LUNG CANCER DETECTION USING IMAGE PROCESSING

A PROJECT REPORT

Submitted by

NAME	REGISTRATION NUMBER
B.Rupendra Reddy	BL.EN.U4AIE19009
D.Sai Tharun Reddy	BL.EN.U4AIE19016
Sandeep Preetham M C	BL.EN.U4AIE19058

in partial fulfillment for the award of the degree

Of

BACHELOR OF TECHNOLOGY

IN

ARTIFICIAL INTELLIGENCE ENGINEERING

AMRITA SCHOOL OF ENGINEERING, BANGALORE

AMRITA VISHWA VIDHYAPEETHAM

BANGALORE-560 035

Dec - 2021

AMRITA VISHWA VIDHYAPEETHAM

AMRITA SCHOOL OF ENGINEERING, BANGALORE,560035

BONAFIDE CERTIFICATE

This is to certify that the project report entitled " **LUNG CANCER DETECTION USING IMAGE PROCESSING ”** submitted by

BL.EN.U4AIE19016	SAI THARUN REDDY
BL.EN.U4AIE19009	BANDI RUPENDRA REDDY
BL.EN.U4AIE19058	SANDEEP PREETHAM

”In partial fulfillment of the requirements for the award of the **Degree Bachelor of Technology** in “**ARTIFICIAL INTELLIGENCE** Engineering “ is a bonafide record of the work carried out under my(our) guidance and supervision at Amrita School of Engineering, Bangalore.

NAME OF SUPERVISOR

Dr.Suja P(CSE)

<<Signature of the Chairperson of the Department with date>>

NAME OF CHAIRPERSON-……..

This project report was evaluated by us on …… (Date...)

Acknowledgment

The satisfaction that accompanies the successful completion of any task would be incomplete without mention of people who made it possible, and whose constant encouragement and guidance have been sources of inspiration throughout the course of this project work.

It is a great pleasure to express our gratitude and indebtedness to our project guide Dr.Suja P, Amrita School of Engineering, Bangalore for his valuable guidance, encouragement, moral support, and affection throughout the project work.

TABLE OF CONTENTS:

TABLE OF FIGURES:

ABSTRACT :

In the field of medicine, identification and treatment of cancer are considered as one of the biggest challenges in the treatment of chronic illness. The survival of patients depends on timely detection and cure. Experts use the CT scan or Computed Tomography Scan images of patients to detect and classify nodules, before proceeding with advanced treatment procedures. The present-day advances in artificial intelligence, machine learning based on deep learning models can be used to develop sophisticated Computer-Aided Diagnosis systems to detect cancerous nodules. The proposed system is based on Convolutional Neural Networks to categorize nodules detected in CT scan images as malignant or benign. Image processing and Neural Networks have been extensively used in the detection and classification of cancerous nodules. Hence CNNs are more appropriate, for the task of nodule detection and classification. CNN's have more properties like multiple feature extraction. When convolution layer, subsampling or pooling layer, fully connected layer such layers are combined, leading to Deep CNNs, it helps in increasing the accuracy of classification. The proposed CNN model will be suitable for the early detection and classification of CT scans images containing nodules with good accuracy, using the domain knowledge of the CT scan images of lungs in the field of medicine and Neural Network.

Keywords:-

Cancer Detection, Image processing, CNN model, CT scan.

CHAPTER 1

INTRODUCTION :

Cancer is the most prevalent terminal disease globally and lung cancer is the second most common cancer in both men and women. About 13% of new cancers are lung cancer, accounting for an estimated 10.6 million deaths in 2020. There are many types of cancers, in those lung cancers is one of the frequently occurring diseases that cause death and is identified in both genders over an estimated death of 1.95 million in 2020. Although surgery, radiation therapy, and chemotherapy have been used in the treatment of lung cancer, the five-year survival rate for all stages combined is only 14%. This has not changed in the past three decades. The survival rate of people suffering from this disease depends on which stage the cancer has been diagnosed. It is difficult to detect because it arises and shows symptoms in the final stage. However, mortality rate and probability can be reduced by early detection and treatment of the disease. In recent years, the people showing the symptoms of lung cancer can be detected by a test known as a CT scan of the chest which has helped the doctors or radiologists to analyze the shape, size, and position of lung nodules or tumors which have spread across the body. CT scan images are used for the detection of lung or pulmonary (means related to lungs) nodules because of their high sensitivity.

.

