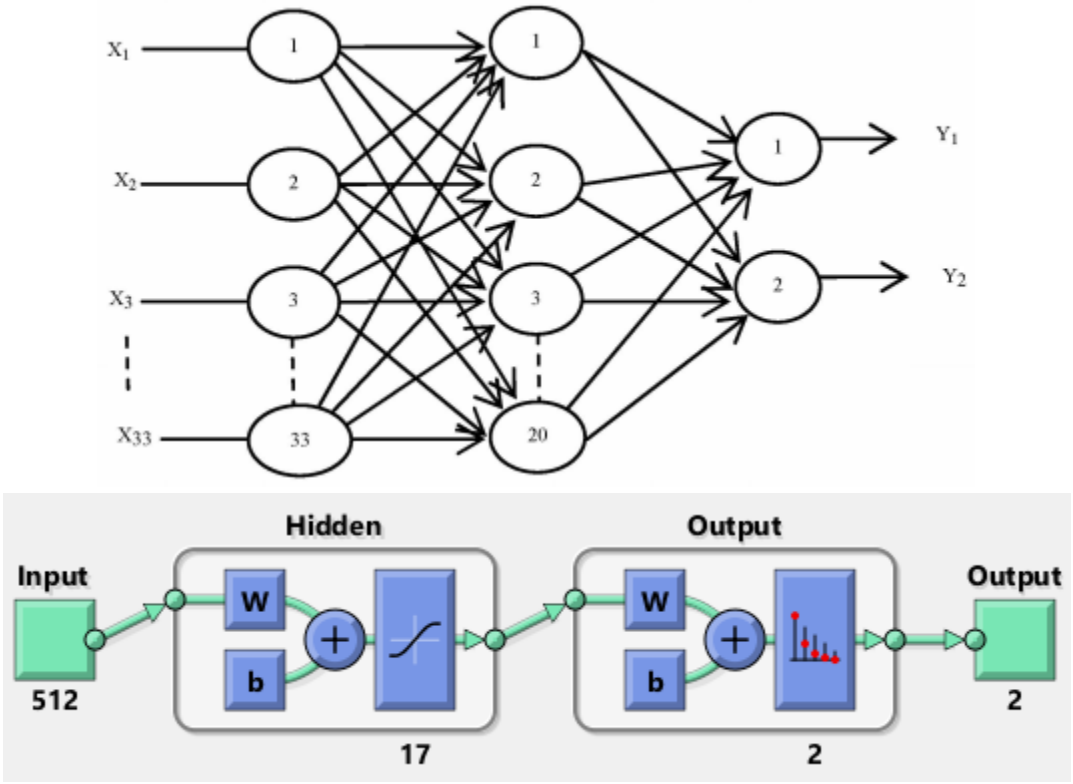


Figure 3: Neural Network design for the system

i. Training and Testing of NN

To train the neural network extracted features are used. The proposed system is designed such that it can detect whether the lungs are affected or not in the first phase of the system. **The system used 70% of the images from the healthy and the unhealthy lung images are used for training while the rest 30% of the images are used for the testing phase so that the system detects the lung cancer accurately. The total number of samples for training phase of NN classifier, were 624. The input of the classifier was**

a) **The data of an image data in the training set in the matrix form of 512 x 624 where 512 corresponds to the features extracted in one image and 624 are the total number of training images (each column of training matrix corresponds to one image).**

b) **The targets of the training set in the matrix form of 2 x 624 where the rows correspond to which image set (that is healthy or unhealthy) does the image belongs to whereas the column defines the total number of images. This matrix is a combination of binary values where it places a**

'1' if the image belongs to folder of healthy images while a '0' is placed in the second-row signifying that it doesn't belong to the 'unhealthy' set.

The system classifies the cancerous and non-cancerous CT scan images after training stage and specified whether the lungs are affected or not. And finally, the system is tested any positive and negative samples and it gives proficient results. The proposed system accurately identifies 99% of the images correctly of whether they are Infectious or Healthy Lungs as shown on figure 4.

