|| Volume 6 || Issue 1 || January 2021 || ISSN (Online) 2456-0774
INTERNATIONAL JOURNAL OF ADVANCE SCIENTIFIC RESEARCH
AND ENGINEERING TRENDS

LUNG CANCER DETECTION USING MACHINE LEARNING APPROACH

Smita Raut[1], Shraddha Patil[2], Gopichand Shelke[3]

[1,2,3] Student, Department of Electronics and Telecommunication Engineering, AISSMS Institute of Information Technology, Kennedy Road Pune, India

4Assistant Professor, Department of Electronics and Telecommunication AISSMS Institute of Information Technology, Kennedy Road Pune, India

-- ***---

Abstract:- The cancer detection is doing with the aid of the skilled expert docs and earlier tiers it may be helpful. The opportunity of human error must be there. It produces the probability of error in the lung cancer detection which necessitate an automatic manner. Afterwards, the paperaims at early detection of cancer through an automatic procedure to decrease human error and making the system greater accurate and error free. In this system we use digital image processing and machine learning algorithm to discover the tumor in the images. Specially there are steps detection manner is performed one is digital image processing and other is machine learning algorithm. In digital image processing image acquisition, grey scale conversion, noise reduction, binarization of picture, segmentation, characteristic extraction, machine studying and the remaining step is most cancers mobile identification. In second step machine learning set of rules this is C 4.5 is used.

Keywords—*Feature Extraction, Adaptive Thresholding, Matching, Multi-Label Classification, CT (computed Tomography), Image Processing (Canny Edge Detection), Machine Learning C4.5 RDMS*

-- ***---

I INTRODUCTION

Cancer is the disease in which cells in the body grows out of control. When cancer stats in the lungs it is called as lung cancer. Lung cancer is the leading cause of cancer death and second most diagnosed cancer in both men and women in United States. Ciggrate smoking is the number one cause of cancer. Lung cancer can also be caused by tobacco, breathing second-hand smoke being exposed to substances such as asbestos or radon at work. There are types of lung cancer and this cancer can be diagonised by doctors with their procedure and to reduce the human efforts or human error for which we have developed a code in which we take the CT scan image and we define the properties and through the various algorithms we can able to detect the image is cancerous or not.In this world not only men but women also suffering from the same dangerous disease. After the detection, the life span of the patient suffering from the lung cancer is very less. If the CT scans have taken in the form of Dicom format, CT scans are taken from studies of 61 patients. Database have 60 images We have proposed a design that reads JPEG converted Dicom Format images of lungs and scans these images for any abnormality through image processing techniques. Once the system has completed the scanning process, it calculates certain features of the abnormality and feeds them into a system which is trained to detect if the abnormality is cancerous. The training system is C 4.5 decision tree machine learning algorithm. The image processing steps include conversion into grayscale, Histogram Equalization, Thresholding and Feature extraction. The machine learning algorithm is trained using 50 images. The output indicates whether the tumor is malignant or benign. Our design was found to be 78% accurate.

We can cure lung cancer, only if you identifying the yearly stage. So here, we use machine learning algorithms to detect the lung cancer. This can be made faster and more accurate. In this study we propose machine learning strategies to improve cancer characterization.

II LITERATURE REVIEW

Software's which are developed and designed are not accessible to any normal patient or else they are not free of cost. It is available offline hence it consumes more space to save the dataset of patients hence it creates the space and time complexity and makes the application bulky.

CNN is a class of deep neural network, but it is done only with the collection of data and it is not labeled. It is most commonly applied to analyze visual imagery. CNN use relatively little pre-processing compared to another image classification algorithm. But it is difficult to get accurate results. Not applicable for multiple images for Lung detection in a short time.

III IMPLEMENTEDMETHOD

System Overview

The figure below shows how the system is going to work, in here first the CT scan image is taken from the website and with the help of DI-COM software. Then the dataset is created from the scraped data and the pre-processing of Data is done on the dataset. After this the datasets are pre-processed by converting grey scale image to binary image and binary image is used to predict the lung cancer. Canny Hash detection is used in this process. These extracted features can be classified using SVM on the basis of area, perimeter and eccentricity.

