LUNG CANCER DETECTION USING MACHINE LEARNING AND IMAGE PROCESSING

[1]KATTA LAKSHMI, [2]MAVALLURU SWATHI, [3]KONDISETTY KAVITHA

[1,2,3]Assistant professor in Audisankara college of engineering and technology,Gudur, AP, India.

Abstract— Lung cancer is a severe illness with the highest rates of sickness and death among all cancers globally. Various methods have been utilized to diagnose and detect lung cancer through information analysis and classification techniques. As the causes of lung cancer remain unknown, prevention is challenging, making early detection of lung tumors the most effective treatment approach. In the detection of lung cancer, image processing and machine learning are employed to identify the presence of cancer in CT scans and blood samples. CT scan reports are more effective than Mammography, with patient images categorized as regular or abnormal, and abnormal images are segmented to focus on the tumor area. Classification is then carried out using features extracted from the dataset. The use of SVM and Image Processing techniques aims to successfully detect lung cancer and its stages accurately.

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

1. INTRODUCTION

Lung cancer is the development of a tumor known as a nodule that originates from cells lining the airways of the respiratory system. These cells are typically visible in chest X-rays and appear as circular objects. However, not all nodules seen in chest X-rays are necessarily cancerous; they could be caused by other conditions like pneumonia, tuberculosis, or calcified granuloma. Detecting lung cancer accurately has been a challenging task in medical imaging for many years. Early identification of lung nodules can significantly improve patient survival rates. Chest X-rays are commonly used for lung cancer detection, but analyzing raw chest X-ray images to identify nodules has become a complex and laborious process. Cancer is a leading cause of death globally, responsible for an estimated 9.6 million deaths in 2018. Around one in six deaths worldwide is due to cancer, with 70% of cancer-related deaths occurring in developing countries. Behavioral and dietary factors such as obesity, low fruit and vegetable intake, lack of physical activity, tobacco use, and alcohol consumption contribute to a significant portion of cancer deaths. Tobacco use is a major risk factor for cancer, accounting for about 22% of cancer-related deaths. Infections like hepatitis and HPV are responsible for approximately 25% of cancer cases in developing countries. Access to cancer diagnosis and treatment services varies widely between developed and underdeveloped countries, with significant financial implications. A novel technique for early lung cancer detection is presented in this paper, involving the collection of CT scans, image enhancement, and feature extraction to indicate normal or abnormal images. The project has been implemented and tested using real CT scan images, aiming to support efficient image processing with noise tolerance and practicality..

2. RELATED WORK

The lung lesion regions obtained from image processing are automatically targeted for refinement using a method that involves multi-constraints to control the segmentation of lesions. Vessels and visceral pleura, which have similar intensities to lung lesions, can sometimes be mistakenly included as adjacent lesions, posing challenges for accurate segmentation. To address this issue, a lung lesion refining method is employed to eliminate incorrect vascular zed regions and other tissues.

The field of medical image processing has seen significant growth and has become an interdisciplinary research area attracting expertise from various scientific and engineering fields. Computer-aided diagnostic processing plays a vital role in clinical practice, especially with the advancements in technology and the use of multiple imaging modalities. Challenges arise in processing and analyzing a large volume of images to obtain high-quality information for disease detection and treatment.

Classifier training involves collecting a large number of image datasets and extracting numerous features from each dataset. In imaging research, various variables are investigated, such as collimation, tube current, reconstruction algorithm, and breathing state in lung CT image acquisition. The goal is to identify the optimal combination of imaging parameters and features to characterize disease processes or clinical questions effectively.

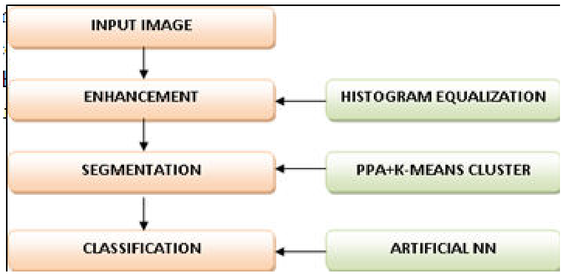
Developing accurate segmentation methods is crucial for diagnostic or prognostic assessments, especially with the advancements in imaging technology. The integration of multi-modal imaging systems requires precise segmentation and quantification of metabolic activities to enhance diagnostic capabilities.

