Machine Learning Approach to Classification of Lung Nodule on CT Images and C-Means Clustering Methods

D. Napoleon
Assistant Professor
Department of Computer Science, Bharathiar University, India

I.Kalaiarasi
Research Scholar
Department of Computer Science, Bharathiar University, India

Abstract- Lung cancer is very dangerous than other types of cancer in human. In recent years lung cancer detection is a challenging factor for researchers. The timely and primary detection of lung cancer is increasing human survival rate which is performed by using image processing techniques. The proposed system describe lung cancer, its types and detection methodology, classification techniques etc., The proposed system consists of many steps which are pre-processing, segmentation, thresholding, feature extraction, classification. Pre-processing can perform by using Weiner filtering and segmentation for separating the nodule from the lung image based on the threshold value. Next extract some features from a segmented region which can be given as an input for classification. Here the classification is done through the artificial neural network. Finally, the performance of the proposed system has been calculated which is helpful to the doctors and radiologists to diagnosing a patient at their early stage of lung cancer.

Keywords – ***CT Image, Lung Nodule Classification, MATLAB, ANN, C-Means cluster***

I. INTRODUCTION

The lung is the main organ in every human which is functioning as a gas exchange called respiration or breathing. Lungs are air-filled organs it is located on thorax which has important parts for the lung to function properly. The trachea process the air inhaled into the lung through the bronches called bronchi. The bronchi divide into two types as smaller and smaller bronchioles which is end up at the alveoli. In alveoli, the inhaled oxygen is let into the blood and carbon dioxide is let from blood to alveoli. The lungs are protected by the pleura and the thin fluid act as a lung smoothly which is helpful to expand while breathing. Finally, on the under surface, the lungs are bordered by the diaphragm (Figure-1).

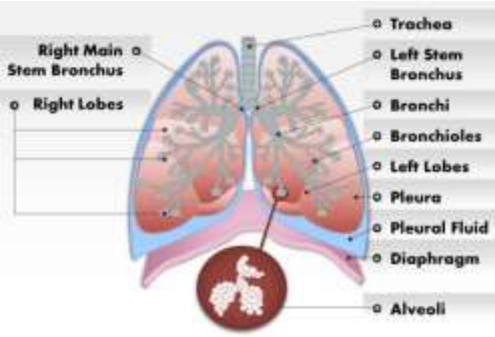


Figure1. Lung Anatomy

1.1 Lung Cancer -

Lung cancer is a type of cancer that starts in lungs which is the cells are uncontrollably growth as tissue to reduce the human breath. In the United States, lung cancer is the leading cause of cancer which affects the normal function of lungs such as receiving oxygen and releasing of carbon dioxide itself. Mostly the reason for lung cancer is smoking and some infectious things, lung cancer is the highest rate among men and women comparatively other cancer types. In Figure.2 the tumour growth in the lymph node which affects the normal lung cells to breath.

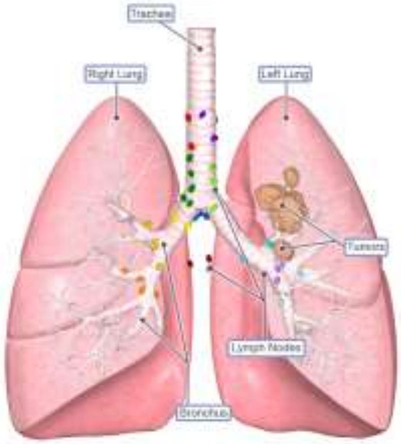


Figure 2: Lung cancer

[1]According to the American cancer society's estimate for lung cancer in united states for the year 2020 is given in Table. 1. Due to this estimation, in the year 2020 how much lung cancer cases and deaths in both men and women in the United States.