CT scanner takes many detailed pictures or images of the internal organ which are combined by computer into the image of slices of the part of the body in which few nodules or the non-pathological structures which are round figured in the shape but, all are not exactly cancerous. So it becomes a very time-consuming and tedious task to analyze the tiny or small nodules for a radiologist. However, radiologists can be assisted by the CAD (computer-aided diagnosis or detection) system. The CAD system processes the digital images to highlight the conspicuous area or section affected by the disease. It helps doctors and radiologists to interpret the medical images. This technology has combined artificial intelligence, machine learning algorithms, and computer vision with image processing.

The CAD system can examine a large number of CT images using image processing techniques such as preprocessing, segmentation, feature extraction, image acquisition, and also deep learning-based methods for better and early diagnosis of pulmonary nodules. This research work presents a novel approach to the CAD system based on CNN (convolutional neural network) and different image processing techniques.

The approach proposed a method where a CT scan image is taken as an input image on which different image processing techniques were applied step by step:

1) Preprocessing of the input image
2) Histogram equalization
3) Segmentation by a thresholding technique
4) Morphological operations which included – image filtering, dilation, image filling.
5) Feature extraction. It is performed using the GLCM algorithm to extract the features of the input image. Different textural features are extracted such as shape, contrast, entropy, and many other features which helped in classification.
6). Finally, the classification of the detected nodules is done using CNN as the classifier. This classification will classify whether the nodule is cancerous or non-cancerous.

The system is trained and the parameters obtained are in the terms of sensitivity, accuracy, specificity, and time taken by the system for each image to detect cancer. Software used for the proposed method is implemented in MATLAB. Working on the CAD system in this research work starts with preprocessing of input CT scan image. Next histogram equalization, lung segmentation using thresholding technique, and different morphological operations are performed on the image.

GLCM (gray-level co-occurrence matrix) algorithm is used for the extraction of the features from the input image. In the last step, classification is done using CNN classifier by training the neural network system and parameters are obtained in the result

CHAPTER 2

SYSTEM DESIGN :

The Proposed System is explained step by step as given below:-

1. Preprocessing:

Preprocessing is the first step of the system. In this CT scan image is taken as the input image. The aim of this step is to convert the image into a grayscale image and to remove the noise from them using a filter. It also increases the contrast of the image giving an enhanced version of the original image. As we know that the medical images are grayscale by default but still the raw CT scan images taken as input images are still converted to grayscale to increase their quality.

2. Histogram Equalization:

After preprocessing of input image histogram equalization is performed. It is a method that uses an image's histogram to adjust the contrast in image processing. It smoothens the histogram of an image and increases global contrast and adjustment of this contrast allows the better intensity distribution. The dark image represents the lower level histogram and the overexposed image indicates a higher level of the histogram. Histogram of chest CT scan input image consists of peaks and valleys representing different regions of the lungs which are equalized and a clear, crisper image with sharp borders and edges is obtained. It also highlights the required objects in the grayscale image giving the enhanced version of the image.

3. Segmentation:

In this step, the image is converted into a binary image or black and white image. As the binary image is a digital image and its pixels can be represented in 0 and 1 which are two discrete levels. Level 1 indicates the presence of data that is white color and level 0 indicates the absence of the data that is black color. The goal of segmentation is to partition the digital image into multiple segments to locate the required objects so that it could be easy to analyze the nodules in the image. In this approach, segmentation is done using Otsu's thresholding technique. The segmentation of the CT scan input image is done by removing those pixels which are below or above the constant level called discrete level (threshold value).

In the presented work, a grayscale picture is utilized for the thresholding process. The binary image obtained from thresholding comprises several benefits like lesser storage space, speedy dispensation velocity, and easiness in exploitation in comparison with a gray level picture that generally includes 256 steps but binary image have two levels 0 and 1 making it easier to analyze for the radiologist the required information for detection of the nodules. segmentation will give the ROI (region of interest) that is the nodules that are conspicuous ones and need to be diagnosed whether they are cancerous or not. ROI helps in feature extraction. Before feature extraction different morphological operations are applied which are explained in the next step.