Area: It is the actual number of pixels present in the cancer image. The defected region represents the number of 1s in the scalar value

Perimeter: It is the actual number of all pixels which are interconnected on the edges of the tumor and it is the sum of all 1 binary bit pixels which are present on the outline of the nodule.

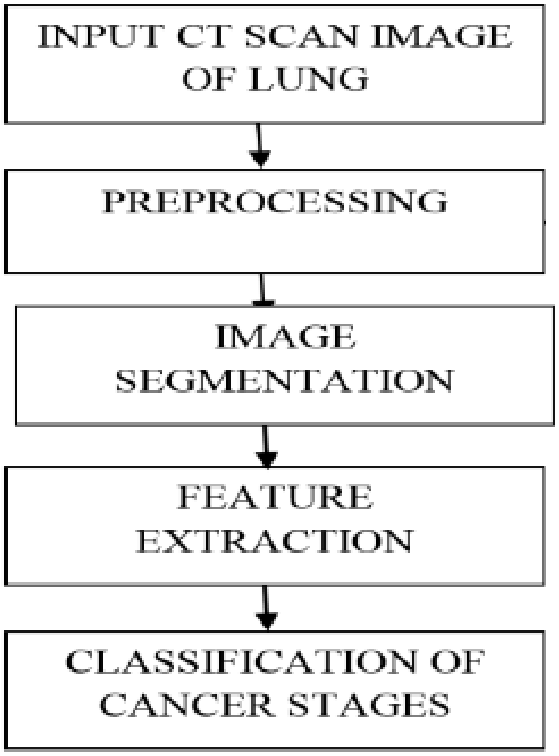
Eccentricity: The roundness or matric value or irregularity index or circularity is to less than one for other shape and one for circular shape.

System Architecture

The above architecture shows the flow of how theprocedure of how the system is going to work and how the interface is built. In the above architecture we can see the different steps that are used for the working of the system and the same are explained below:

1] Pre-processing: In pre-processing, the input CT image is being processed to improve the quality of image. In this some operations are performed on image in which certain details and data of image is enhanced. This enhanced version will contribute in further steps of any robotized system. So, it is beneficial to do some operations of pre-processing.

2]Image Segmentation: Image segmentation is the process in which a digital image is partitioned into multiple segments. in case of images segments corresponds to pixels or super pixels. Segmentation is done is to make the representation of an image into more simplified way or something that is more meaningful and easier to analyze.



3] Data Thresholding: In image processing, Otsu's method is used to automatically perform clustering-based image thresholding. It performs the reduction of a grey level image to a binary image. The algorithm works by assuming that there are two classes of pixels present in image following bi-modal histogram which includes foreground pixels and background pixels, it then computes the optimum threshold value which separates the two classes. It works by storing intensities of pixels in array. Total mean and variances used to calculate threshold value.

In ML C4.5, graythresh () function is used to perform Otsu Thresholding.

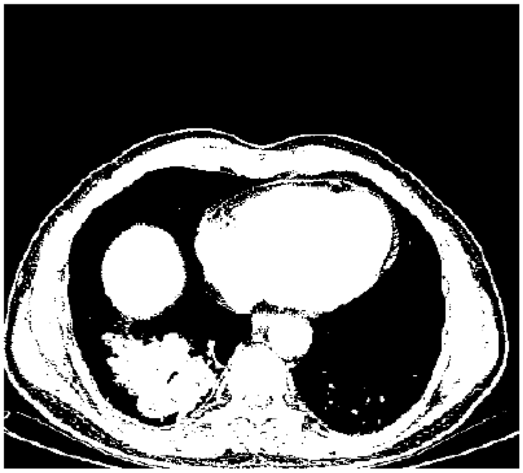
Syntax:

level = graythresh(K);

Above line will create a threshold value which is stored in level.

img = im2bw (I, level);

level is passed to im2bw () function which converts the image into binary.