Table – 1. Cancer Cases and Deaths in the United States

	Total	In Men	In Women
New Cases of Lung Cancer	**228,820**	**116,300**	**112,520**
Deaths from Lung Cancer	**135,720**	**72,500**	**63,220**

1.2. Types of Lung Cancer-

Lung cancer is of two types Small Cell Lung Cancer (SCLC) and Non-Small Cell Lung Cancer (NSCLC). The types are based on the size of the tumour. [2] SCLC is related to cigarette smoking which is very aggressive and rapidly growing among all the types.

Non-small cell lung cancer types and % of cases and its arising place.

Table -2**.** NSCLC

Types of NSCLC	Cancer Cases in %	Arising place in lungs
Squamous-cell carcinoma	**25-30%**	**In bronchial tube**
Adenocarcinoma	**40%**	**Alveolar cells**
Large cell (undifferentiated) carcinoma	**5-10%**	**Central part of a lung**

Lung cancer is one of the leading causes of cancer deaths in this world compared with other types of cancer. At present the lung cancer has been estimated in world-wide there are 1.61 million new cases of lung cancer in every year, with the death rate is 1.38 million [3]. At every year there are 63000 new cases of lung cancer is reported in

India [4]. Lung cancer is an uncontrollable growth of abnormal cells that are on both sides of the lung. Normally the nodule [or mass] is found in the lung which is representing as benign and malignant depending on that size. The size is from 0.2 inches [5 millimetres] to 1.2 inches [30 millimetres] and above. Therefore the size may increase or not, if it increases>30 millimetres then it became cancerous. So the detection of the nodule and is classified as benign or malignant is always a challenging task for the doctors to diagnose the disease at an early stage [5]. Nowadays there are several existing techniques to classify the lung nodule as malignant that are CT, MRI, X-Ray and sputum cytology and so on. The image processing techniques are used to detect the pulmonary nodule as normal or abnormal which is helpful to doctors, diagnose the patient at their early stage and increase the survival rate of them. Here the related works are described in section (2), the Methodology in section (3). In section (4) discussion and finally the conclusion in section (5).

II. RELATED WORKS

Many researchers are done their work for lung cancer is to give better accuracy in detecting and classifying the cancer nodule. Normally lung segmentation is more important for diagnosing a patient because depends on the segmented lung affected region only the doctors can treat a patient. There are two segmentation algorithms are very familiar and effective in segmenting the region. Prionjit Sarker [6], using the k-means and c-means algorithm for calculating the lung affected region and finding the tumour stage. Dzulkifli Bin Mohammad and M. Masroor Ahmed proposed k-means clustering algorithm with Anisotropic Diffusion Model of Perona-Malik for tumour detected and segmented from the affected region. A. Amutha and R.S.D Wahidabanu [7], proposed level set-Active contour model as a method for detecting the tumour.G.Bhat et al, Proposed artificial neural network-based system for detecting and classifying the lung nodule as benign and malignant[8]. N.A. Memon et al proposed the system for thresholding method which is select the threshold value based on background pixel of an object [9]. A.M Yametkar and R.D.Patane et al.[10] used Bayesian classifier for lung nodule detection as cancerous and non-cancerous.

III. PROPOSED ALGORITHM

The main aim of this proposed system is detecting lung pulmonary nodule with better accuracy. In this system, the methods which are focussing automatic detection and classifying nodule as benign or malignant. Here the whole system of pulmonary nodule classification into the following steps.1. Image Acquisition 2. Image pre-processing 3.Segmentation 4.Thresholding 5.Feature Extraction 6.Neural Network classification.

a. Image Acquisition

Image acquisition is characteristics of an object represented by digitally encoded which is used to describe the internal part of an object. In the medical field, computed tomography is an imaging procedure which is done by scanning inside the human body throw some machine. In this work, the system has been collected more number of CT images from various sources which are having both types of images such as benign and malignant.

b. Image Pre-processing

i. Image Conversion (RGB to grayscale)

Image conversion is converting an image from RGB to greyscale by using rgb2gray Matlab function. It has been done by eliminating the hue and saturation while retentive luminance.