4. Morphological Operations:

The morphological operations are performed before feature extraction. In a morphological operation, the adjustments of pixels in the images are done which is based on its neighborhood pixel's value. These operations are used to extract the image components which are useful in the description of region shapes and give better representation. The operations used in this research are filtering dilation and image filling.

a). Filtering :

It is a very useful and effective technique to reduce the noise in the image. This image filtering allows the morphological operators such as dilate, erosion open, and close to being applied on the image with these filters. With image filtering, the image regions can grow or shrink and can fill in or remove the image region boundary. The Sobel edge detector is used for filtering in this research work.

b). Dilation :

The addition of pixels to the boundaries of the object in the image is known as dilation. A structuring element is added to the input image to get the output image. The number of pixels added or removed depends on the shape and size of the element.

c) Image filling:

This operation fills the holes which can be defined as the background regions in the input binary image.

5). Feature Extraction:

Different features are extracted from the detected nodules in the image, using the GLCM (Gray Level co-occurrence matrix) algorithm. The GLCM algorithm will extract the textural features of the input image. The gray co-occurrence matrix is created and several statistics are derived from the matrix which analyzes the textural features in the images. The features like energy, entropy, homogeneity, contrast can be extracted after the segmentation process is done in the region of the lung, for analyzing some diagnoses to identify cancer in the lung correctly. The extraction of the texture features using GLCM has been considered a significant technique and it has been used in several applications of remote sensing for analysis of the texture. GLCM points to the Gray level Co-occurrence matrix. It is of the second-order statistics, so information with regards to pixels of pairs is collected by the GLCM algorithm. GLCM uses the graycomatrix function. The graycomatrix function creates a gray-level co-occurrence matrix (GLCM) by calculating how often a pixel with the intensity (gray-level) value *i* occurs in a specific spatial relationship to a pixel with the value *j* GLCM exhibits how pixel brightness occurs in an image.

By default, the spatial relationship is defined as the pixel of interest and the pixel to its immediate right (horizontally adjacent), but you can specify other spatial relationships between the two pixels. Each element (*i*,*j*) in the resultant glcm is simply the sum of the number of times that the pixel with a value *i* occurred in the specified spatial relationship to a pixel with value *j* in the input image. By default, the graycomatrix function creates a single GLCM, with the spatial relationship, or *offset*, defined as two horizontally adjacent pixels. However, a single GLCM might not be enough to describe the textural features of the input image.

To create multiple GLCMs, specify an array of offsets to the graycomatrix function. These offsets define pixel relationships of varying direction and distance. For example, you can define an array of offsets that specify four directions (horizontal, vertical, and two diagonals) and four distances. In this case, the input image is represented by 16 GLCMs. When you calculate statistics from these GLCMs, you can take the average.

we specify these offsets as a *p*-by-2 array of integers. Each row in the array is a two-element vector, [row_offset, col_offset], that specifies one offset. row_offset is the number of rows between the pixel of interest and its neighbor. col_offset is the number of columns between the pixel of interest and its neighbor.

After you create the GLCMs, you can derive several statistics from them using the graycoprops function. These statistics provide information about the texture of an image. The following lists the statistics we can derive.

a). Energy - This feature is used for optimization or minimization or maximization. It performs gradient-descent and computes its lowest value and gives the desired output for image segmentation.

$$\text{Energy} = \sqrt{\sum_{i,j=0}^{N-1} p_{i,j}^2}$$

b). Entropy - This feature is used to characterize the texture of an input image which describes the measure of the degree of the randomness.

$$\text{Entropy} = \sum_i p_i \log_x i$$

c). Contrast - Here contrast is the ratio of difference of maximum intensity value and minimum intensity value upon the sum of maximum and minimum intensity value. It Measures the local variations in the gray-level co-occurrence matrix.

d). Homogeneity This statistic is also called the Inverse Difference Moment. It measures image homogeneity. It is sensitive to the presence of near diagonal elements in the GLCM. When all elements in the image are the same it gives maximum value. GLCM homogeneity and contrast are strong but inversely correlated in terms of equivalent distribution in the pixel pair population. It means if contrast increases homogeneity decreases while energy is kept constant.

$$Homogeneity = \sum_{i,j=0}^{N-1} \frac{P_{ij}}{1+(i-j)^2}$$

Classification:-

For the classification of the input images, a deep learning method is used that is a class of deep neural networks known as CNN (convolutional neural network). It is used to analyze visual imagery and has different applications such as image classification, image and video recognition, object detection, feature, and medical image analysis. In the last step, the approach of CNN is applied which can categorize and locate the cancerous nodules. The input image is fed into the network which consists of different layers including the input layer, hidden layer, and the output layer. Simple diagram of the artificial neural network showing three layers: input layer, hidden layer, output layer.

