JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Lung Cancer Detection using Machine Learning(CNN)

Prof.Junaid Mandviwala, Abdullah Khan, Sania Khan
junaid@eng.rizvi.edu.in , abdullah@eng.rizvi.edu.in, saniak@eng.rizvi.edu.in
Department of Computer Engineering,
Rizvi College of Engineering, Mumbai, India

Abstract : The Automatic defects detection in CT images is very important in many diagnostic and therapeutic applications. Because of high quantity data in CT images and blurred boundaries, tumor segmentation and classification is very hard. This work has introduced one automatic lung cancer detection method to increase the accuracy and yield and decrease the diagnosis time. The goal is classifying the tissues to three classes of normal, benign and malignant. In MR images, the amount of data is too much for manual interpretation and analysis. During past few years, lung cancer detection in CT has become an emergent research area in the field of medical imaging system. Accurate detection of size and location of lung cancer plays a vital role in the diagnosis of lung cancer. The diagnosis method consists of four stages, pre-processing of CT images, feature, extraction, and classification, the features are extracted based on DTCWT and PNN. In the last stage, PNN employed to classify the Normal and abnormal. Machine learning based lung cancer prediction models have been proposed to assist clinicians in managing incidental or screen detected indeterminate pulmonary nodules. Such systems may be able to reduce variability in nodule classification, improve decision making and ultimately reduce the number of benign nodules that are needlessly followed or worked-up.

IndexTerms -CNN, PNN, Pulmonary modules, Lung cancer, DTCWT, Theurapeutic**.**

I. INTRODUCTION

Cancer is a group of diseases characterized by the uncontrolled growth and spread of abnormal cells. If the spread is not controlled, it can result in death. Lung cancer was the most common cancer in worldwide, contributing 2,093,876 of the total number of new cases diagnosed in 2018. 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 SURVEY

[1]T. Sowmiya, M. Gopi, M. New Begin, L.Thomas Robinson - In this paper they described Cancer as the most dangerous diseases in the world. Lung cancer is one of the most dangerous cancer types in the world. These diseases can spread worldwide by uncontrolled cell growth in the tissues of the lung. Early detection of the cancer can save the life and survivability of the patients who affected by this diseases. In this paper we survey several aspects of data mining procedures which are used for lung cancer prediction for the patients. Data mining concepts is useful in lung cancer classification. We also reviewed the aspects of ant colony optimization (ACO)

technique in data mining. Ant colony optimization helps in increasing or decreasing the disease prediction value of the diseases. This case study assorted data mining and ant colony optimization techniques for appropriate rule generation and classifications on diseases, which pilot to exact Lung cancer classifications. In additionally to, it provides basic framework for further improvement in medical diagnosis on lung cancer.

[2]Ada[1], Rajneet Kaur[2] (2013) - In this paper uses a computational procedure that sort the images into groups according to their similarities. In this paper Histogram Equalization is used for preprocessing of the images and feature extraction process and neural network classifier to check the state of a patient in its early stage whether it is normal or abnormal. After that we predict the survival rate of a patient by extracted features. Experimental analysis is made with dataset to evaluate the performance of the different classifiers. The performance is based on the correct and incorrect classification of the classifier. In this paper Neural Network Algorithm is implemented using open source and its performance is compared to other classification algorithms. It shows the best results with highest TP Rate and lowest FP Rate and in case of correctly classification, it gives the 96.04% result as compare to other classifiers.

3]Dasu Vaman Ravi Prasad (2013) - In this paper image quality and accuracy is the core factors of this research, image quality assessment as well as improvement are depending on the enhancement stage where low pre-processing techniques is used based on Gabor filter within Gaussian rules. Following the segmentation principles, an enhanced region of the object of interest that is used as a basic foundation of feature extraction is obtained. Relying on general features, a normality comparison is made. In this research, the main detected features for accurate images comparison are pixels percentage and masklabeling.

