Lung Cancer Detection Using Machine Learning Algorithms

Basit Ashraf
Dept of CSE
Desh Bhagat University Mandi, Gobindgarh, Punjab.

Abstract- *Creating and assessing machine learning models for the identification of lung cancer is the main topic of this research paper. One major medical condition that must be identified is lung cancer and treated right away to save the patient's life. The use of machine learning techniques has been widely applied in the healthcare industry to create predictive models for many ailments, including brain strokes, heart attacks, and lung cancer. In this work, we present an algorithm for the early identification and prediction of lung cancer. We utilized a dataset containing data on important variables associated with lung cancer, such as age, gender, anxiety, smoking status, and others, in order to develop a prediction model.The dataset underwent preprocessing in order to manage missing values, balance it, and handle categorical features. To create our predictive model, we used a variety of categorization methods, including Support Vector Machine (SVM), XGBoost, Random Forest, and Decision Trees. Many criteria were used to evaluate the models, including recall, accuracy, F1-score, and precision. The Support Vector Machine algorithm beat other models, according to our data, with 89.9% accuracy, 87.3% F1-score, 84.7% precision, and 90.5% recall.*

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 diagnosed 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 lifespan 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 60images 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

II. LITERATURE REVIEW

Softwares 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. Because of its high death rates, cancer is still regarded as a dangerous illness in the twenty-first century. The greatest rate of morbidity and death is seen in lung cancer of any cancer kind . It is possible to classify lung cancer into two primary categories: NSCLC (non-small cell lung cancer) and SCLC (small cell lung cancer). Other subtypes of non-small-cell lung cancer are lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC). Roughly 85% of cases of lung cancer are these kinds of tumors . Because of its high death rates, cancer is still regarded as a dangerous illness in the twenty-first century. The greatest rate of morbidity and death is seen in lung cancer. of any cancer kind. It is possible to classify lung cancer into two primary categories: NSCLC (non-small cell lung cancer) and SCLC (small cell lung

cancer). Other subtypes of non-small-cell lung cancer are lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC). Roughly 85% of cases of lung cancer are these kinds of tumors. In addition to the diagnosis of benign and malignant lung cancers, a more detailed classification of these diseases the prognosis of lung cancer is greatly influenced by factors including LUSC, LUAD, and SCLC. The effectiveness of treatment and, consequently, the patient survival rate are directly impacted by accurately classifying lung cancer in the early stages of diagnosis. Non-invasive diagnostic imaging techniques such as computed tomography (CT) and positron emission tomography (PET) are frequently used for both clinical diagnosis in general and the diagnosis of lung cancer in particular. The gold standard for classifying lung cancer is immunohistochemical examination. But in order to do this, a tissue sample must be taken, which is an intrusive process that carries a risk of a delayed diagnosis.and hence an increase in the patient's suffering. Advances in the field have made it possible to conduct numerous studies on automatic lung cancer diagnosis in artificial intelligence research. The use of data in lung cancer type categorization is divided into three main categories: pathological images, CT, and PET imaging data.High-quality CT pictures are made available to participants in the well-known data science community Kaggle, whose goal is to distinguish between benign and malignant lung nodules. Excellent deep learning algorithms for these tasks are consistently produced by Kaggle contests .Advances in the field of automatic lung cancer diagnosis have expanded the scope of investigations beyond the distinction between benign and malignant nodules, and the datasets used in these studies are no longer restricted to CT scans. In addition to the diagnosis of benign and malignant lung cancers, a more detailed classification of these diseases the prognosis of lung cancer is greatly influenced by factors including LUSC, LUAD. The "Comparison of Lung Cancer Detection Algorithms" study was suggested by Günaydin and associates . The prevalence of lung cancer in both genders was examined by the study's authors. They employed a variety of classifiers to identify abnormalities, including Artificial Neural Networks, Principal Component Analysis, K-Nearest Neighbors, Support Vector Machines, Naïve Bayes, Decision Trees, and Machine Learning techniques. He evaluated the specificity, sensitivity, and precision of the data using machine learning methods. Lung cancer was analyzed and segmented using an active spline model by Joon et al. Using X-ray images, this technique has made it possible to obtain lung X-ray images. First, while the preprocessing stage is running, noise should be detected using a median filter. Additional K-means and fuzzy C-means clustering are employed to collect features during the segmentation stage. The X-ray image is segmented in this work, leading to the achievement of the desired feature retrieval outcome. The Support Vector Machine (SVM) method of data classification was used to create the proposed model. MATLAB is used to simulate the impact of the cancer. system of detection. This study aimed to detect and categorize lung cancer using both malignant and normal pictures.The idea for the study "Multi-Classifier Structure for lung cells category" came from Dash and colleagues. In this study, the investigators searched for lung cancer cells using High-Resolution Computed Tomography (HRCT) scans. Several classifiers and the discrete wavelet transform are employed by the classifier to identify the first option on the input image. The input photo's functions were taken out and used to feed the Semantic Network Classifier, Ignorant Bayes Classifier, and Choice Combination, which resulted in the Accurate Choice.