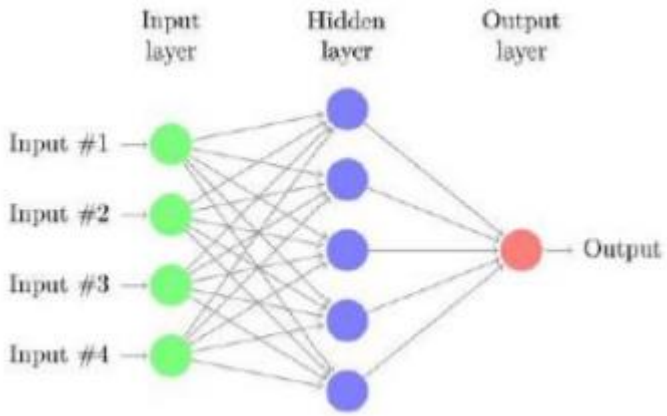


Fig 2.1 - CNN Architecture

The hidden layer consists of multiple layers including the convolutional layer, pooling layer, ReLU (rectified linear unit) layer (it is an activation function), and fully connected layer in them. The system starts with an input image and creates a feature map by applying different types of filters to it. A pooling layer is applied to each feature map and an activation function so that non-linearity gets increased. The pooled image is flattened into one long vector. Now the vector is taken as input and fed into the fully connected artificial neural network. The network processes the features and the fully connected layer provides the classification and categorizes the tumor-affected images and the normal images.

The system is trained through the forward and backward propagation for many epochs and till we get a well-defined neural network with the feature detectors and trained weights. When the minimum mean square error is achieved by the system the network will stop updating its weights. The training performance is based on the mean square error with respect to epochs.

CHAPTER 3

IMPLEMENTATION :

The proposed model is implemented in Matlab. For the implementation, real CT scan images are obtained from internet sources, and then images are converted to JPEG grayscale images using the Matlab commands. Features extraction and detection are implemented in Matlab GUI and classification is implemented using the image processing techniques.

Firstly we import the CT scanned to our code. Based on that CT image we classify whether the image is normal or lung cancer affected.

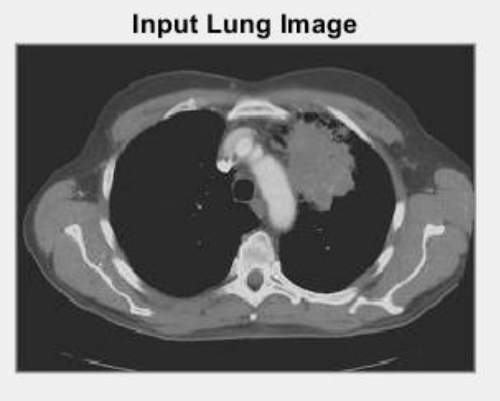


Fig 3.1 - CT scan image

Next, the CT image is converted to the grayscale image to perform the mathematical operations, Histogram of the chest CT scan input image consists of peaks and valleys representing different regions of the lungs which are equalized, and a clear, crisper image with sharp borders and edges is obtained.

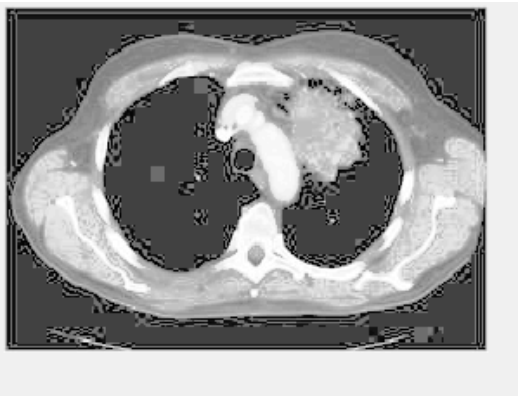


Fig 3.2 - Grey-scale image

And in the segmentation step, the image is converted into the binary image or black and white image as shown in the below figure. As the binary image is a digital image and its pixels can be represented in 0 and 1 which are two discrete levels.

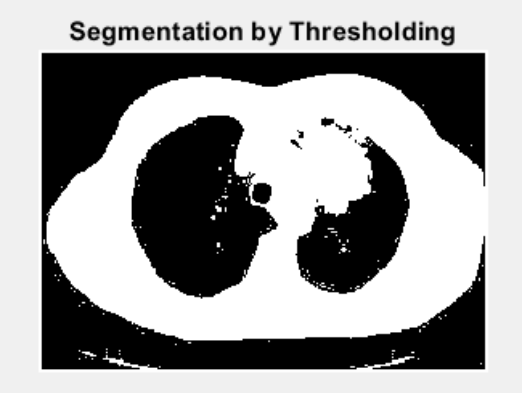


Fig 3.3 - Segmented Image

Now it is ready to carry out further mathematical operations, now we can morphological operations to extract features. The image is passed through a filter to enhance the information needed to be shown. And With the image filtering, the image regions can grow or shrink and can fill in or remove the image region boundary.