3. PROPOSED WORK

The proposed system for detecting and predicting lung cancer aims to identify the disease in its early stages and forecast its progression, ultimately leading to a significant increase in patient survival rates. The objectives of this system include minimizing the number of testing rules, reducing the time and cost associated with unnecessary medical tests, enhancing the accuracy of lung cancer prediction and detection, utilizing fewer attributes for predicting cancer, detecting cancer in its early stages, and improving patient survivability beyond five years.

4. MODULE DESCRIPTION

This project contains 3 modules:

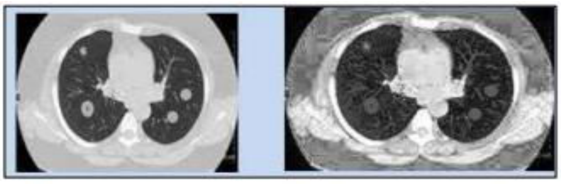


Breaking down an image into smaller, more significant parts makes it easier to work with in subsequent steps. These individual parts, when put together, form the complete image. Segmentation relies on various factors like color or texture present in the image. The primary goal of segmentation is to extract information for simpler analysis, and it is a valuable tool in image processing methods.

4.1 ENHANCEMENT PROCESS

4.1.1 Histogram Equalization

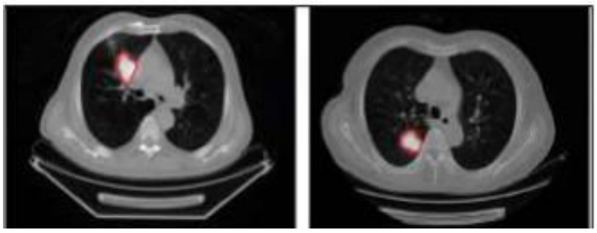
Histogram equalization is a method used to improve contrast in images. While it typically enhances contrast in various images, including CT-scans, there are instances where it may not be beneficial and could actually worsen the contrast. By adjusting the intensity values in the histogram, this technique can help to evenly distribute intensities and improve contrast in areas with lower local contrast. It is particularly useful for images with similar brightness levels in both the background and foreground. Histogram equalization can be especially effective in enhancing bone structure visibility in x-ray images and improving details in photos that are either overexposed or underexposed. One key advantage of this technique is its straightforward nature and the fact that it can be reversed.



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

4.2 SEGMENTATION

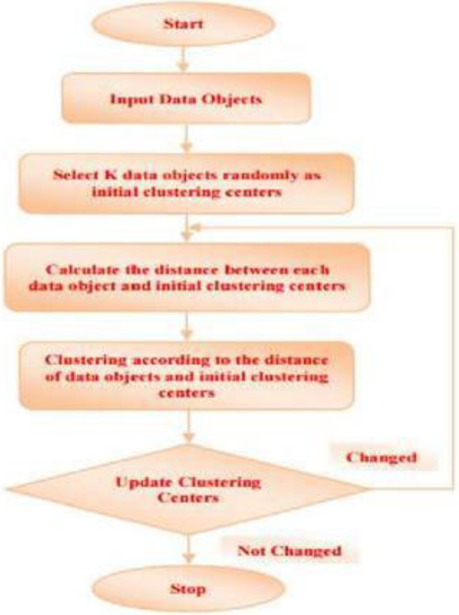
Segmentation plays a crucial role in image processing as it involves dividing an image into smaller, more meaningful parts for easier analysis. This process relies on various factors such as color or texture within the image. The ultimate goal of segmentation is to extract information that can facilitate further procedures and analysis in image processing techniques.



4.2.1 K Means Algorithm

K-means is a basic unsupervised learning algorithm used for clustering data. It categorizes a dataset into a predetermined number of clusters by placing k centers, one for each cluster, in a strategic manner to achieve optimal results. The process involves assigning each data point to the nearest center, recalculating new centroids based on these assignments, and repeating this process until the centers no longer move.

The next step involves assigning each point in a dataset to the nearest center. Once all points have been assigned, the initial grouping is completed and an initial cluster formation is established. Following this, new centroids are recalculated as the centers of the clusters formed in the previous step. These new centroids are then used to reassign each data point to its nearest new center. This process continues in a loop, during which the k centers gradually shift their positions until they reach a point where no further movement occurs.