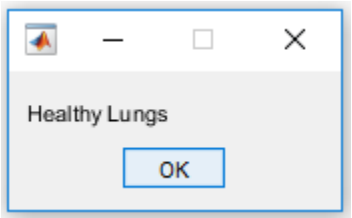


Figure 4: Result of NN Classification

Out of the 70% Training set, 75% was kept for training while 5% was kept for validation and rest 20% was kept for the testing within the NN classifier. The confusion matrix displayed in figure 5 helps explains the details further. The last entry of the 4th matrix determines the accuracy of the Neural Networks implemented for the proposed system. Overall, 99.8% of the predictions are correct and 0.2% are wrong.

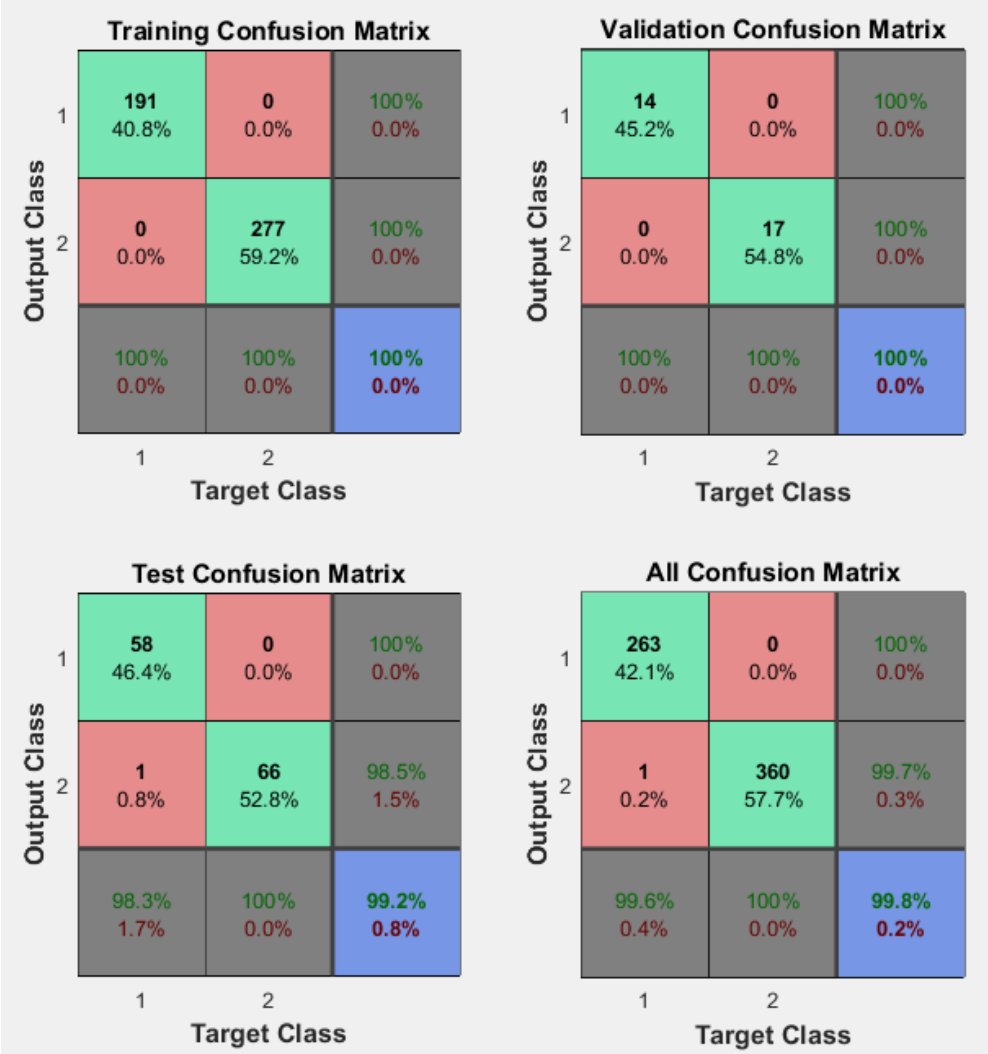


Figure 5: Division of Training set within Neural Network

Error Histogram is displayed in figure 6.