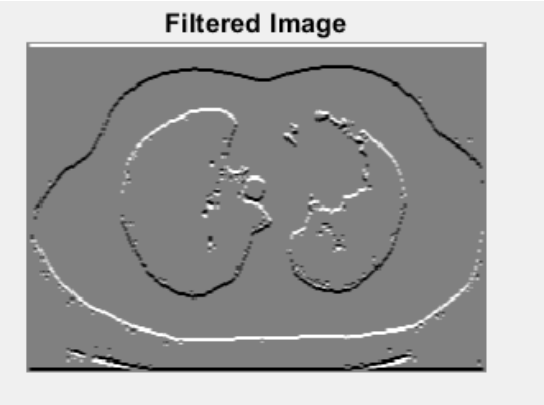


Fig 3.4 - Filtered image

The next step in the morphological operation is dilution to the image which is used to the addition of pixels at the boundaries. A structuring element is added to the input image to get the output image. The number of pixels added or removed depends on the shape and size of the element.

Fig 3 .5 - Dilated image

The last step in morphological operation in our project is filling the holes. This operation fills the holes which can be defined as the background regions in the input binary image.

Fig 3.6 - image filling

Then after in the feature extraction step, we create a GLCM (Gray Level Co-occurrence Matrix) GLCM texture that picks up the relation between two pixels at a time, called the Neighbour and the reference pixel. The GLCM algorithm will extract the textural features of the input image.

The gray co-occurrence matrix is created and several statistics are derived from the matrix which analyzes the textural features in the images, after the segmentation process After you create the GLCMs, you can derive several statistics from them using the graycoprops function. These statistics provide information about the texture of an image. The following lists the statistics we can derive.

Fig 3.7 - Extracted Features

Then getting the features like Entropy, construct, Energy we find above and by training the CNN model with the dataset that we load the CNN model will predict whether the Image is Non-cancerous or not. A convolutional neural network (CNN) is a type of artificial neural network that has been designed for reduced processing requirements. The layers of a CNN consist of an input layer, an output layer, and a hidden layer that includes multiple convolutional layers, pooling layers, fully connected layers, and normalization layers. The removal of limitations and increase in efficiency for image processing results in a system that is far more effective, simpler to trains limited for image processing and natural language processing. A Convolutional neural network (CNN) is a neural network that has one or more convolutional layers and is used mainly for image processing, classification, segmentation, and also for other autocorrelated data.

Our proposed system followed data acquisitions, data formatting, model training, testing, and prediction, described in the below sections.

A. Data acquisition:

The images obtained from the database are taken to the data formatting. whether to predict the image is a normal image or lung cancer-affected image.

B. Data formatting:

The obtained dataset was RGB images with .jpeg format. The images were resized to maintain a uniform aspect ratio of one with pixel size for the CNN operation. All the pixel values for the images were converted to range of (0, 1) to make convergence faster.

C. Model Training, Testing, and Prediction:

A linear stack of layers was used to create the Convolutional Neural Network (CNNs or ConvNets)for the image classification and recognition. Training and testing images were passed through convolutional layers with kernel filters, max pooling, and fully connected layers. The softmax function was applied to classify the given object. The loaded image after preprocessing steps goes to the CNN model. In that model we have already trained and tested the dataset so when we load it will detect the image and predict the result. The below diagram represents the complete flow of our project.