. [4]S Vishukumar K. Patela and Pavan Shrivastavab (2012) - In this paper authors mostly focus on significant improvement in contrast of masses along with the suppression of background tissues is obtained by tuning the parameters of the proposed transformation function in the specified range. The manual analysis of the sputum samples is time consuming, inaccurate and requires intensive trained person to avoid diagnostic errors. The segmentation results will be used as a base for a Computer Aided Diagnosis (CAD) system for early detection of cancer, which improves the chances of survival for the patient. In this paper, authors proposed gabor filter for enhancement of medical images. It is a very good enhancement tool for medical images.

[5]Fatma Taher1,*, Naoufel Werghi1, Hussain Al-Ahmad1, Rachid Sammouda2 (2012) - This paper presents two segmentation methods, Hopfield Neural Network (HNN) and a Fuzzy CMean (FCM) clustering algorithm, for segmenting sputum color images to detect the lung cancer in its early stages. The manual analysis of the sputum samples is time consuming, inaccurate and requires intensive trained person to avoid diagnostic errors. The segmentation results will be used as a base for a Computer Aided Diagnosis (CAD) system for early detection of lung cancer which will improve the chances of survival for the patient. However, the extreme variation in the gray level and the relative contrast among the images make the segmentation results less accurate, thus we applied a thresholding technique as a pre-processing step in all images to extract the nuclei and cytoplasm regions, because most of the quantitative procedures are based on the nuclear feature. The thresholding algorithm succeeded in extracting the nuclei and cytoplasm regions. Moreover, it succeeded in determining the best range of thresholding values. The HNN and FCM methods are designed to classify the image of N pixels among M classes. In this study, we used 1000 sputum color images to test both methods, and HNN has shown a better classification result than FCM, the HNN succeeded in extracting the nuclei and cy-toplasm regions. In this paper authors uses a rule based thresholding classifier as a preprocessing step. The thresh-olding classifier is succeeded in solving the problem of in-tensity variation and in detecting the nuclei and cytoplasm regions, it has the ability to mask all the debris cells and to determine the best rang of threshold values. Overall, the thresholding classifier has achieved a good accuracy of 98% with high value of sensitivity and specificity of 83% and 99% respectively.

III. PROBLEM STATEMENT AND OBJECTIVE

Self Education on Deep Learning A large part of this project contains a lot of self education, initially we new very little about deep learning, as part of this project the we should have a good grasp on deep learning concepts.

User Research and Evaluation The project should have a user-centred design aspect. This means that the system should be designed to help certain users. In this project the user would be medical professionals who work in diagnosing lung cancer.

System: Upload CT Scans The system should be capable of getting CT Scans from Users that will be utilized by the Deep Learning Model.

System: Detection of Lung Cancer The system should be able to detect the lung cancer within the CT scan images that users have uploaded.

System: Display Results The system should be able to give information that our user can appropriately understand and gain insight from it.

IV. SCOPE

Doctor's who work in this field are prone to observer fatigue from viewing so many CT scan images. The research on that suggests that observer fatigue increases the risk of errors that can be made by doctors while analysing these scans. Many images in a CT scan also are irrelevant to Doctors e.g. for 200-300 images only 3 scans would show cancer depending on the stage of the patient. A more efficient machine learning model would be capable of alleviating these additional challenges.

- To decrease the number of rules for testing.
- To reduce the time and cost required for various excessive Medical Tests.
- To increase the accuracy of performance of Lung Cancer Prediction and Detection System.
- Use less number of attributes for prediction of Cancer.
- Early stage detection of cancer.

V.PROPOSED SYSTEM

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 preprocessed 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.
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 preprocessing.

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 bimodal 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.
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,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 oject is more circular that value is closer to 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 greylevel 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.

