Lung Cancer Detection Using Image Processing and Machine Learning HealthCare

Wasudeo Rahane
Information Technology Department,
NBN Sinhgad School of Engineering,
Pune, India
wasudeo.rahane@sinhgad.com

Himali Dalvi
Information Technology Department,
NBN Sinhgad School of Engineering,
Pune, India
dalvihimali1997@gmail.com

Yamini Magar
Information Technology Department,
NBN Sinhgad School of Engineering,
Pune, India
magaryamini526@gmail.com

Anjali Kalane
Information Technology Department,
NBN Sinhgad School of Engineering,
Pune, India
anjee9545@gmail.com

Satyajeet Jondhale
Information Technology Department,
NBN Sinhgad School of Engineering,
Pune, India
satyajeetjondhale@gmail.com

***Abstract*— Lung cancer is one of the significant reasons for death among India. Many diagnosis and detection of lungs cancer has been done using various data analysis and classification techniques. Since the cause of lung cancer stay obscure, prevention become impossible, thus early detection of tumor in lungs is the only way to cure lung cancer. Hence, lung cancer detection system using image processing and machine learning is used to classify the presence of lung cancer in a CT- images and blood samples. In spite of CT scan reports are more effective than Mammography, therefore patient CT scan images are categorized in normal and abnormal. The abnormal images are subjected to segmentation to focus on tumor portion. Classification done on features extracted from the images. The efficient method to detect the lung cancer and its stages successfully and also aim to have more accurate results by using SVM and Image Processing techniques.**

Keywords—ROI(Region of Interest),SVM(support Vector Machine), CT(computed Tomography), RDMS (Relational Database Management System, JSP (Java Script Pages)

I.INTRODUCTION

Lungs cancer is one of the leading cause of the death among the world. Every year many of the peoples die because of lung cancer than other types of cancer. 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 diagnosis is done in early stages then the chances of patient survival is more to increase the patient survival rate which is needed to detect cancer as early as possible. Therefore to get the correct and instant result we can apply the modern techniques by using the image processing and machine learning domain. By increasing the quantity of replica used for the procedure, will improve the accuracy. Correct identification and prior prediction of cancer can extend the rate of survival. The previous techniques comprise study of Mammography, Computerized Tomography Scan, Magnetic Resonance Imaging images. The professional physicians identify the disease and determine the stages of cancer by professionalism. The treatment includes some surgical procedures, chemical treatment to kill or halt the replication and stop of cancerous cell, radiotherapy and targeted therapy. These analysis is very long, expensive and part of body affected with pain/arduous. Therefore, to reduce this process by using various image processing algorithm. CT scan images and blood samples are obtain from hospitals. Computerized Tomography reports are less noisy as compared to MRI and X-Ray reports.

II.RELATED WORK

Literature Review:

*Ashwini Rejintal, Aswini N. [1],*A number of research has been undertaken on Lung Cancer Detection and classification by using various Image Processing algorithms using MATLAB as a software. MATLAB is best for designing an algorithm, but MATLAB is low speed, so for implementing your algorithm you must convert that to any object oriented programing

language. So all this process is more time consuming.*Mena Bansil.[2]*,Some of the classification and pattern recognition algorithms which embedded in the CAD system were still suffering from some redundancy criteria for large number of histopathological datasets, but this shortage can be resolved if large and accessible repository with efficient data search engine implemented in the system. *Mohammand Abu Yousuf* .[3-4],Some of the system proposed to detect the lungs cancer by using Artificial Neutral Network(ANN). provide poor accuracy. it is Relatively easy to use and great for complex/abstract problems like image recognition but Increasing accuracy by a few percent can bump up the scale by several magnitudes. *Abbas Khosravi*[4-5],The deep learning training is very slow and when adding new feature tomake it as slower where in comparison the linear model would be much faster to be trained. *Dhanesh D. Lokhande[6],*Magnetic Resonance Imaging(MRI), create highly detailed images of structures in the body. A computed tomography (CT) scan also same as MRI, CT scan is more advanced, powerful x-ray that takes 360-degree images of internal organs, the spine. CT scans show the bones of the spine much better than MRI.

Existing work:
Some of the software are design to detect the lungs cancer but they are not accessible to the normal patient or else they are not free of cost. Because of they are available offline hence it consume more space to store all the dataset hence it increase space complexity and make that application heavy.