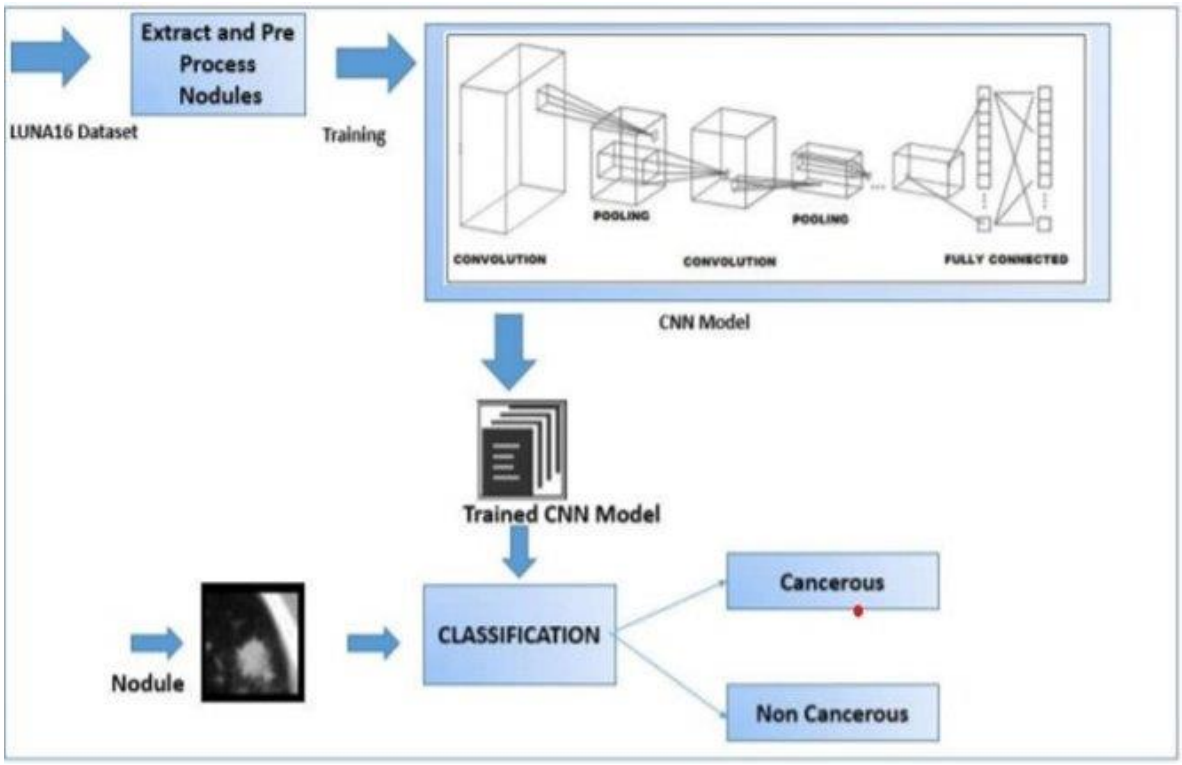
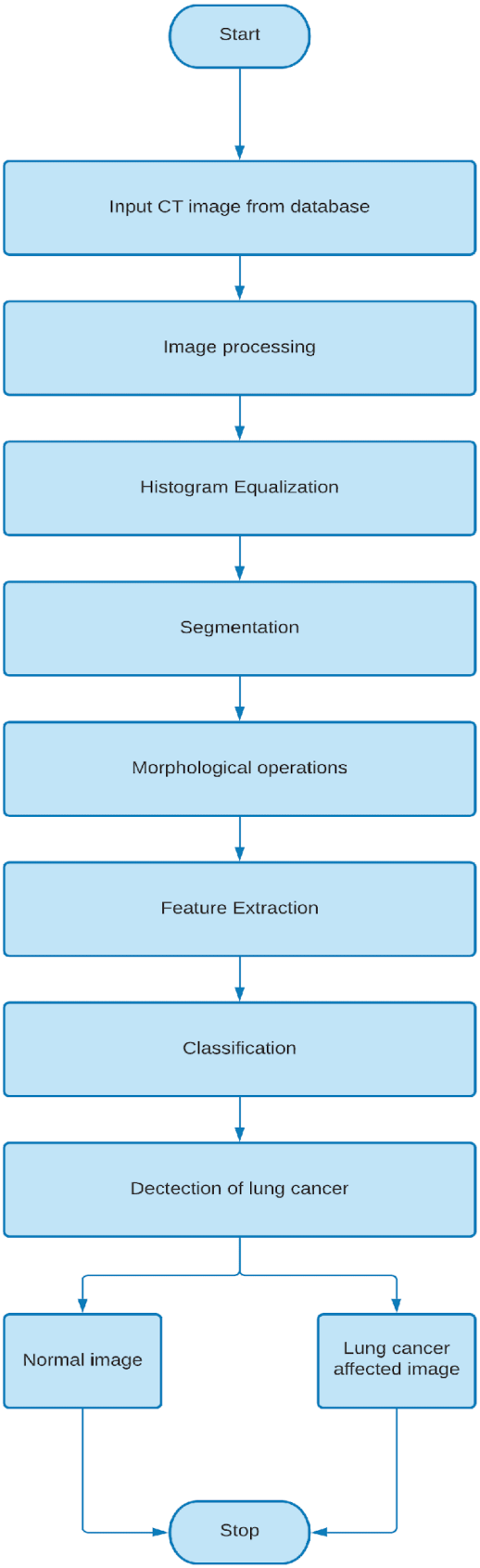


