Journal of Engineering Sciences
Vol 11, Issue 7,July/2020
ISSN NO:0377-9254

EARLY LUNG CANCER DETECTION USING MACHINE LEARNING AND IMAGE PROCESSING

Dr.R.Jothilakshmi
Information Technology Department,
RMD Enfinnering College,
(Affliated to Anna University)
Chennai,India
rjl.it@rmd.ac.in

Ramya.M
Information Technology Department,
RMD Enfinnering College,
(Affliated to Anna University)
Chennai,India
ramyamurali20@gmail.com

Prajwala.G
Information Technology Department,
RMD Engineering College,
(Affliated to Anna University)
Chennai, India
prajwala.praju9@gmail.com

Ramya Geetha.S.V
Information Technology Department,
RMD Engineering College,
(Affliated to Anna University)
Chennai, India
ramyageeth02@gmail.com

Abstract— Lung cancer is a serious disease with the most morbidity and mortality rates of any cancer worldwide. Many diagnosis and detection of lungs cancer has been carried out with the help of various information analysis and classification techn iques. Since the purpose of lung cancers stay obscure, prevention is impossible, hence early detection of tumor in lungs is the most effective way to treat lung cancer. Hence, in lung cancer detection, image processing and machine learning is used to classify the presence of lung cancer in a CT- scans and blood samples. In spite of CT scan, reports are more efficient than Mammography, therefore patient CT scan pictures are categorized in regular and abnormal. The abnormal images are subjected to segmentation to focus on tumor portion. Classification is performed on features extracted from the dataset. The efficient method to detect the lung cancer and its stages successfully and also the goal to have more accurate responses by the use of SVM and Image Processing techniques.

Keywords— Dataset, Lung cancer detection, Mammography, CT-scans, SVM, Image Processing.

I. INTRODUCTION

Lung cancer is the growth of a tumor, referred to as a nodule that arises from cells lining the airways of the respiratory system. These cells are often in bright contrast in chest X-rays and take the shape of a circular object. However, these nodules that can be seen in a chest X-ray may not necessarily be a tumor or lung cancer; it can be due to some other disease such as pneumonia, tuberculosis or calcified granuloma. As such, the detection of lung cancer has been a tiresome task in medical image analysis over the past few decades. If lung nodules can be identified accurately at an early stage, the survival rate of the patients can be increased significantly. According to health industry, chest X-rays are considered to be the most widely used technique for the detection of lung cancer. However, because it is difficult to identify lung nodules using raw chest X-ray images, analysis of such medical images has become a tedious and a much sophisticated task. Cancer is the second primary cause of death worldwide, and is liable for an estimated 9.6 million deaths in 2018. In the entire world, about 1 in 6 deaths is due to cancer. Approximately 70% of deaths from cancer occurs in developing or under-developed countries. About one third of deaths from cancer are due to the five leading behavioral and dietary risks: high body mass index, low fruit and vegetable consumption, lack of physical activity, tobacco use, and alcohol use. Tobacco use is the vital risk factor for cancer and is liable for about 22% of cancer deaths. Cancer affecting infections, such as hepatitis and human papilloma virus (HPV), are liable for about 25% of cancer cases in developing or under-developed countries. Late-prediction and remote diagnosis and its treatment are common. In 2017, only

26% of low-income countries testified having pathology services commonly existing in the public sector. More than 90% of developed countries specified treatment services are obtainable compared to less than 30% of under-developed countries. The fiscal effect of cancer is important and is growing. The total yearly financial cost of cancer in 2010 was valued at around US$ 1.16 trillion. This paper presents a novel technique that can be used to detect lung cancer in its early stage. The first phase starts with taking a collection of CT scans (both normal and abnormal) from the available database. The second phase applies several techniques of image enhancement, to get better level of quality of the images. The third phase contains the technique to get the common features from improved image which provides indicators of normality or abnormality of images. The project has been implemented completely (in matlab) and tested with real CT scan images. The objective is to support efficient image processing feature extraction. Apparently, to accommodate real image data, the image processing tool must own important features like being noise-tolerant, resourceful, practical, and appropriate to use.

II. Related work

A target is choosed automatically in the lung lesion regions we obtained from image processing. The multi-constraints are proposed to control the lesion segmentation. As the intensity of vessels and visceral pleura is close to that of the lung lesion, they are sometimes considered to be part of the adjacent lesions. These tissues are giant obstacles for lesion segmentation. Lung lesion refinement, a lung lesion refining method is used to get rid of the incorrect vascularised regions and other tissues.