III.PROPOSED WORK

In this proposed system, our main aim to detect the lungs cancer and its stages using CT scan images. The user will upload his CT scan report on website and hence the query is generated at the client side and send it to the sever side. Some algorithm are performed on it and generated output is again send it to the client side and guide him about whether the user is having cancer or not? And in which stage?

Following are algorithms of proposed system as shown in Fig.1.

Image acquisition:-
The first step we gone follow in this system is acquiring CT scan images of user. As compared to X-ray and MRI images CT scan replica have less noise therefore CT scan image is taken as input. To get better accuracy and less distortion, these images are used. The input images are acquired from ELCAP Public Lung Image DB which contains near about 200 lung images of both cancerous and noncancerous patient CT scan report. We will also take blood sample as a input with CT Scan images which is the one of the perimeter to check the lungs cancer detection while performing Machine learning algorithm which will be explained further. In the obtained images more noise is observed. Pre-processing of images is used to improve the contradiction and clarity of images. Hence, various techniques such as Grayscale conversion, noise reduction, and binarization techniques are applied to get image in required form.

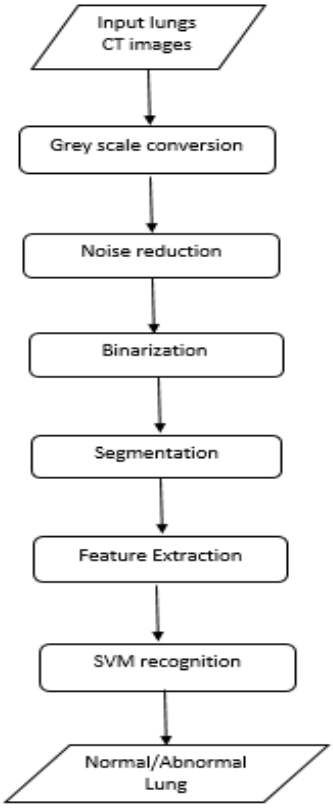
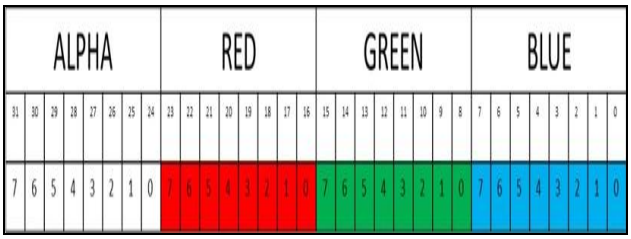


Fig.1 Flow Chart of Algorithms

Grayscale conversion:
In this step an RGB image is converted to Grayscale image. A pixel is made of 4 components such as alpha, red, green, blue .where alpha determine the transparency while red, blue, and green determines the color of pixel. We can represent these 4 components as, A stands for alpha, R stands for Red stands for Green stands for Blue. Each of these components (ARGB) having the values in between 0 to 255;where 0 means the components is absent and 255 means component is fully present. We can represent the value between 0 to 255 using 8-bits. Hence we need 32bit to represents the pixel.

For converting the color image into grayscale we find the average of it; and replace the R,G,B pixel value with Average

[Average:- $\frac{R+G+B}{3}$]



Noise Reduction:
Noise reduction is the mechanism of discarding noise from grayscale image. In Proposed system median filter is used to remove the noise. Therefore median filter will arranged nearby pixels as per the intensity and the following median value turn out to be recent value for the central pixel and it is a non- linear function using in image processing. This filter is more effective whiling reducing noise, maintain edges and it also removing 'salt and pepper' type noise.

Binaization:
Binarization is the process of converting the grayscale image (which is having pixel range from 0 to 255) into binary image (which is having (0, 1)) by a threshold value of 175. The pixel value which is less than 175 are turned to white (0) and pixel value which is greater than 175 are turned to black (1).

Let input image *is p(m,n), T* can be defined as threshold value and the output image is *q(m,n)* of thresholding process therefore mathematically it can be expressed as:- *q(m,n)=1 ,if p(m,n)* $\geq T$ otherwise 0.

Segmentation
Image segmentation act as a part of screening in medical imaging field. Segmentation algorithm divides the replica into multiple meaningful segments. In computerized vision system and recognition, the digital image is divided into many segments. The main objective of segmentation is to make simple and transform the delegation of CT scan replica into more informative and examine it easily in details. Segmentation of Image is used to reduce unnecessary information in the image and locate object, boundaries like lines, curves etc. in images. In the proposed system, segmentation process consist of some steps. Firstly, it transforms the original/real image into edge only image. The transformed edge only image into dilated image and filled image and at last finally both (left and right lungs) are segmented.