Thresholding

4] Edge Detection: Sobel filter is used for calculating gradient for edge detection. In IP special('Sobel') is used for sobel filtering.

Syntax:

H=fspecial('Sobel')

This function returns a 3-by-3 filter h that highlights horizontal edges using the smoothing effect by approximating a vertical gradient value. To highlight vertical edges, the filter h' is transposed.

[1 2 1

0 0 0

-1 -2 -1]

5]Feature-extraction:The features that are considered to be extracted in project are as follows: -

1) Perimeter: It is a scalar value that gives the actual number of the outline of the nodule pixel. It is obtained by the summation of the interconnected outline of the registered pixel in the binary image.

2) Area: It is a scalar value that gives the actual number of overall nodule pixel. It is obtained by the summation of areas of pixel in the image that is registered as 1 in the binary image obtained.

3) Eccentricity: It helps us to understand roundness of the object. This matric value or roundness or circularity or irregularity index (I) is to 1 only for circular and it is <1 for any other shape. Here it is assumed that, more circularity of the object. When the object is more circular the value is closer to 1.

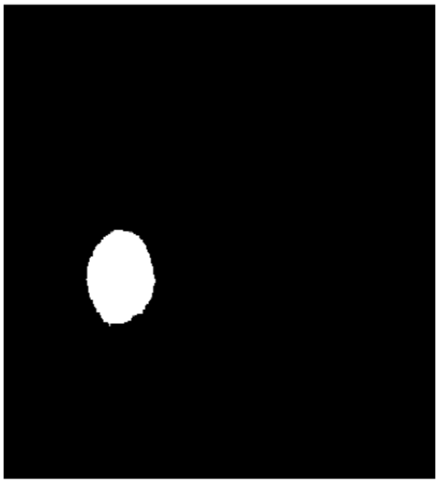
5.1]Grey-Level Co-Occurrence Matrix:A statistical mathematical method of examining feature texture that considers the spatial relationship of pixels in an image is the grey-level co-occurrence matrix (GLCM), also known as the grey-level spatial dependence matrix. The GLCM functions works by finding the texture of a specific image by calculating how frequently pairs of pixels with specific intensity values and in a specified spatial relationship occur in an image, creating a GLCM, and then extracting statistical information from this matrix.

Graycomatrix is a function used in MATLAB for feature extraction.

Syntax:

glcms = graycomatrix (I, Name, Value...)

Above function creates a gray-level co-occurrence matrix (GLCM) from image I.



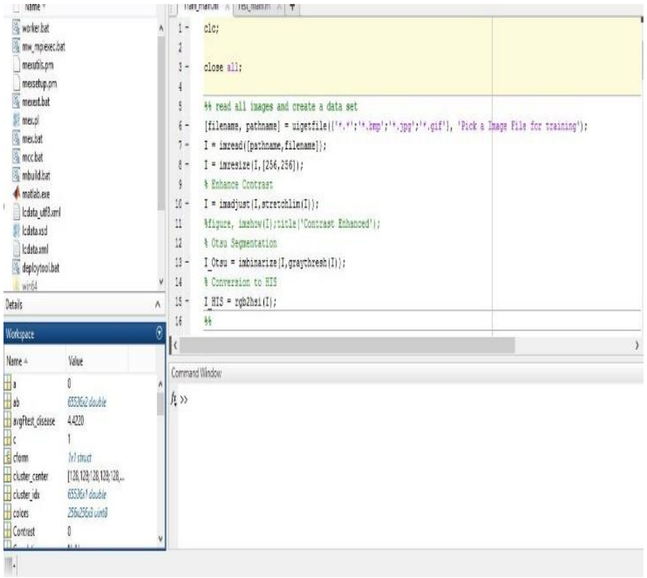
Extracted Image

IV RESULTS AND ANALYSIS

Processing the CT scan image through MATLAB we have analyzed properties through which we can find the difference in between the cancerous image and normal lung image. By going through all the image, we get a difference of all the properties and we have chosen those properties where we can get a maximum difference and from that we come to the result and found that the image is cancerous or not.