ii. Normalization

Normalization is used to resize the image by using Matlab function *'Imresize'*. After we receiving input image which is resizing the image with the value of 256 x 256 pixels.

iii. Adding Noise

This function Imnoise used to adds zero-means, Gaussian white noise with the variance of 0.01 to a grayscale image I.

iv. Image Restore

This function used to filter the grayscale image with the 2-d adaptive pixel-wise low-pass wiener filter. The Wiener filter can reduce the mean squared error of the estimated random process and desired random process. The Matlab function *wieiner2* is used here to removing noise for noise-free images.

v. Edge detection

Edge is represented as the location in the image is sudden changes in their contrast level at each pixel. There are two types of methods that are 1.gradient 2.Laplacian. Here in the proposed system, the Sobel edge detection method is used which is belongs to gradient-based edge detection method. The Sobel method presents how smoothly the image changes in each pixel by calculating the gradient of image intensity at each pixel within the image. ThSobellab function *edge(--,' Sobel')* is used as an edge detection algorithm. So in Figure. 3 some outputs for above the following pre-processing techniques given.

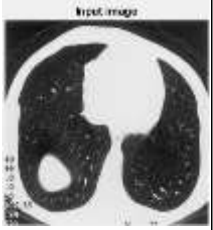
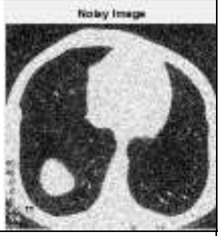
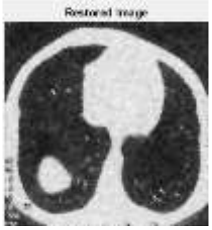
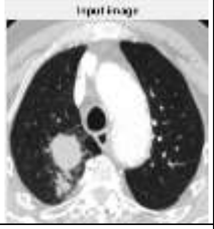
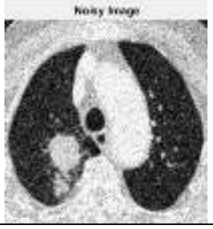
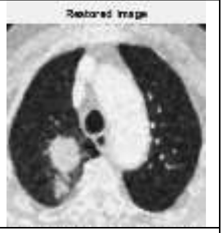
Image(a)			
	Input image	Noisy image	Restored image
Image(b)			
	Input image	Noisy image	Restored image

Figure 3: Pre-processing outputs of images (a) and (b)

c. Thresholding

Thresholding is a very important and powerful tool for segmenting the image. The main usage of this threshold is it's having some advantage like less number of storage space, high speed, ease of manipulation and compared with a greyscale image which usually contains 256 levels[11]. Here the global threshold of Otsu's method is using with the Matlab keyword *Greythresh* which is used to convert the greyscale image into a binary image. Therefore the two-level of pixels that are allocated to above or below the specified threshold. The two values of 0 and 1[12].

$$g(x,y) = \begin{cases} 1 & f(x,y) > T \\ 0 & f(x,y) \leq T \end{cases} \qquad (1)$$

Where g(x, y) = Output image

f(x, y) = Input image

T = Threshold value

Thresholding is always based on two values 0 and 1 that is a black and white pixel. Here the pixel is equal to within-class variance is maximization between-class variance[13].Otsu's method based on the frequency and mean value, the equation to be given below. Following formulas are representing frequency and mean value calculation while thresholding.

1.Frequency:

$$w = \sum_{i=0}^{T} P(i), \quad P(i) = \frac{ni}{N} \tag{2}$$

Where

N= representing total number of pixel

N_i = number of pixels in level i

2.Mean:

$$\mu = \sum_{i=0}^{T} iP(i)/\omega \tag{3}$$

The variation of the mean value for each class from the intensity mean of all pixel:

Between classes-variance σb^2

$$\sigma_b^2 = \omega_0 \, (\mu_0 - \mu_t)^2 + (\mu_0 - \mu_t)^2, \tag{4}$$

Substituting, $\mu_{t=\omega_0}\mu_0 + \omega_1\mu_1,$ (5)

$$\sigma_b^2 = \omega_0 \, \omega_1 (\mu_1 - \mu_0)2 \tag{6}$$

$\omega_0 \omega_1 \mu_0 \mu_1$ are the values of frequency and mean values. So the threshold has been achieved through this type of Otsu's method is very effective which made segmentation clearly [14].