Fig 3.8 Block diagram of CNN model

The process of workflow is shown below figure :



CHAPTER 4

RESULTS :

The proposed approach was implemented in MATLAB. The common tools to analyze the performance of the system are sensitivity, specificity, and accuracy which gives the desired result. These parameters are obtained after the CNN approach is applied for the classification. These parameters can be described by knowing the basic terms false positive, false negative, true positive, and true negative which were used to calculate the result. A brief description of the parameters obtained through them is as follows:

Sensitivity is defined as the relationship between the true positive and false negative results obtained whereas true positive (TP) is the actual nodules that are cancerous and are correctly identified by the system. False-negative (FN) is when the result obtained is a negative result but the expected result is a positive result. This means nodules that are missed by the system. Thus it can be calculated as given below:

Sensitivity = TP/ TP+FN

Specificity is the relationship between the true negatives and false positives where false positive (FP) is when the result obtained is a positive result but the expected result was to be a negative result. This means the non-cancerous nodules are detected as cancerous nodules. True negative (TN) are the nodules that are non-cancerous and are correctly identified as non-cancerous by the system. Thus it can be calculated as given below:

Specificity =TN/ TN+FP

The false-positive rate is the rate obtained after every single scan and can be calculated by the formula given below.

False-positive rate= 1 – specificity Accuracy of the system is analyzed by the below-given formula that is the sum of true positive and true negative upon the total sum of all the terms that are truly positive, true negative, false positive, and false negative.