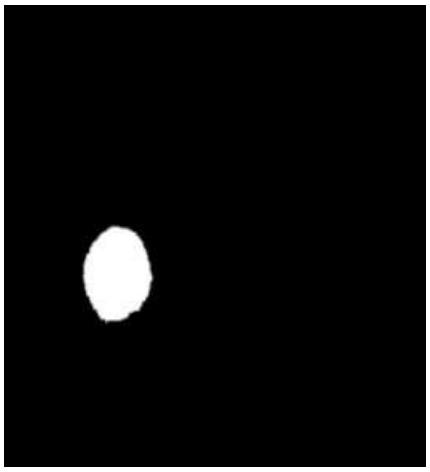

Fig 1. Extracted Image

VI.Tools required

1.Python
2.Anaconda
3.Numpy
4.Pandas
5.Matplotlib
6.Tensorflow
7.Keras
8.OpenCV
9.Floydhub
10.Jupyter Notebook

VII.RESEARCH METHODOLOGY

There have been a number of lung cancer risk models developed and validated that one may consider to be a form of CADx tool (6-9). Typically based on logistic regression, such tools aim to provide an overall risk of the patient having cancer based on patient meta-data such as age, sex and smoking history and nodule characteristics such as nodule size, morphology and growth, if a previous CT was available.

Diagnosis of lung cancer includes the following stages:

1. Images captured
2. Preprocessing of images
3. Image segmentation
4. Feature extraction
5. Principle component analysis
6. Neural network classifier
7. Diagnosis result

1.Images captured or collected : Primarily, cancer and non-cancer patient's data or CT-Scan images will be collected from different diagnostic centers. The digitized images are stored in the DIACOM format with a resolution of 8 bits per plane

2.Preprocessing of images : The image Pre-processing stage in this system begins with image enhancement which aims to improve the interpretability or sensitivity of information included in them to provide better input for other programmed image processing techniques. Image enhancement techniques can be divided into two wide types: Spatial domain methods and frequency domain methods. On the other hand, when image enhancement techniques are used as preprocessing tools for other image processing techniques, the quantifiable measures can determine which techniques are most suitable. In the image enhancement stage we will be using the Histogram

3.Image Segmentation : Image segmentation is a crucial process for most image analysis consequent tasks. Especially, most of the existing techniques for image description and recognition are highly depend on the segmentation results. Segmentation splits the image into its constituent regions or objects. Segmentation of medical images in 2D has many beneficial applications for the medical professional such as: visualization and volume estimation of objects of concern, detection of oddities, tissue quantification and organization and many more. The main objective of segmentation is to simplify and change the representation of the image into something that is more significant and easier to examine. Image segmentation is usually used to trace objects and borders such as lines, curves, etc. in images. More accurately, image segmentation is the process of allocating a label to every pixel in an image such that pixels with the same label share certain pictorial features. The outcome of image segmentation is a set of segments that collectively cover the entire image, or a set of edges extracted from the image i.e. edge detection. In a given region all pixels are similar relating to some distinctive or computed property, such as texture, intensity or color. With respect to the same characteristics adjacent regions are significantly different.

One of two basic properties of intensity values Segmentation algorithms are based on: discontinuity and similarity. In the first group we partition the image based on abrupt changes in intensity, such as edges in an image. The next group is based on segregating the image into regions that are alike according to a predefined criterion. Histogram thresholding methodology comes under this group.