Below Figure.4 input images shows the output image of greythresh thresholding or Otsu's thresholding.

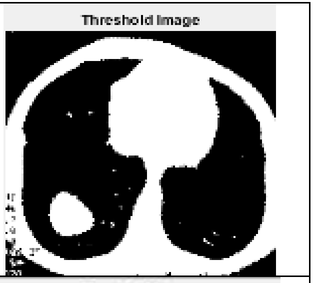
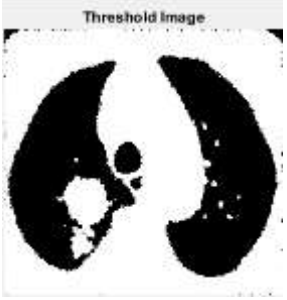
Image(a)	
Image(b)	

Figure 4. Threshold image of (a) and (b)

d. Morphological Operation

The unwanted portion of an image can be eliminated through the structuring element which is an important portion of the morphological operation of dilation and erosion. The Matlab function *Strel* has been used to remove the unnecessary pixel of a black pixel in an image with the help of *origin*. The *origin* is the center pixel of the structuring element, which is used to identifying the pixel when the image being processed. So the black(0) pixel is removed from the original image for further processing

In this work, the morphological operations are performed for better clarity in an image. The morphological means based on the size and shape of the input image that can be processed. There are two types of operations are performed which is dilation and erosion. The uses of dilation can adds pixel to the boundaries of an object in an image and erosion to remove the pixels from boundaries of an object in an image. A structuring element called a mask is used in a morphological operation. The dilation and erosion are '*thin*' and '*thicken*' the boundaries of the object in an image through adding and removing a pixel to an image boundary. The formal definition of a dilation using two sets A and B are given below.

$$A \oplus B = \{z \mid (\hat{B})_z \cap A \neq \emptyset\} \qquad (7)$$

$\hat{B}$ is reflecting B that means dilation of A is performed by B through this reflection of B. Then using z shifting the B over A.This displacement became overlap of A and B at least one element. This process gives the dilation result and B known as the structuring element.

$$A\theta B = \{z \mid (B)_z \cap A^c \neq \emptyset\} \qquad (8)$$

So the morphological operation on the thresholded image makes to remove the unwanted parts of image background[15]. Here the output image of the morphological operation of input images (a) and (b) is given below in Figure 5.