Accuracy = TP+TN/ TP+TN+FN+FP

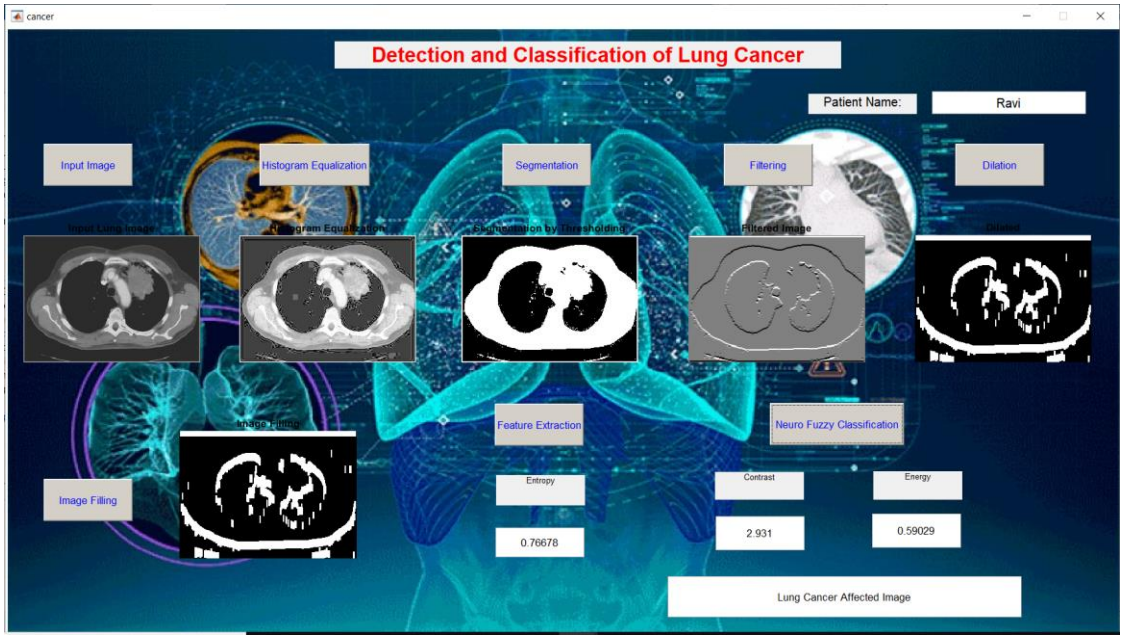


Fig 4.1 - GUI for lung cancer detection

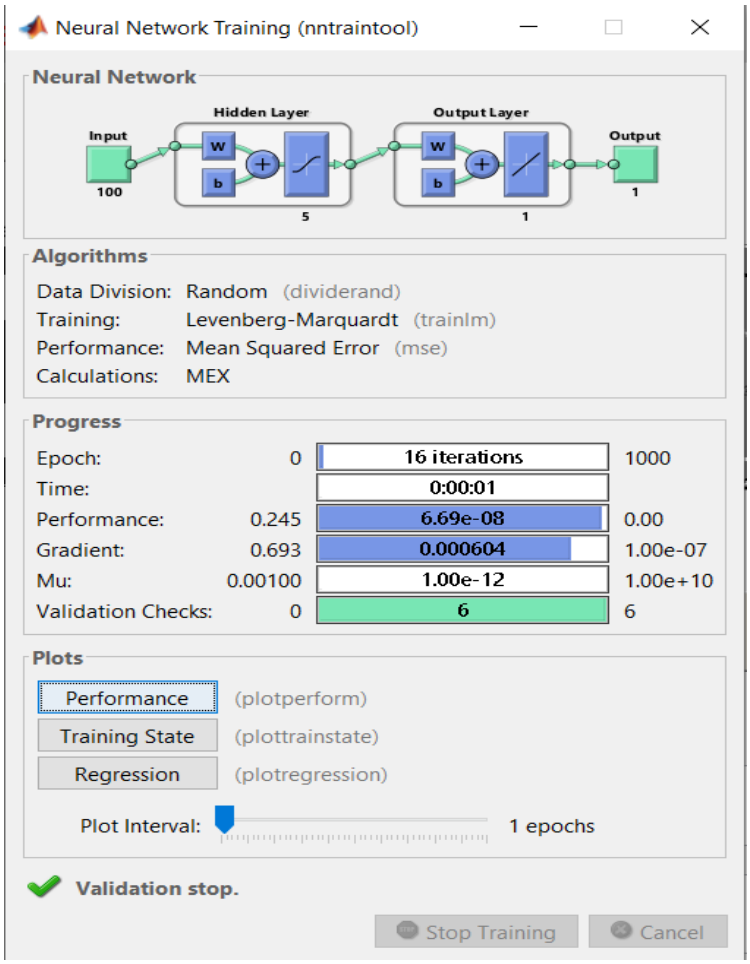


Fig 4.2 - Neural network toolbox

The proposed algorithm uses computer vision and a neural network toolbox for implementation. The interface of lung cancer detection is shown in the project. In the figure, all the techniques performed and the tumor portion marked image is illustrated. The system will be trained to generate the desired results. The training of the neural network system is shown in the above-given Figure

CHAPTER 5

CONCLUSION & FUTURE WORK: -

For lung cancer detection image processing is used. There are three steps for the detection of cancer. To detect the presence of cancer CT scan images are used. Further, the pre-processing is composed of two processes. Image enhancement and image segmentation are two processes. For human viewers, the interpretability of information in the image is improved by the image-enhancing step. There are many enhancement algorithms such as Gabor filter, fast Fourier transform, log Gabor filter, and auto enhancement. In pre-processing the second step is Image segmentation. The purpose of image segmentation is to partition the image into meaningful regions and to identify the object or relevant information from the digital image. The output from the segmentation process is going to feature the extraction stage. Features such as area, perimeter, and irregularity are found out in feature extraction. On the basis of the extracted features, the abnormality in the lung is found out by the cancer cell identification module. The approach of GLCM and CNN will be used in this research work for localizing and characterizing cancer portions from the CT scan image. The proposed approach is implemented in MATLAB and results are analyzed in terms of accuracy, sensitivity, and specificity. For future work, the proposed methods can also be applied to some other cancer types like brain tumors, skin cancer, stomach cancer, breast cancer, etc.

CHAPTER 6

REFERENCES :

Hatuwal, Bijaya & Thapa, Himal. (2020). Lung Cancer Detection Using Convolutional Neural Network on Histopathological Images. International Journal of Computer Trends and Technology. 68. 21-24. 10.14445/22312803/IJCTT-V68I10P104.

Chunran, Y., Yuanvuan, W., & Yi, G.X. (2018). Automatic Detection and Segmentation of Lung Nodules on CT Images. 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 1-6

https://www.ijert.org/research/lung-cancer-detection-using-machine-learning-IJERTCONV7IS01011.pdf