This output is divided into two parts training and testing. In training part, we have defined properties and in testing we have tested the image and come to the result of this project.

Training output:



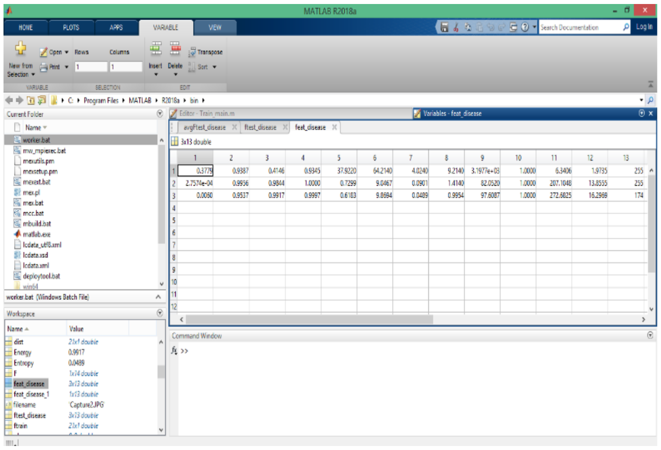
```
clc;

close all;

%% read all images and create a data set
[filename, pathname] = uigetfile({'*.*';'*.bmp';'*.jpg';'*.gif'}, 'Pick a Image File for training');
I = imread([pathname,filename]);
I = imresize(I,[256,256]);
% Enhance Contrast
I = imadjust(I,stretchlim(I));
%figure, imshow(I);title('Contrast Enhanced');
% Otsu Segmentation
I_Otsu = imbinarize(I,graythresh(I));
% Conversion to HIS
I_HIS = rgb2hsi(I);
%%
```

Name	Value
a	0
ab	65536x2 double
avgFtest_disease	4.4220
c	1
cform	1x1 struct
cluster_center	[128,128;128,128;128,...
cluster_idx	65536x1 double
colors	256x256x3 uint8
Contrast	0

In training data code, the properties are defined in a left side corner with the help of this properties we have found the result. Below given image it is a table of thirteen properties which were given in the above image in that left side corner table.

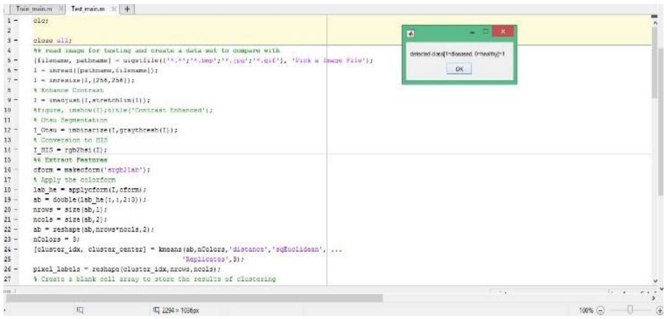


feat_disease — 3x13 double

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0.3779	0.9387	0.4146	0.9345	37.9220	64.2140	4.0240	9.2140	3.1977e+03	1.0000	6.3406	1.9735	255
2	2.7574e-04	0.9956	0.9844	1.0000	0.7299	9.0467	0.0901	1.4140	82.0520	1.0000	207.1048	13.8555	255
3	0.0060	0.9537	0.9917	0.9997	0.6183	9.8694	0.0489	0.9954	97.6087	1.0000	272.6825	16.2969	174

Name	Value
dist	21x1 double
Energy	0.9917
Entropy	0.0489
F	1x14 double
feat_disease	3x13 double
feat_disease_1	1x13 double
filename	'Capture2.JPG'
ftest_disease	3x13 double
ftrain	21x1 double