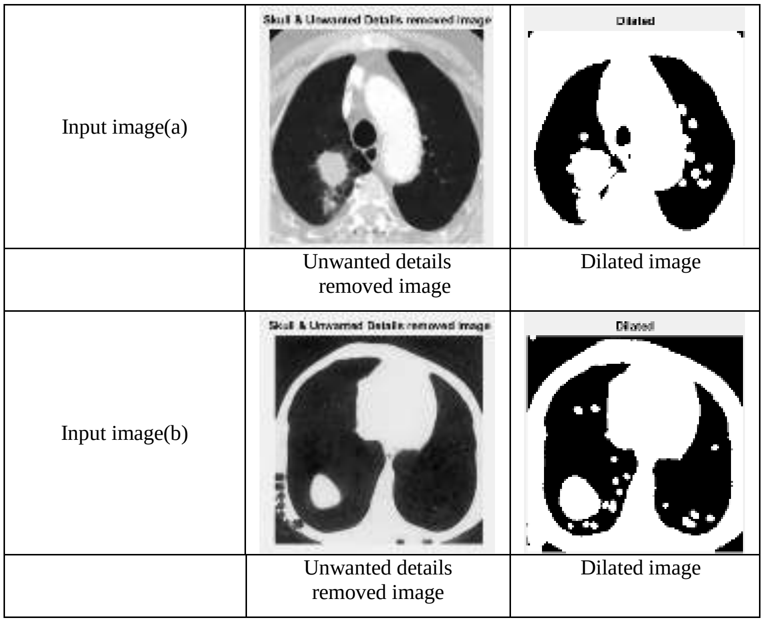

Input image(a)
Skull & Unwanted Details removed image
Dilated
Unwanted details removed image
Dilated image
Input image(b)
Skull & Unwanted Details removed image
Dilated
Unwanted details removed image
Dilated image

Figure 5. Morphological operations of (a) and (b)

e. Segmentation

Segmentation is partitioning an image into multiple regions containing a pixel with similar characteristics. The main usage of segmentation is if u want to analyse something in inside an image is easier through this image segmentation process. Mainly it used to locate an object and boundary of image such as lines, curves and assigning a label for every pixel in an image. The pixel with same label share some characters and collection of a region of pixels are identified by mask or image label. By dividing an image into several regions which can be used to analysing a particular object in a whole image. So using of segmentation is only the interesting region has been used for analysing something instead of using an entire image. Image segmentation in medical images gives the result about affected region or area of particular imaging type which is made helpful to diagnosis a disease at their earlier stage. The approach of the segmentation process is edge-based and region-based. So the edges that define region while the discontinuity in pixel value. Another one is similarities in the region in an image, so some techniques are following these approaches are region growing, thresholding, clustering.

In the proposed system the C-means clustering is used for segmenting an image into the region. In the clustering process, the various techniques are available to find a region of interest (ROI). The C-means clustering gives better result in clustering are grouping the interested region in an image for segmentation. In c-cluster, the data points and clusters centers are considered for calculation. It works by assigning the membership to data points to corresponding cluster center.C-means gives a better result than K-Cluster in the segmentation process.[16][17].

Algorithm for C-clusters

1. First, select the cluster 'C' randomly
2. Calculate the membership.
3. Compute the cluster center
4. Finally, repeat step 2 and 3 until the value of the objective function is minimum

In Figure. 6 describe the output of the C-means cluster of images (a) and (b).

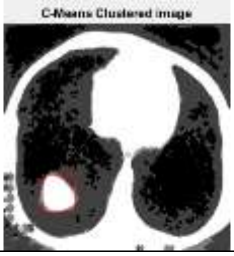
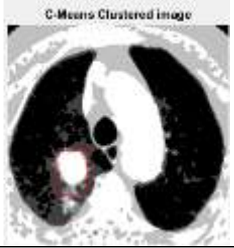
Input image(a)	
	C-means clustered Image(a)
Input image(b)	
	C-means clustered Image(b)

Figure 6. Segmented image of (a) and (b)

f. Feature Extraction

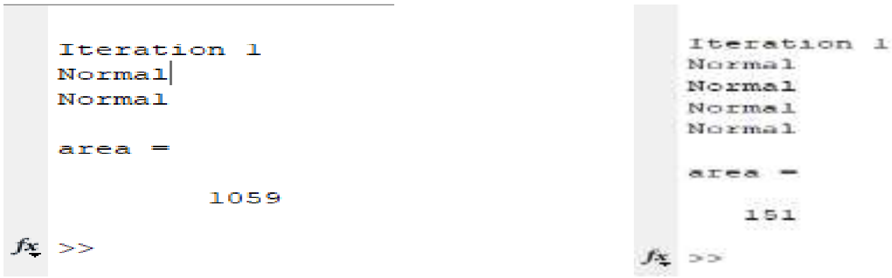
Feature extraction has an important role in recognizing the correct or true pulmonary nodule as benign or malignant. The feature extraction means that some kind of information or properties which are extracted from that particular segmented image. Always the good features can be used to identify the true positive nodule accurately.