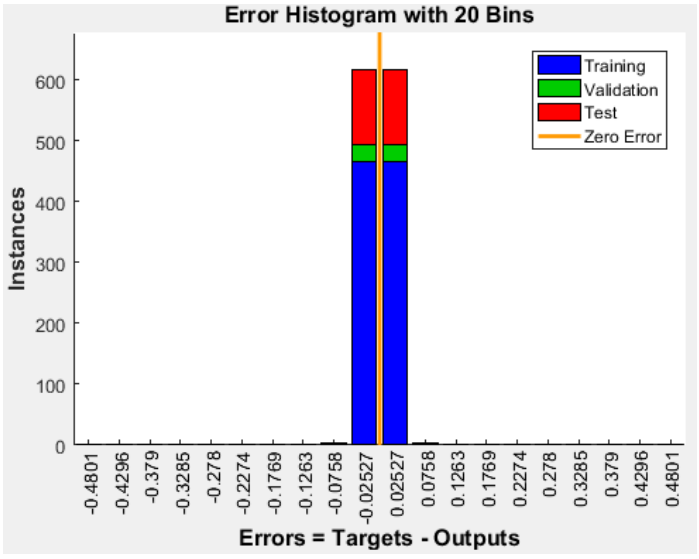


Figure 6: Error Histogram of Neural Networks

The confusion matrix of the implemented neural network executes the following statistics:

$$c = 0.016$$

$$cm = \begin{bmatrix} 263 & 1 \\ 0 & 360 \end{bmatrix}$$

$$per = \begin{bmatrix} 0.0028 & 0 & 1.0 & 0.9972 \\ 0 & 0.0028 & 0.9972 & 1.0 \end{bmatrix}$$

c	Confusion value = fraction of samples misclassified
cm	2-by-2 confusion matrix
per	2-by-4 matrix where each row represents the percentage of false negatives, false positives, true positives, and true negatives for the class and out-of-class

E. Discrete Wavelet Transform (DWT)

Once the first phase of the Neural Network Classification is done, we now need to verify the presence of the nodule. For that we use the already pre-processed image and apply the lossless 5/3 discrete wavelet transform (DWT) on it.

In 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).

The Discrete Wavelet Transform (DWT) became a very versatile signal processing tool after Mallat proposed the multi-resolution representation of signals based on wavelet decomposition. Wavelets allow both time and frequency analysis of signals simultaneously because of the fact that the energy of wavelets is concentrated in time and still possesses the wave-like (periodic) characteristics. As a result, wavelet representation provides a versatile mathematical tool to analyze transient, time-variant (non-stationary) signals that are not statistically predictable especially at the region of discontinuities – a feature that is typical of

images having discontinuities at the edges. The DWT decomposes a digital signal into different sub-bands so that the lower frequency sub-bands have finer frequency resolution and coarser time resolution compared to the higher frequency sub-bands. The results of the implemented 5/3 DWT is shown in figure 7.

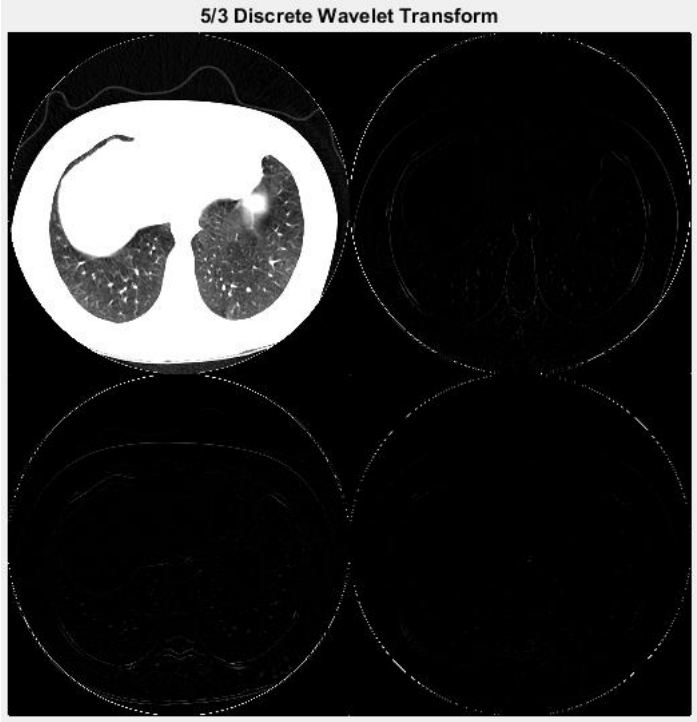


Figure 7: 5/3 Discrete Wavelet Transform

F. Deep Neural Networks (DNN)

After the DWT we have to perform Deep Neural Network (DNN) for more efficient and effective Nodule detection in the images. Once the infected lungs have been identified we need to move higher ahead in the research.