Medical image processing has experienced intense growth, and has been an interdisciplinary research field attracting skill from many scientific and engineering fields. Computer-aided diagnostic processing has proved to play a significant role in clinical routine. Complemented by a rush of new development of high technology and use of several imaging modalities, many challenges arise; for example, how to process and analyze a substantial volume of images so that high quality information can be obtained for disease detection and treatment. This classifier training involves collection of a large Number of image data sets and then extraction of a large number of features from each data set .In an imaging research setting, there are typically many variables being investigated, for example, variables in lung CT image acquisition are collimation, tube current, reconstruction algorithm, and breathing state . For each different imaging protocol, there are also many different quantitative features being extracted to search for the optimal combination of imaging parameters and features to characterize the disease process or clinical question to be answered. This requires a complex data model and queries. Once the meaningful variables are selected for use by system to perform a particular diagnostic, the queries become less complex since only those variables need be retrieved. Therefore, an accurate image segmentation method, other than the conventional region of interest analysis, is often needed for diagnostic or prognostic assessment. This functional characterization has a higher potential for proper assessment due to recent advances imaging. Indeed, this higher potential has renewed interest in developing much more accurate segmentation methods to turn hybrid imaging systems into diagnostic tools. Specifically, after the adoption of multi-modal imaging systems, optimal approaches for precise segmentation and quantification of metabolic activities were crucial.

III. Proposed work

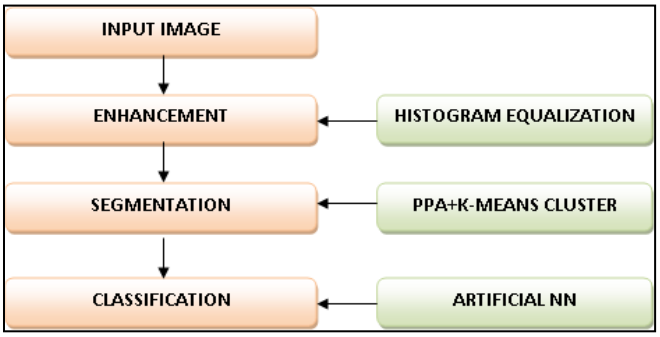
The proposed lung cancer detection and prediction system helps to detect the lung cancer in the early stages and also to predict the stages of lung cancer. Therefore the survival rate of patient will increase significantly.

Objectives of this system are as follows:

- To reduce the number of rules for testing.
- To reduce the time and cost essential for several unnecessary medical tests.
- To intensify the precision of performance of Lung Cancer Prediction and Detection System.
- Use fewer number of attributes for prediction of cancer disease.
- To detect the cancer at its initial stages.
- Increasing the survivability of the patient more than 5 years.

IV. Module Description

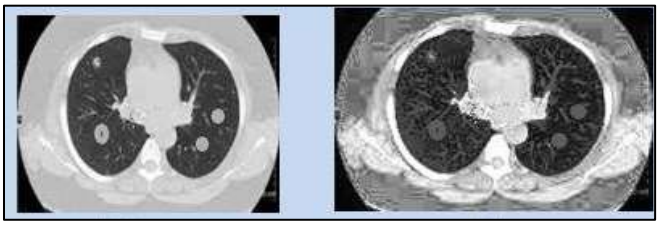
This project contains 3 modules:



4.1 ENHANCEMENT PROCESS

4.1.1 HISTOGRAM EQUALIZATION

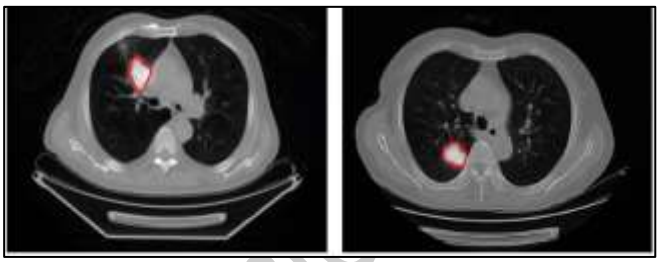
Histogram equalization is used to enhance contrast. It is not necessary that contrast will always be increase in CT-scans. There may be few cases were histogram equalization can be worse. In those cases the contrast is compromised. Histogram equalization procedure usually progresses the total contrast of various images, particularly when the operational data of the image is denoted by close contrast values. Through this modification, the intensities can be better dispersed on the histogram. This agrees for areas of lower local contrast to gain a higher contrast. Histogram equalization achieves this by efficiently spreading out the most recurrent intensity values. This technique is beneficial in images with backgrounds and foregrounds that are both bright or both dark. In particular, the technique can lead to improved views of bone structure in x-ray images, and to better details in photographs that are over or under-exposed. A significant advantage of the technique is that it is a fairly direct technique and an invertible operator.