When we obtaining an initial location of pulmonary nodule then we calculate some features with candidate nodule. Here basically the lung nodule is circular but we want to compare with blood vessels in the lung. Because the nodule and blood vessels are similar grey level value in the lung. So the roundness of nodule has been calculated through one formula given below.

$$F = \frac{4\pi S}{L^2} \quad (9)$$

Here S is the area of the region and L is the perimeter which is one type feature extracted from a segmented image, and the region of interest is close to the shape of circular when F close to 1. So the region is more circular than the blood vessels (slender structure). Depending on this formula the roundness is always should less than the T (Threshold) for lung nodules[18].

In this proposed system the Regionprops function is used for this feature extraction process in which 21 features are extracted only some kind of features like area, perimeter, eccentricity and nodule diameter are used as features to classify the lung nodule. Here the nodule classification is done the feature called area which is extracted from the segmented image through the Regionprops function. Depends on this area calculation that the nodule has been described as normal or abnormal. In the proposed system the two input images are used to classify as benign and malignant which are (a) and (b) and their result of area calculation is given below in Figure:7.



```
Iteration 1
Normal
Normal

area =

        1059

>>
```

```
Iteration 1
Normal
Normal
Normal
Normal

area =

   151

>>
```

Figure 7.Area calculation of (a) and (b)

g. Classification

In the classification phase, the extracted features are used as an input for classifying the correct nodule. The process of classification is classifying the lung nodule as benign or malignant depending on the values of the features. The classification methods are two types which are four-type nodule classification and two-type nodule classification. The four-type nodule classification depends on their shape, appearance and location and the two-type of nodule type classification is depends on benign and malignant[19]. In the proposed system the lung nodule classification using the feed-forward neural network of artificial neural network. The artificial neural network is a collection of the mathematical structure looking like a real structure of the human brain[20].

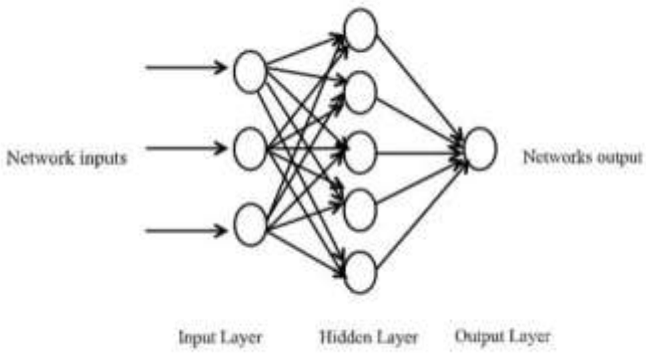


Figure 8. Architecture of ANN

Normally the neural network has three layers which are input layer and hidden layer then the output layer. The input features are flow only in the forward direction such as from the input layer to hidden layer and hidden layer to output layer. The inputs can be multiplied with weight and the bias value which is performed by the sigmoid transfer function. Here the input layer considers the twenty features and one output layer with 10 hidden layers.

After creating and training the network with known training CT images and unknown testing CT images. Through these details, the training is started with means squared error and epoch. In this proposed system the 7000 epochs are defined which is stopped the training when the time is reached this epoch or it automatically stopped when the mean squared error reaches zero. The classification based on the affected area of lung region which is done by this feed-forward neural network with a better result. The test image and their classification result are in Figure: 9 which describes the input images are either benign or malignant, it should specify the name called class1 and class2.