Feature Extraction:
Feature Extraction is an essential step/stage that uses algorithms and techniques to recognize the patterns of an replica. The Segmented output are given as input for the feature extraction. The following features are covered under feature extraction such as Area, Perimeter and Eccentricity and these all are scalar quality.

These features are declared as:
Area: Area that gives absolute variety of overall lump pixel element within the extracted ROI. Transformation function generates array of region of interest (ROI) that contains pixels with 255 values.

Area= a = (a l, m , P ROI[Area] = l , Q ROI[Area] = m) where l and m are the pixels among the figures. P ROI[ ] is vector contain ROI x region, Q ROI[ ] is vector contain ROI y region.

Perimeter: Perimeter will gives an actual number of nodule elements which is a scalar quantity. It will tell the length of extracted ROI boundary. The transformation function is created with alteast one pixel which contain 0 value under array of edge that contain pixel with 255 values.

Perimeter = X =(Xl,m , P edge[X] = l, Q edge[X] = m)
Where, P edge [ ] and Q edge[ ] are vector represents the region of the lth and mth pixel respectively forms curve.

Eccentricity: This metrical value for the nodule can be decided if circularity or irregularity portion will be 1 then it is in circular shape and if circularity or irregularity will be not greater than 1 then it is in different shapes.

$$Eccentricity = \frac{Length\ of\ major\ axis}{Length\ of\ minor\ axis}$$

Support Vector Machine:
SVM is a supervised learning model that evaluate data and predicate pattern using Classification technique. SVM classifier divides the texture into two groups or classes i.e., normal and abnormal images. It is used to trace the nodule successfully. SVM is a margin classifier (hyperplane) which separated the two groups that is why it is also known as non-probabilistic binary classifier. Training data points which is nearest to classifier is called Support Vector and hence this maximum classifier is known as Support Vector Machine. The distance between the nodules of cancer and hyperplane is as far as possible. Due to the SVM classification, calculate the Stages of lung cancer.

In the Proposed System, SVM classifier classifies the positive and negative samples of lung cancer images. In this system, SVM classification for lung cancer can be detect, instead of using previous techniques such as decision trees, K-nearest neighbor, LASSO regression, artificial neutral network and random forests. SVM is less time-consuming than other algorithms.

SVM classifier used to classify the linear and non-linear regions. A linear separation, classifier is used to separate the affected and non-affected regions within the Image. It basically uses soft and hard margin. These linear

equations are homogeneous. In non-linear separation, we will separate the affected portion or region by representing non-linear form. These linear equations are homogeneous.

IV. IMPLEMENTATION

To create better access to medical care and to make lungs cancer detection system more strong, the system uses JAVA. It is an open source, object oriented, system independent programing language. Java would be wonderful in sense of script line comparison and the visualization. Java is devise independent hence we can operate the system anywhere. In this proposed system we are developing the user friendly website which will help the user to detect the cancer and its stage with accuracy. For framework development we use the java script pages (JSP) as a scripting language.

Following are the GUI and framework of proposed system.