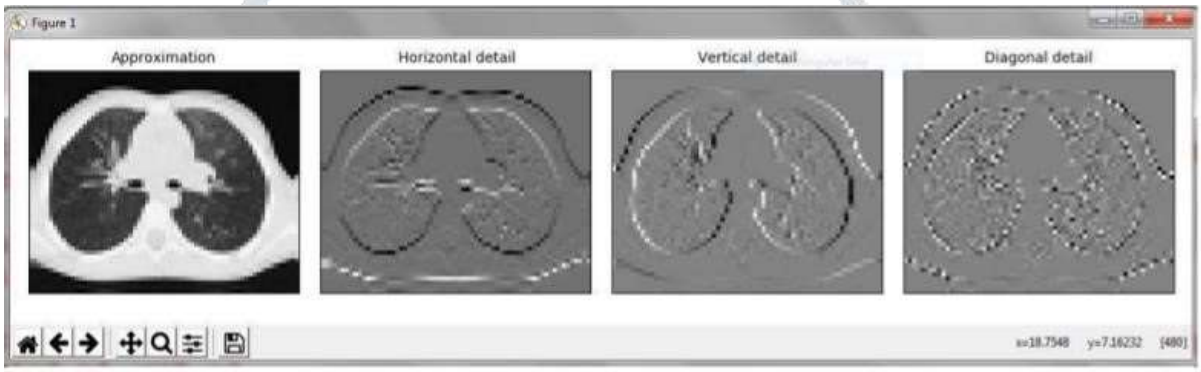

Fig 2.Dual tree complex wavelet Transformation

Feature Extraction: Image features Extraction stage is a crucial stage that uses algorithms and methods to detect and separate various preferred portions or shapes of an inputted image. The following two methods are used to predict the probability of lung cancer presence: binarization and GLCM, both methods are based on facts that
strongly related to lung anatomy and information of lung CT imaging.

A. Binarization Approach For detection of cancer binarization approach has been applied for detection of cancer. In binarization we extract the number of white pixels and check them against some threshold to check the normal and abnormal lung cells. After this process the condition is check whether number of white pixels of a new image is less than the threshold then it indicates that the image is normal, or else if the amount of the white pixels is greater than the threshold, it specifies that the image in abnormal. Merging Binarization and GLCM methods together will lead us to take a decision whether the case is normal or abnormal.

B. GLCM (Grey Level Co-occurrence Method) The GLCM is a process of tabulating different combinations of pixel brightness values called as grey levels which occurs in an image. In this first step is to create gray-level co-occurrence matrix from image in MATLAB. In second step we normalize the GLCM using the following formula Where: i is the row number and J: is the column number From this we compute texture events from the GLCM.

C. Masking Approach Inside lungs masses are appeared as white connected areas inside ROI (lungs), masking approach depends on this. As they increase the percent of cancer presence increases. Also combining Binarization and Masking approaches together will help us to take a decision on whether the case is normal or abnormal according to the mentioned assumptions in the previous two approaches, we can make a conclusion that if image has number of black pixels greater than white pixels then that image is normal or otherwise we can say that the image is abnormal.
PCA (Principle Component Analysis) PCA is a technique to normalize the data in image. Real-world data sets generally display associations among their variables. These associations are frequently linear, or at least practically so, making them agreeable to common examination techniques. One such technique is principal component analysis ("PCA"), which rotates the original data to new coordinates, making the data as "even" as possible. The features mined are delivered to the PCA data mining for better sorting
The following steps takes place in PCA:-

i. Calculating the mean and standard deviation of the features in the image.
ii. Subtracting the sample mean from each observation, and then dividing by the sample standard deviation. This scales and centers the data.
iii. Then we calculate the coefficients of the principal components and their relevant changes are done by finding the Eigen function of the sample covariance matrix.
iv. This matrix holds the coefficients for the principal constituents. The diagonal elements store the modification of the relevant principal constituents. We can mine the diagonal.
v. The maximum variance in data results in maximum information content which is required for better classification

Neural Network Classifier Supervised feed-forward back-propagation neural network ensemble used as a classifier tool. Neural network contrasts in different means from traditional classifiers like Bayesian and k – nearest neighbor classifiers. Linearity of data is one of the major variances. Other existing classifiers like Bayesian and k – nearest neighbor entails linear data to work properly. But neural network works as well for nonlinear data because it is simulated on the reflection of biological neurons and network of neurons.