(a) Original Image (b) Original Image after Histogram Equalization

4.2 SEGMENTATION

Segmentation is the most significant part in image processing. Fence off an entire image into several smaller parts which is something more meaningful and easier for further procedure. These several parts that are combined will cover the whole image. Segmentation also depends on other factors or features that are contained in the image. It may be either color or texture. The main aim of segmentation is to deduce the information for easier analysis. Segmentation is also useful for Image processing techniques.

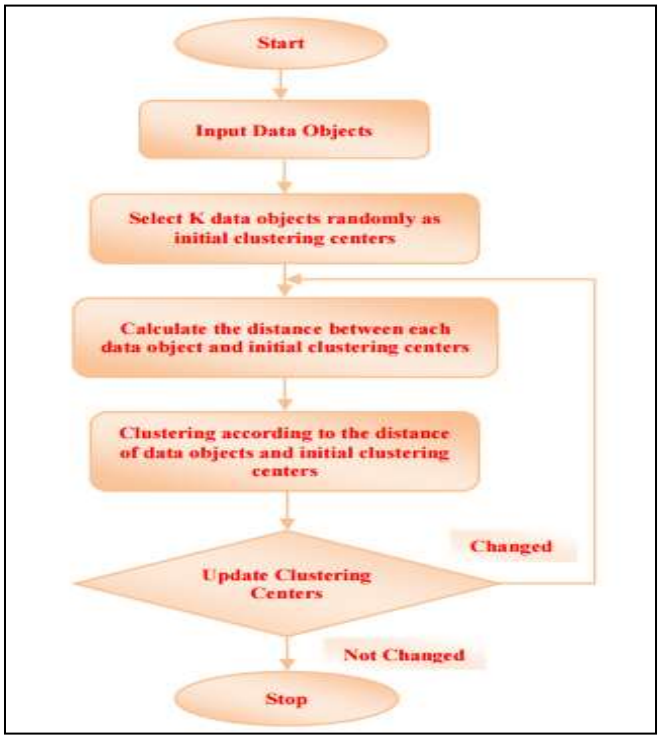


4.2.1 K MEANS ALGORITHM

K-means is one of the simplest unsupervised learning algorithms that solves the well-known clustering problem. The method follows a simple and easy way to categorize a given data set via a specific number of clusters (assume k clusters) fixed apriori. The key idea is to state k centers, one for each cluster. These centers must be placed in a cunning way because different location leads different result. So, the better option is to locate them as much as possible far away from each other.

The following phase is to take each point belonging to a particular data set and associate it to the nearest center. When no point is remaining, the first phase is accomplished and an early group age is done.

At this point we need to re-calculate k new centroids as center of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more.



4.2.2 PRINCIPAL COMPONENT ANALYSIS (PCA) ALGORITHM

Still there is an ever growing needs for techniques related to the dimensionality reduction and classification. A novel algorithm called Principal Component Analysis algorithm (PCA) is presented in our proposed work. The work partially implements k-means algorithm and then employs the principal pattern analysis algorithm, consequently evaluating the feature patterns. The inbtensity with which each principal-pattern contributes to the intensity-pattern can be denoted on a set of orthogonal axes that span a formerly presented pattern space.

PCA can be thought of as an unsupervised learning problem. The whole procedure of gaining principle components from a raw dataset can be abridged in six parts:

- Take the entire dataset containing of d+1 dimensions and ignore the labels such that our new dataset becomes d dimensional.
- Calculate the mean for every dimension of the entire dataset. The mean of matrix **A** would be $\bar{A}$.
- Calculate the covariance matrix of the entire dataset.

$$cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(Xi-\bar{x})(Yi-\bar{y})$$

- Calculate eigenvectors and the equivalent eigenvalues.
- Let $\boldsymbol{A}$ be a square matrix, $\mathbf{v}$ a vector and λ a scalar that satisfies $\boldsymbol{A}v = \lambda v$, then λ is called eigenvalue associated with eigenvector $\mathbf{v}$ of $\boldsymbol{A}$. The eigenvalues of $\boldsymbol{A}$ are roots of the representative equation:

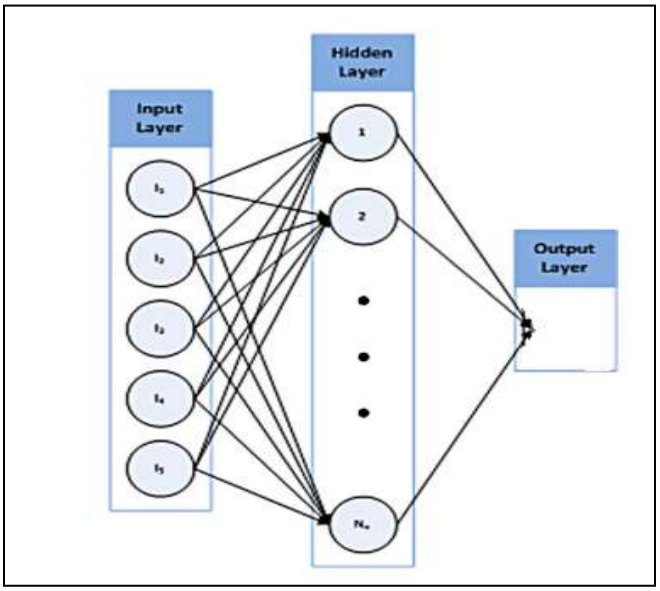
$$\det(A-\lambda I) = 0$$

- Categorize the eigenvectors by reducing eigenvalues and choose k eigenvectors with the biggest eigenvalues to form a d × k dimensional matrix **W.**
- Use this d × k eigenvector matrix to transform the samples onto the new subspace via the equation $\boldsymbol{y} = \boldsymbol{W'} \times \boldsymbol{x}$ where $\boldsymbol{W'}$ is the *transpose* of the matrix $\boldsymbol{W}$.

4.3 CLASSIFICATION

4.3.1 ARTIFICIAL NEURAL NETWORK (ANN)

ANN Classification is the process of learning to sort samples into dissimilar classes by finding common characteristics between samples of known classes. Artificial neural networks are comparatively simple electronic networks of neurons based on the neural arrangement of the brain. They process each record individually, and study by comparing their classification of the record (i.e., largely arbitrary) with the known actual classification of the record. Neural networks are usually structured in layers. Layers are made by a number of interconnected 'nodes' which comprise an 'activation function'. Patterns are presented to the network via the 'input layer', which communicates to one or more 'hidden layers' where the real processing is done through a system of weighted 'connections'. The hidden layers then connects to an 'output layer'.



.

V. CONCLUSION

Image edges help us to recognize objects. In this proposed technique, the cancerous part in the lung is identified successfully using CT scan images. Physicians use the naked eye to identify the progress and spread of cancerous nodule in the lungs from the CT scan descriptions. The professional physicians diagnose the disease and detect the stage of cancer by experience. The treatment comprises of surgery, chemotherapy, radiation therapy and targeted therapy. These treatments are extensive, expensive and excruciating. Hence, an effort is made to atomize this process to identify the lung cancer using image processing techniques. CT scan images are obtained from many hospitals. These scans (images) includes less noise as compared to X-ray and MRI images. An image improvement technique is developed for earlier disease detection; the time factor is considered to discover the abnormality concerns in target images. The CT captured images are then processed. Gabor filter and watershed segmentation gives great results for pre-processing stage. Canny Operator provides with better results for edge detection while comparing to other edge detection methods.

VI. REFERENCES

[1] Ada, Rajneet Kaur" Feature Extraction and Principal Component Analysis for Lung Cancer Detection in CT scan Images" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 3, March 2013.

[2] Neha Panpaliya, Neha Tadas, Surabhi Bobade, Rewti Aglawe, Akshay Gudadhe" A Survey On Early Detection And Prediction Of Lung Cancer" IJCSMC, Vol. 4, Issue. 1, January 2015, pg.175 – 184.

[3] C. Jeya Bharathi, Dr. P. Kabilan" Analysis and Edge Detection of Lung Cancer – Survey" International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 4 Issue: 5.

[4] Arvind Kumar Tiwari" Prediction Of Lung Cancer Using Image Processing Techniques: A Review" Advanced Computational Intelligence: An International Journal (ACII), Vol.3, No.1, January 2016.

[5] 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:23196890)(online),2347-5013(print)VolumeNo.3,Issue No.5, pp : 309-3149(2014)

[6] Dasu Vaman Ravi Prasad,"Lung cancer detection using image processing techniques", International journal of latest trends in engineering and technology.(2013)

[7] 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.

[8] 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

[9] Morphological Operators, CS/BIOEN 4640: "Image Processing Basics", February 23, 2012.

[10] 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.

[11]American Cancer Society, "Cancer facts & figures2010" http://www.cancer.org/acs/groups/content/@epidemiologysurveilance/documents/document/acspc026238.pdf (2010).

[12] "Multilevel Thresholding Based on Histogram Difference," in 17th International Conference on Systems, Signals and Image Processing. 2010.

[13] 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.

[14] S.Shah, "Automatic Cell Images segmentation using a Shape-Classification Model", Proceedings of IAPR Conference on Machine vision Applications

[15] World Health Organisation, https://www.who.int/news-room/fact-sheets/detail/cancer, September 2018