Deep-learning networks are distinguished from the more commonplace single-hidden-layer neural networks by their **depth**; that is, the number of node layers through which data passes in a multistep process of pattern recognition. In deep-learning networks, each layer of nodes trains on a distinct set of features based on the previous layer's output. The further you advance into the neural net, the more complex the features your nodes can recognize, since they aggregate and recombine features from the previous layer. The proposed DNN classifier has three layers.

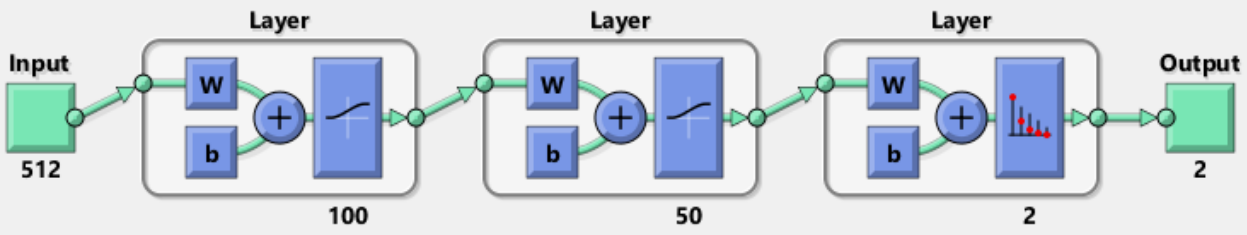


Figure 8: DNN classifier for the system

i. Training and Testing of DNN

To train the DNN extracted features from DWT are used. The proposed system is designed such that it can detect whether the infected lungs have a nodule present or not. **The system used 70% of the images from the unhealthy lungs are used for training while the rest 30% of the images are used for the testing phase so that the system detects the lung cancer nodule accurately. The total number of samples for training phase of DNN classifier, were 360. The input of the classifier was**

a) **The data of an image data in the training set in the matrix form of 512 x 360 where 512 corresponds to the features extracted in one image and 360 are the total number of training images (each column of training matrix corresponds to one image).**
b) **The targets of the training set in the matrix form of 2 x 360 where the rows correspond to which image set (that is Nodule or No Nodule) does the image belongs to whereas the column defines the total number of images. This matrix is a combination of binary values where it places a '1' if the image belongs to folder of Nodule images while a '0' is placed in the second-row signifying that it doesn't belong to the 'No Nodule' set.**

The system classifies the cancerous nodule CT scan images after training stage and specified whether the infected lungs have a nodule present or not. And finally, the system is tested with any positive and negative samples and it gives proficient results. The proposed system accurately identifies 100% of the images correctly of whether the infected lungs have a nodule present or not as shown in figure 9 and 10.

Figure 9: Nodule Detected

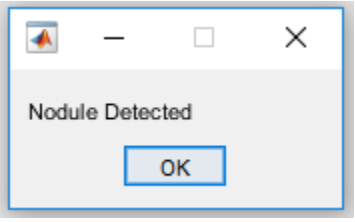


Figure 10: Result of DNN Classification

From the matrix displayed in figure 11, the last entry determines the accuracy of the Deep Neural Networks implemented for the proposed system. Overall, 100% of the predictions are correct and 0.0% are wrong.