Start
Input Data Objects
Select K data objects randomly as initial clustering centers
Calculate the distance between each data object and initial clustering centers
Clustering according to the distance of data objects and initial clustering centers
Update Clustering Centers
Changed
Not Changed
Stop

4.2.2 Principal Component Analysis (Pca) Algorithm

Techniques related to dimensionality reduction and classification are still in high demand. Our proposed work introduces a novel algorithm known as the Principal Component Analysis algorithm (PCA). The principal pattern analysis algorithm and the partial implementation of the kmeans algorithm are used in the work to evaluate the feature patterns. A set of orthogonal axes that span the previously presented pattern space can be used to indicate the intensity with which each principal pattern contributes to the intensity pattern.

PCA is a learning problem that can be approached unsupervised. The entire strategy of acquiring standard parts from a crude dataset can be compressed in six sections:

Our new dataset will be d dimensional if we take the entire dataset with d+1 dimensions and disregard the labels. Ascertain the mean for each element of the whole dataset. Matrix A's mean 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 A be a square matrix, ν a vector and λ a scalar that satisfies Aν = λν, then λ is called eigenvalue associated with eigenvector ν of A. The eigenvalues of A are roots of the representative equation:

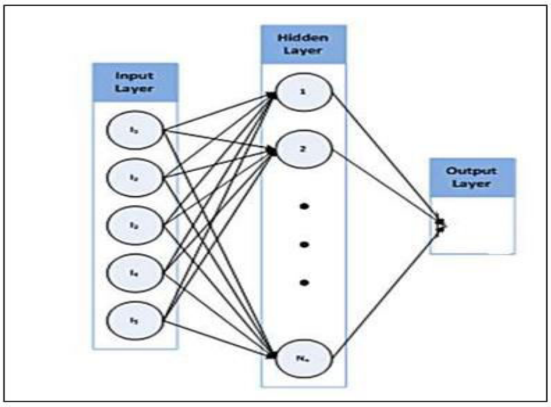
$$\det(A-\lambda I) = 0$$

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 y = W′ × x where W′ is the transpose of the matrix W.

4.3 CLASSIFICATION

4.3.1 Artificial Neural Network (ANN)

ANN Grouping is the method involved with figuring out how to sort tests into unique classes by tracking down normal attributes between tests of known classes. Counterfeit brain networks are similarly straightforward electronic organizations of neurons in view of the brain plan of the mind. They process each record separately, and concentrate by looking at their characterization of the record (i.e., generally erratic) with the known genuine order of the record. Layers typically structure neural networks. An "activation function" consists of a number of interconnected "nodes" that create layers. Through a system of weighted "connections," patterns are presented to the network by the "input layer," which communicates with one or more "hidden layers" for actual processing. After that, the hidden layers connect to an "output layer.".



5. CONCLUSION

Picture edges assist us with perceiving objects. In this proposed strategy, the harmful part in the lung is distinguished effectively utilizing CT examine pictures. From the descriptions provided by the CT scan, doctors can see with their own eyes how a cancerous nodule in the lungs has progressed and spread. The expert doctors analyze the sickness and identify the phase of disease by experience. Surgery, chemotherapy, radiation therapy, and targeted therapy are all parts of the treatment. These treatments take a long time, are expensive, and are painful. As a result, efforts are made to fragment this procedure in order to use image processing methods to identify lung cancer. CT filter pictures are acquired from numerous clinics. These sweeps (pictures) incorporates less commotion when contrasted with X-beam and X-ray pictures. For earlier disease detection, a method for improving images is developed; the time factor is considered to find the anomaly worries in target pictures. After that, the CT images are processed. Gabor channel and watershed division gives incredible outcomes for preprocessing stage. Watchful Administrator furnishes with improved results for edge identification while contrasting with other edge discovery techniques.

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/@epidemiologys urveilance/documents/document/acspc026238.pdf (2010).

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

AUTHORS

Katta Lakshmi has received M.Sc computer science Krishna Chaitanya PG college, Vikrama simhapuri university in year 2020. She is currently working as Assistant Professor. She has teaching experience:3.5 years(KCDC 1.5 year, Audisankara College of Engineering and Technology 2 years) and she Guided: 5 PG students and 20 UG students. Her research area is machine learning.

Mavalluru. Swathi has received her MCA degree from JNTUA in 2013.she is dedicated to teaching field from last 10 years. She has guided 5 P.G and 10 U.G students. Currently she is working as Assistant professor in Audisankara College of Engineering and Technology, Gudur.

Kondisetty Kavitha has received B.Tech in CSE from Swetha Engineering College, JNTUA and M.Tech Degree in from Swetha Engineering College, JNTUA in 2013 and 2016 respectively. Currently she working as Assistant professor in Audisankara college of engineering and technology since last 5 years.