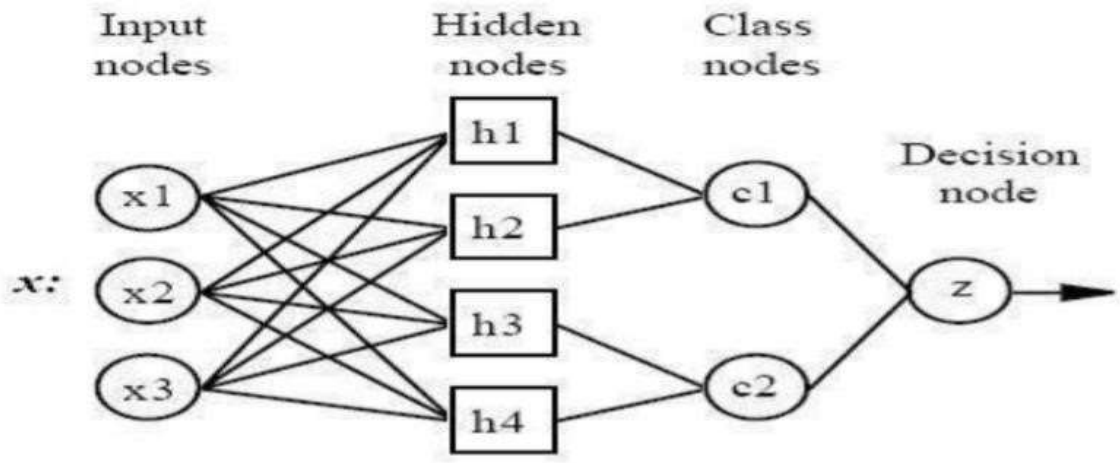

Fig 3. Architecture of neural network

Training the neural network with wide range of input data will increase the detection accuracy, in other words the system will get biased with a small set of data or large set of similar data. Hence neural network classifier needs a large set of data for training and also it is time consuming to train to reach the stable state. But once it is trained it works as fast and quick as biological neural network by transmitting signals as fast as electrical signals.Input layer, internal hidden layer and output layer are the three layers of the architecture of the neural network. The nodes in the input layer are linked with a number of nodes in the internal hidden layer. Each input node connected to each node in the internal hidden layer. The nodes in the internal hidden layer may connect to nodes in another internal hidden layer, or to an output layer. And the output layer consists of one or more response variables. Following are the general Steps performed in Neural Network Classifier:-

Creating feed-forward back propagation network.

Training neural network with the already available samples and the group defined for it.

The input image mined PCA consistent data as the test samples, fires the neural network to check whether the particular selected input sample has cancer or not.

From the results which are obtained from the neural network and the samples trained in network classification rate is calculated using some mathematical formulas. 7.

Diagnosis Result After completion of all the processes in the last stage i.e. in the diagnosis stage or in diagnosis result the proposed system show whether the image is in normal or in abnormal state.

8. Prediction process There is no remedy for cancer after completely affected. Death is inevitable. So the ability to predict Lung cancer plays an important role in the diagnosis process. In this paper we have proposed an effective Lung cancer prediction system . This lung cancer risk prediction system should prove helpful in detection of a person's predisposition for lung cancer. We will be considering various risk factors which includes-age, gender, hereditary, previous health examination, use of antihypersensitive drugs, smoking, food habit, physical activity, obesity, tobacco, genetic Risk, environment, mental trauma, uptake of red meat, balance diet, hypertension, heart disease, excessive alcohol, radiation therapy and chronic lung disease algorithm for the prediction process.

RESULT

We have successfully achieved our goal of implementing a model to detect lung cancer from ct scans using machine learning.The model gives an accurate result with an accuracy of 99%.The model was extensively trained with the dataset that contained cancerous and non-cancerous ct scan images.After learning to distinguish between the images the model was extensively tested with images that contained cancerous and non-cancerous output.We found that the model was able to distinguish between the cancerous and non-cancerous images in minimal time.This model will help doctors distinguish lung cancer in the early stage which is difficult to detect as the size of the tumour is very small and often negotiated as benign in most of the cases which risks the disease to deteriorate and advance in late stages.This could prove costly for the patient and also fatal.The output after implementing the project are