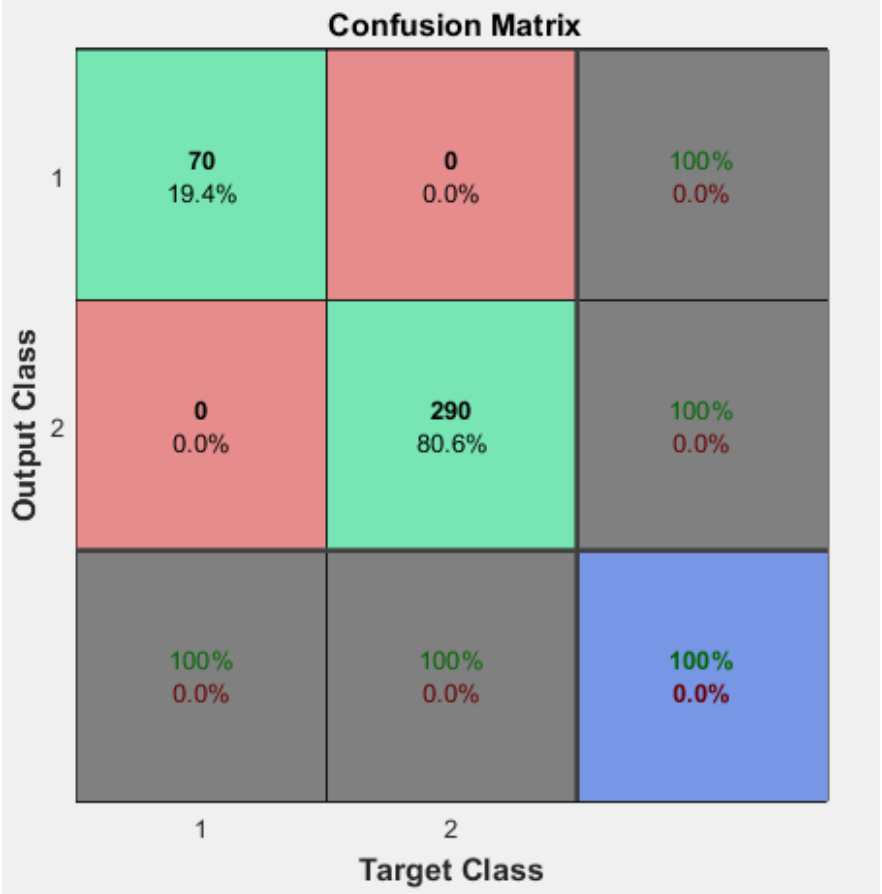


Figure 11: Confusion Matrix for Deep Neural Networks

The confusion matrix of the implemented deep neural network (DNN) executes the following statistics:

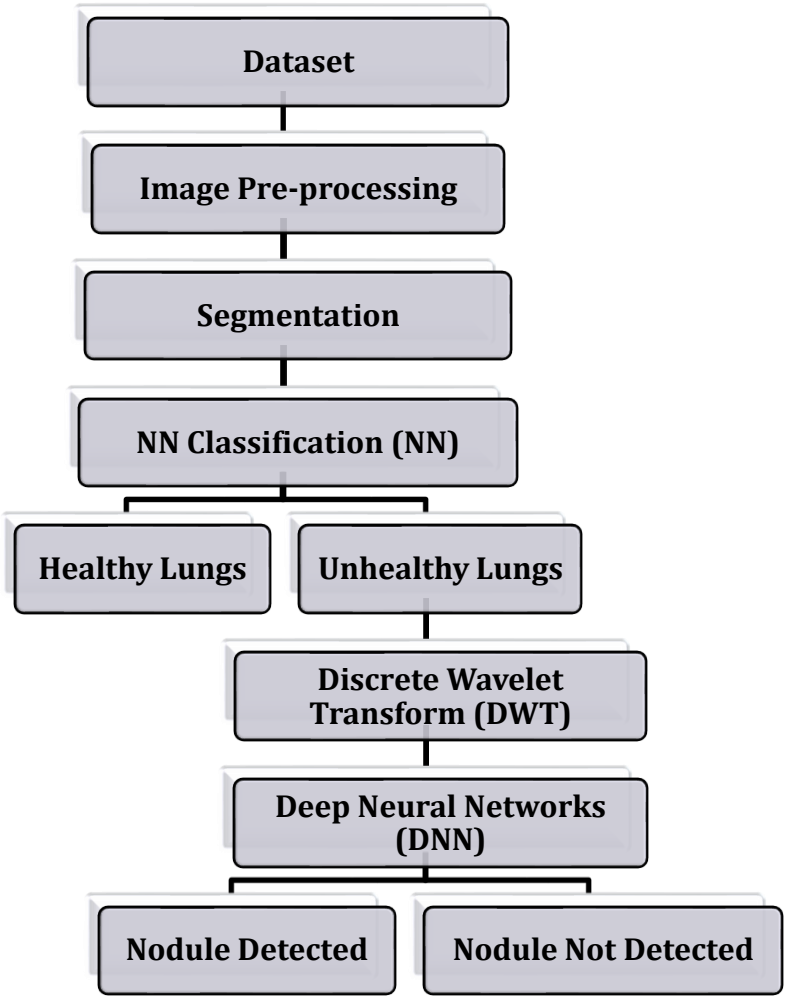
$$c = 0$$

$$cm = \begin{bmatrix} 70 & 0 \\ 0 & 290 \end{bmatrix}$$

$$per = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

c	Confusion value = fraction of samples misclassified
cm	2-by-2 confusion matrix
per	2-by-4 matrix where each row represents the percentage of false negatives, false positives, true positives, and true negatives for the class and out-of-class