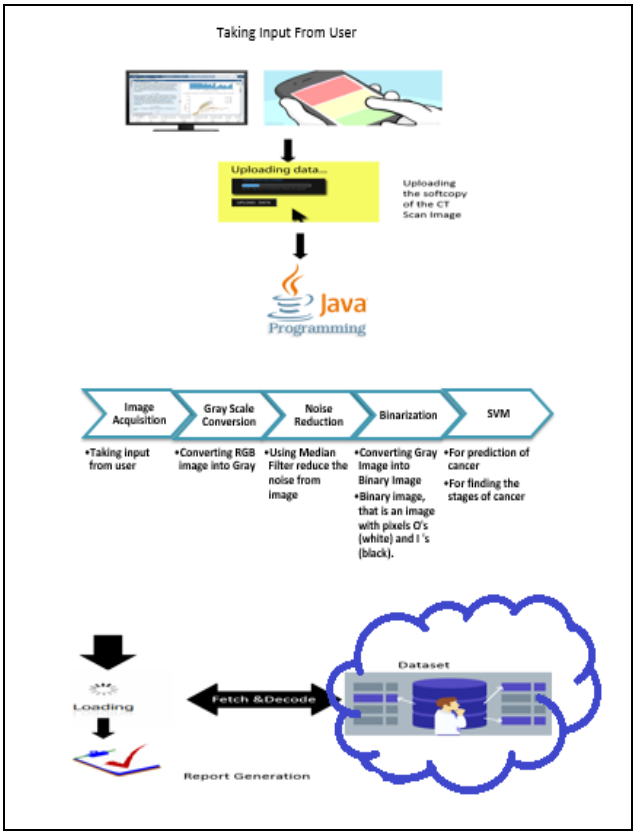


Fig.2 Framework for the Proposed System

JSP is a sever side programing technology which is used for building web-based applications; it allow the dynamic creation of independent framework. JSP is used to access JAVA APIs, which allow JDBC API to access project databases. JSP support dynamic content which will be helpful for the developers to insert java code in HTML pages using of particular JSP tag. Amazon cloud is used to store the lung cancer image dataset and framework will be accessible 24/7 from it.

MySQL is having High Performance, Round-the-clock Uptime, Reduced Total Cost of Ownership, The Flexibility of Open Source because of this advantages we use MySQL as our database to store the images of the CT scan report and blood samples. Amazon internet Services (AWS) is a secure cloud services platform, which providing database storage functionality. We are using an amazon cloud to store the information.

Graphics User Interface

V. CONCLUSION

Lungs cancer is a serious disease that described by unlimited growth of cells in tissues of the lungs. Detection of lungs cancer in earlier stages is very crucial it can save many lives. In our proposed system we are describing the lungs cancer and its stages using different image processing and machine learning algorithms such as, grayscale conversion, noise reduction and binarization. All this algorithms are used for the pre-processing of the given CT scan image. ROI is defined from main CT scan image. For pre-processing stages, median filter and segmentation gives accurate result. From the extracted ROI some features are extracted i.e., Area, Perimeter, and Eccentricity. This characters are helpful for defining the lungs cancer at earlier stages. For grouping purpose Support Vector Machine (SVM) classifier classifies the positive and negative samples of lung cancer images in this system.

FUTURE SCOPE

In Future work, proposed system will help to suggest to diagnose cancer in different organs of human being. Proposed technique can be implemented for specific cancer i.e., group of diseases which help in reducing the growth of abnormal cells or spreading to other parts of body.

REFERENCES

[1]. 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

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

[3]. 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

[4]. A.K. AliZubaidi, F.B Sideseq, A.faeq and M.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),* pp.219-224,March 2017

[5]. S.M. Salaken, A. Khosravi, A. Khatami, S. Nahavandi and M.A. Hosen," Lung Cancer Classification Using Deep Learned Features on Low Population Dataset", *IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE),*2017

[6]. A.M Rossetto and W. Zhou," Deep Learning for Categorization of Lung Cancer CT Images",*IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*,2017

[7]. A.S.Deshpande,D. D.Lokhande,R.P. Mundhe and J.M.Ghatole,"Lung cancer detection with fusion of CT and MRI Images using Image Processing and Machine Learning", *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET),* Vol.4,pp.763-767 March 2015

[8]. S.P.Patil, R.Sanyal and R.Prasad,"Efficient video coding in region prediction in online video surveillance",*International Conference on Image Processing, Computer Vision, and Pattern Recognition(IPCV),*pp.210-216 Jan 2015

[9]. U.K.Kumar, M.B.Nikhil and K.Sumangali,"Prediction of breast cancer voting classifier technique",*Smart Technologies and Management for computing, Communication,Controls,Energy and Materials(ICSTM),*pp.108-114 Aug2017

[10]. B.J.Nair, K.J Anju and A. Jeevakumar,"Tabacco smoking induced lung cancer prediction by Ic-micrornas secondary structure prediction and target comparison",*2nd International Conference for convergence in Technology(I2CT)*,pp.854-857 April 2017

[11]. S.Muthuselvan, K.S.Sundaram and Prabasheela,"Prediction of breast cancer using classification rule mining techniques in blood test datasets",International Conference on Information Communication and Embedded Systems (ICICES),pp.1-5 Feb 2016

[12]. U.Ojha and S.Goel,"A Study on Prediction of breast cancer recurrence using data mining techniques",7th International Conference on Cloud Computing, Data Science and Engineering- Confluence, pp.527-530 June 2017

[13]. R.Wulandari,R.SigitandS.Wardhana,"",InternationalElectronicsSymposium on Knowledge Creation and Intelligent Computing (IES-KCIC),pp.120-126 2017