PERFORMANCE ANALYSIS

Accuracy: 0.997

Precision: 1.000

Recall: 0.997

F-Measure: 0.997

Fig 4. Accuracy Table

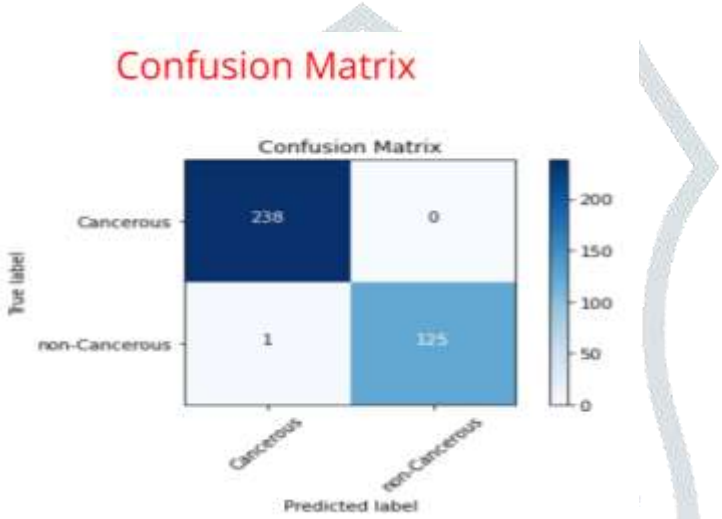

Fig 5. Confusion Matrix

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. When doctors find small nodules (less than 3mm) the current practice suggests that they should wait and rescan in 6-12 weeks to see signs of growth. Depending on the tumor, a tumor can grow up to double its size and evolve to a more advanced form of cancer. It is also important to note that the second most frequent diagnosis is small tumors. The project demonstrates that it would be possible for Doctor's to use deep learning applications to aid their decision making process regarding whether a patient with a small tumor should perform a biopsy or rescan in a few weeks which to a patient could mean early treatment and a better prognosis.

Lung cancer is an extremely complex problem to solve however with early detection a patient has a high increase of survivability. The diagnostics data suggests that the highest frequency of people who get diagnosed have already an advanced form of cancer and the second most frequent cohort are have been accidentally diagnosed and have the cancer that is hardest to determine and is easiest to cure. The research leads the author to believe that deep learning could be a powerful tool in diagnosing very small and very hard to determine nodules to aid medical decision making process.

REFERENCES

[1]T. Sowmiya, M. Gopi, M. New Begin L.Thomas Robinson "Optimization of Lung Cancer using Modern data mining techniques." International Journal of Engineering Research ISSN:2319- 6890)(online),2347-5013(print)VolumeNo.3,Issue No.5, pp : 309-3149(2014)
[2]Ada[1], Rajneet Kaur[2] "Early Detection and Prediction of Lung Cancer Survival using Neural Network Classifier", (IJAIEM)Volume 2, Issue 6, June 2013
[3]Dasu Vaman Ravi Prasad,"Lung cancer detection using image processing techniques", International journal of latest trends in engineering and technology.(2013)
[4]S Vishukumar K. Patela and Pavan Shrivastavab, "Lung A Cancer Classification Using Image Processing", International Journal of Engineering and Innovative Technology Volume 2, Issue 3, September 2012.
[5]Fatma Taher1,*, Naoufel Werghi1, Hussain Al-Ahmad1, Rachid Sammouda2, "Lung Cancer Detection Using Artificial Neural Network and Fuzzy Clustering Methods," American Journal of Biomedical Engineering 2012, 2(3): 136-142
[6]Morphological Operators, CS/BIOEN 4640: "Image Processing Basics", February 23, 2012.
[7]Almas Pathan, Bairu.K.saptalkar, "Detection and Classification of Lung Cancer Using Artificial Neural Network", International Journal on Advanced Computer Engineering and Communication Technology Vol-1 Issue :2011.
[8]American Cancer Society, "Cancer facts & figures2010"
http://www.cancer.org/acs/groups/content/@epidemiologysurveilance/documents/document/a cspc026238.pdf (2010).
[9]Multileve-l Thresholding Based on Histogram Difference," in 17th International Conference on Systems, Signals and Image Processing. 2010.
[10]Nunes, É.d.O. and M.G. Pérez., Nunes, É.d.O. and M.G. Pérez., "Medical Image Segmentation by Multilevel Thresholding Based on Histogram Difference," in17th International Conference on Systems, Signals and Image Processing. 2010.
[11]S.Shah, "Automatic Cell Images segmentation using a Shape-Classification Model", Proceedings of IAPR Conference on Machine vision Applications, pp. 428-432,Tokyo, Japan,(2007)