Proposed Model Flowchart



Comparison to the Existing System

The proposed system introduces a unique pre-processing followed by a strong segmentation method and image transformation which helps train both the sets of classifiers efficiently as compared to other existing systems and achieve better performance for Lung Cancer Nodule Detection (LCND) system. The proposed system provides more accurate result compare than other existing system shown in the following Table 1.

Table 1: Comparison to the Existing System

Lung Cancer Detection System	Accuracy (%)
Lung Cancer Detection using Curvelet Transform 90% and Neural Network	90 %
Automatic Detection of Lung Cancer in CT Images	96 %
Early Detection and Prediction of Lung Cancer Survival using Neural Network Classifier	96.04 %
Gray Coefficient Mass Estimation Based Image Segmentation	83 %

Technique For Lung Cancer Detection Using Gabor Filters	
Identifying Lung Cancer Using Image Processing Techniques	80 %
Detection of Lung Cancer from CT Image Using Image Processing and Neural Network	96.67 %
Proposed System	**99 %**

Questions:

1. **If system does not localize the nodule and detects it then how we come to know that nodule is maligin or not maligin?**
   The malignancy of the tumor is distinguished by the amount of white area present in the lung portion of the image. If it exceeds a specific value it is malignant.
2. **For accuracy we should test the nodule using matrix acceptable in the research community**
   Multiple images have been tested and each time the detection has been done correctly. Roughly 7GB of CT scan images have been used for training and testing.
3. **E.F FAR (false acceptance rate) sensitivity, specificity, true false negativity and positivity rate**.
   These specifications have been highlighted above in the confusion matrix.
4. **By using proper equations explain how you have calculated the accuracy**
   The confusion matrix determines the accuracy of the classifiers. It can be calculated by taking average of the values lying across the main diagonal of the confusion matrix.
   Classification Accuracy:

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

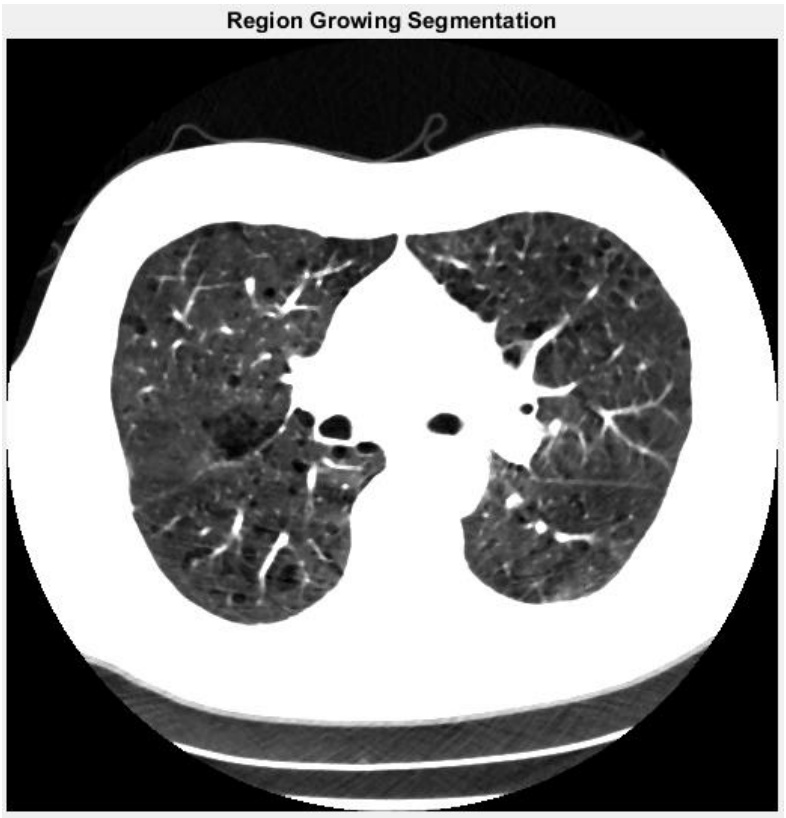
   It can be calculated by taking average of the values lying across the main diagonal of the confusion matrix.