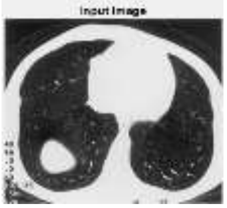
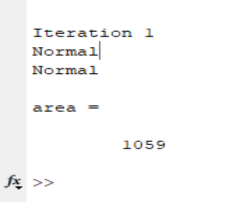
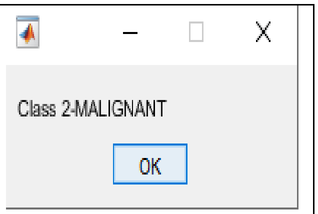
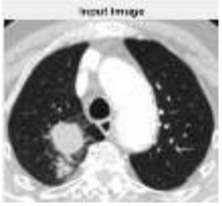
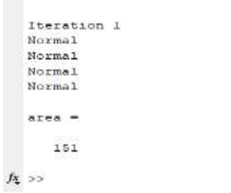
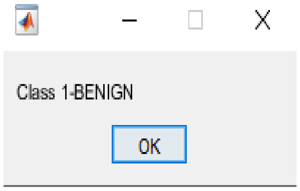
Input image(a)		Iteration 1 Normal Normal area = 1059 >>	Class 2-MALIGNANT OK
	Input image	Area calculation	Classification
Input image(b)		Iteration 1 Normal Normal Normal Normal area = 151 >>	Class 1-BENIGN OK
	Input image	Area calculation	Classification

Figure 9. Classification of normal and abnormal lung images

IV. DISCUSSION

First, before the pre-processing the input image has been capturing and performing some action for further processing. Here the Matlab function imread has been used to get the input image which is resizing using imresize command. In the pre-processing step, the noise has been removed through the Weiner filter which is called adaptive noise removal filtering. It returns the estimated value of additive noise before doing the filtering which is used to removing noise. Here the additive noise means the Gaussian white noise which is the power of noise image and it has been calculated with the variance. Normally the Weiner filter is used to smoothing the image based on the variance and it works better when the noise is gaussian white noise. It is always better than linear filtering based on noise removal.

In the segmentation step, the image can be threshold using the greythresh command in Matlab. The greythresh belongs to Otsu's thresholding technique which is used to segmenting the objects from an image. During this process, the image should be in a binary image, for this the greyscale image is converted into a binary image. After segmentation, the morphological operations are performed by using erosion and dilation. The main purpose of this morphological operation is removing the unwanted portion from an image and separating the needful parts from an entire image. In feature extraction, the features can be extracted and calculated for classifying the nodule. In the classification step, the neural network can be used to classify the nodule by using some extracted features from a previous step. Here the features are trained and tested which is used to correct the classification of a lung nodule.

Finally, the system performance has been calculated by using a formula

$$\text{Correct classification}= \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \qquad (10)$$

$$\text{Incorrect classification}= \frac{FP+FN}{TP+TN+FP+FN} \times 100\% \qquad (11)$$

TP=True Positive, TN=True Negative, FP=False Positive, FN=False Negative

Depends on these four categories of classification result only the performance can be calculated and diagnosing the patient. Through this classification, the nodule can classify correctly as benign or malignant which is used by the doctors to diagnose the patients at an earlier stage and increase their survival rate.

The proposed system has been used very effective thresholding, segmentation techniques and powerful feature extraction techniques. Compare to other existing systems the proposed system gives better result in classifying the lung nodule. The proposed system gives an accuracy of 96.35%.

IV. CONCLUSION

Lung cancer is the most dangerous than other cancer types among people. By earlier diagnosis, it increases the survival rate of lung cancer patients. Here the artificial neural network has been used to classify the lung nodule as benign or malignant which is very helpful to the doctors in identifying the nodule at their early stage. Initially, the pre-processing and segmentation techniques are performed and after segmentation, the features have been extracted and calculated for the classification process. In classification feed-forward, the neural network of ANN techniques has been used to classifying the lung nodule. The performance of the proposed system obtains the result with a success rate of 96.35%. This technique can play a very important role to detect lung cancer and through this technique to reduce the percentage of lung cancer in humans. Therefore its time consuming and not expensive too.

REFERENCES

[1] Yu T[1],Zhong D[1], " Clinical Development of Immunotherapy for Small Cell Lung Cancer" Dec 20;21(12):918-923,2018.