Testing Output:



```
clc;

close all;
%% read image for testing and create a data set to compare with
[filename, pathname] = uigetfile({'*.*';'*.bmp';'*.jpg';'*.gif'}, 'Pick a Image File');
I = imread([pathname,filename]);
I = imresize(I,[256,256]);
% Enhance Contrast
I = imadjust(I,stretchlim(I));
%figure, imshow(I);title('Contrast Enhanced');
% Otsu Segmentation
I_Otsu = imbinarize(I,graythresh(I));
% Conversion to HIS
I_HIS = rgb2hsi(I);
%% Extract Features
cform = makecform('srgb2lab');
% Apply the colorform
lab_he = applycform(I,cform);
ab = double(lab_he(:,:,2:3));
nrows = size(ab,1);
ncols = size(ab,2);
ab = reshape(ab,nrows*ncols,2);
nColors = 3;
[cluster_idx, cluster_center] = kmeans(ab,nColors,'distance','sqEuclidean', ...
                                      'Replicates',3);
pixel_labels = reshape(cluster_idx,nrows,ncols);
% Create a blank cell array to store the results of clustering
```

detected class(1=diseased, 0=healthy)=1

In Testing we have given a healthy and diseased as two states. If the testing image is cancerous then it will give output as a diseased =1, and if it is non-cancerous image then it gives output as a healthy=1.

V CONCLUSION

Cancer is potentially fatal disease. Detecting cancer is more challenging for doctors. Detection of cancer in its early stages is curable. The main aim of this system to predict the cancer in its early stage so that patient treatment must be on time. By using digital image processing and machine learning we have proposed a system which is automatically detect the cancer cell by using machine learning algorithm. This research shows that application of deep learning has the potential to significantly increase the classification accuracy for the low population, high dimensional lung cancer dataset without requiring any hand-crafted, case specific features.

REFERENCES

[1] Mr. Vijay A. Gajdhane, Prof. Deshpande L.M. “Detection of Lung Cancer Stages on CT scan Images by Using Various Image Processing Techniques” IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 16, Issue 5, Ver. III (Sep – Oct. 2014), PP 28-35 www.iosrjournals.org.

[2] Xinliang Zhu, Jiawen Yao, Xin Luo, Guanghua Xiao, Yang Xie, Adi Gazdar and Junzhou Huang “Lung Cancer Survival Prediction from Pathological Images and Genetic Data - An Integration Study” 978-1-4799-2349-6/16/$31.00 ©2016 IEEE

[3] Syed Moshfeq Salaken, Abbas Khosravi, Amin Khatami, Saeid Nahavandi, Mohammad Anwar Hosen “Lung Cancer Classification Using Deep Learned Features on Low Population Dataset” 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE)

[4] Abbas K. AlZubaidi, Fahad B. Sideseq, Ahmed Faeq, Mena Basil " Computer Aided Diagnosis in Digital Pathology Application: Review and Perspective Approach in Lung Cancer Classification," Annual Conference on New Trends in Information & Communications Technology Applications-(NTICT'2017) 7 - 9 March 2017

[5] Sheenam Rattan, Sumandeep Kaur, Nishu Kansal, Jaspreet Kaur” An optimized Lung Cancer Classification System for Computed Tomography Images” 2017 Fourth International Conference on Image Information Processing (ICIIP)

[6] B.A. Miah and M.A. Yousuf, “Detection of Lung cancer from CT image using Image Processing and Neural network”,2nd International Conference on Electrical Engineering and Information and Communication Technology (ICEEICT), May 2015

[7] S. Singh, Vijay and Y. Singh, “Artificial Neural Network and Cancer Detection” National Conference on Advances in Engineering, Technology &Management (AETM)”, pp.20-24,2015

[8] R. Agarwal, A. Shankhadhar and R.K. Sagar,” Detection of lung cancer using content based medical image retrieval”,5th International Conference on advanced computering and communication technologies,pp.48-52,2015