$$Accuracy = \frac{TruePositives + FalseNegatives}{TotalNumberofSamples}$$

$$Accuracy\ for\ NN = \frac{263 + 360}{624} = 0.998$$

$$Accuracy\ for\ DNN = \frac{70 + 290}{360} = 1.0$$

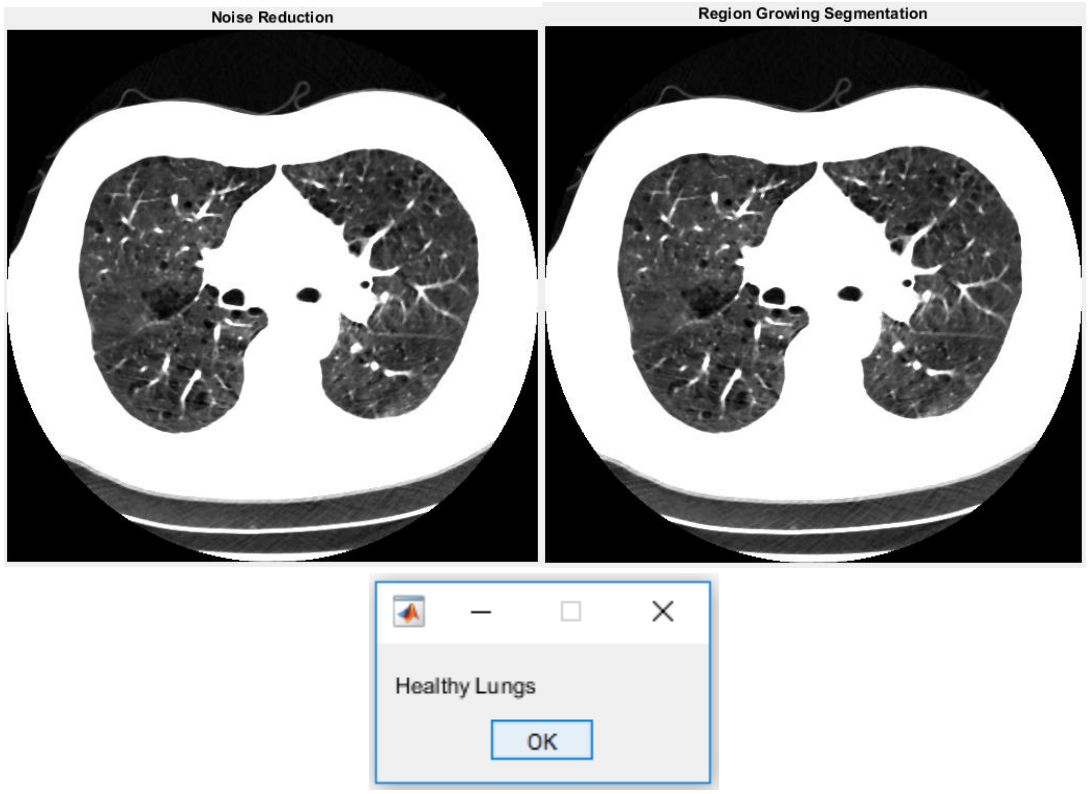
5. **How did you apply region growing segmentation and what results you get?**
   The RG is implemented using a specified seed point from the image. The region is iteratively grown by comparing all unallocated neighboring pixels to the region. The difference between a pixel's intensity value and the region's mean, is used as a measure of similarity. The pixel with the smallest difference measured this way is allocated to the respective region. This process stops when the intensity difference between region mean and new pixel become larger than a certain threshold.



6. **Why it is only detecting the nodule in images? Why is it not locating the place of the nodule?**

   Many studies in the research have proposed such an algorithm which locates the nodule. In the proposed system, the main aim was to detect the nodule and increase the accuracy of the system as compared to the already implemented studies. The proposed system offers a unique approach with

7. **Explain results of segmentation, classification and detection separately.**



8. What is innovation in this research or what is your contribution?

The proposed system has increased the accuracy up to 99.8 % which uses a unique set of pre-processing, segmentation and image transformation properties and 2 layers of classifiers to finally detect the nodule.