[2] Livingston RB, Moore TN, Heilbrun L, et al, "Small-cell carcinoma of the lung: combined chemotherapy and radiation: a Southwest Oncology Group study", Ann Intern Med ;88:194-9,1978 [PubMed] [Google Scholar]

[3] Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM., "Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008" Int J Cancer ;127:2893-917,2010.

[4] Ganesh B, Sushama S, Monika S, Suvarna P, "A Case-control Study of Risk Factors for Lung Cancer in Mumbai, India", Asian Pac J Cancer Prev;12:357-62. 2011 [PUBMED]

[5] The international early lung cancer action program investigators, "Survival of patients with stage I lung cancer detected on CT screening,", N Engl J Med., 355, pp. 1763-1771, 2006.

[6]. Prionjit Sarker1, Md. Maruf Hossain Shuvo2, Zakir Hossain3, and Sabbir Hasan4, "Segmentation and Classification of Lung Tumor 3D CT Image using K-means Clustering Algorithm" 4th International Conference on Advances in Electrical Engineering (ICAEE), Dhaka, Bangladeshakshmi, 28-30, September-2017..

[7] A.Amutha and R.S.D Wahidabanu, "A Novel Method for Lung Tumor Diagnosis and Segmentation using Level Set- Active Contour Modelling", European Journal of Scientific Research, Vol.90, No.2, pp.175-187, November 2012.

[8] G.Bhat, V.G Biradar, H.S.Nalini, "Artificial Neural Network based cancer cell classification(ANN=C3)", Computer Engineering and Intelligent System, Vol. 3, No.2, 2012.

[9] N.A. Memon et. al, "Segmentation of Lungs from CT Scan Images for Early Diagnosis of Lung Cancer," World Academy of Science, Engineering and Technology. 2006.

[10] A.M Yametkar and R.D.Patane, "Lung Cancer Detection and Classification by Using Bayesian Classifier", Proceedings of IRF International Conference, pp.7-13 Feb.2014.

[11] Gonzalez R.C., Woods R.E., "Digital Image Processing using MATLAB", Upper Saddle River, NJ Prentice Hall, 2008.

[12] 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.

[13] Huang Q., Gao W., Cai W., "Thresholding technique with adaptive window selection for uneven lighting image", Pattern Recognition Letters, Elsevier, p. 801-808,2004.

[14] Khin Mya Mya Tun, Aung Soe Khaing, "Feature Extraction and Classification of Lung Cancer Nodule using Image Processing Techniques" International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 3 Issue 3, March – 2014.

[15] Sudha.V, Jayashree.P.," Lung Nodule Detection in CT Images Using Thresholding and Morphological Operations", International Journal of Emerging Science and Engineering (IJESE) ISSN: 2319– 6378, Volume-1, Issue-2, December 2012

[16] Shengdong Nie, Lihong Li, Yuanjun Wang."A Segmentation Method for Subsolid Pulmonary Nodules Based on Fuzzy C-means Clustering" 2012 5th International Conference on BioMedical Engineering and Informatics (BMEI 2012).

[17] Kan Chen, Bin Li, Lian-Fang Tian, Jing Zhang, "Segmentation of Pulmonary Nodules Using Fuzzy Clustering Based on Coefficient of Curvature", Seventh International Conference on Image and Graphics,2013.

[18] Aggarwal, T., Furqan, A., & Kalra, K,"Feature extraction and LDA based classification of lung nodules in chest CT scan images" International Conference on Advances in Computing, Communications and Informatics (ICACCI). DOI:10.1109/icacci.2015.7275773.

[19] Xinqi Wang,[1] Keming Mao,[1,*] Lizhe Wang,[2] Peiyi Yang,[1] Duo Lu,[1] and Ping He[3] ,"An Appraisal of Lung Nodules Automatic Classification Algorithms for CT Images" journal List Sensors (Basel), v.19(1); PMC6338921.2019 Jan

[20] Mark Hudson Beale., Martin T. Hagan., Howard B. Demuth, "Neural Network Toolbox™ User"s Guide", R2014a Matlab, MathWorks, Inc.