III. IMPLEMENTED METHOD

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

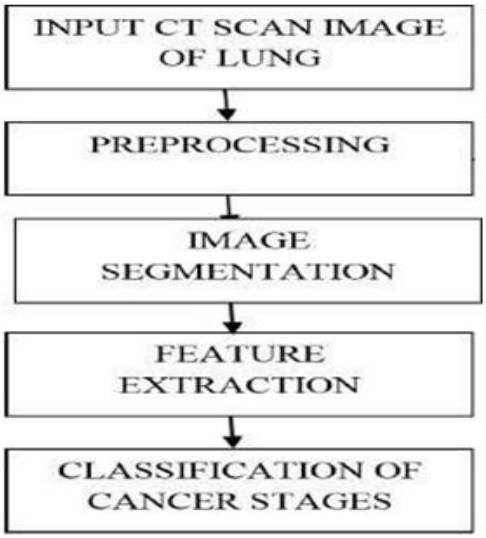
Perimeter: It is the actual number of all pixels which are interconnected on the edges of the tumor and it is the sum of all 1 binary bit pixels which are present on the outline of the nodule.

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:

- **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.
- **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.



- **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

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.

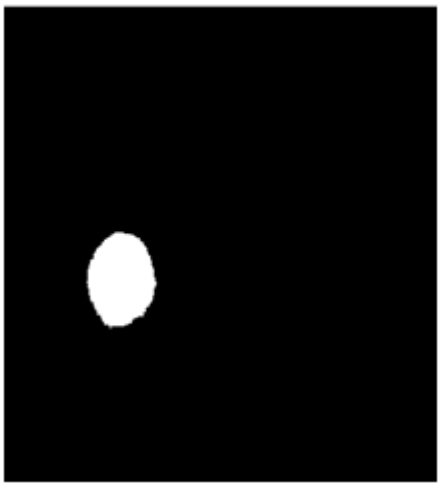
[1 2 1
0 0 0
-1 -2 -1]

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 registeredpixel 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 clos

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 occurin 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.



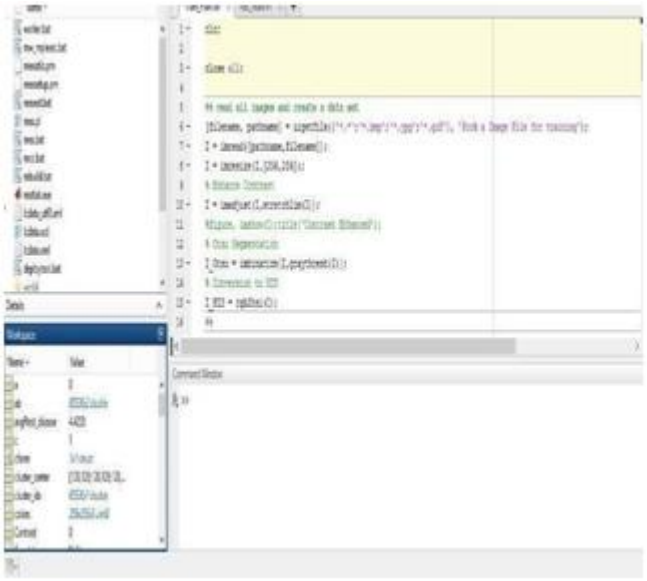
Extracted Image

IV. RESULTS AND ANALYSIS

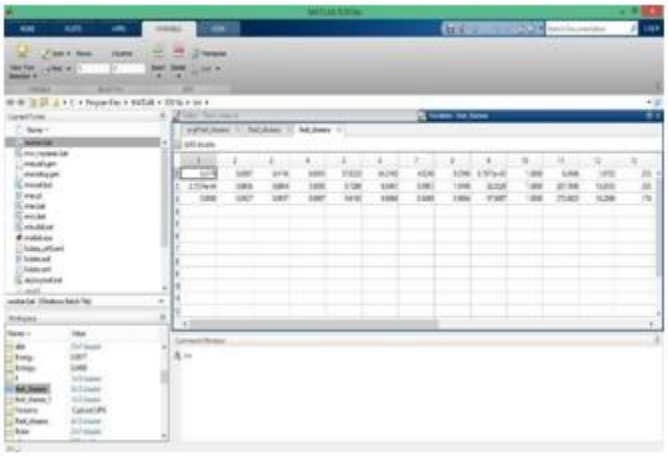
Processing the CT scan image through MATLAB we have analyzed properties through which we can find the difference in between the cancerous image and normal lung image. By going through all the image, we get a difference of all the properties and we have chosen those properties where we can get a maximum 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:



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.



Testing Output:

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.

VI. FUTURE SCOPE

Expanding the datasets used to train strategies utilizing machine learning to treat lung cancer prediction is indeed a crucial avenue for future research. By incorporating a wider range of information on potential risk factors, such as demographic data, genetic markers, environmental exposures, and lifestyle factors, More detailed and precise predictive models can be created by academics. Moreover, the scalability of machine learning algorithms means that they can benefit from larger datasets. Training the system with a vast amount of data can enhance its accuracy and generalization capabilities, enabling it to better predict lung cancer diseases across diverse populations and scenarios.

Additionally, exploring various combinations of algorithm for machine learning and techniques could lead to further improvements in predictive performance. Ensemble methods, which combine multiple models to produce a more robust prediction, could be particularly promising in this context.Overall, future research should prioritize the expansion of datasets, integration with real-time data sources, and exploration of sophisticated algorithmic methods to increase the precision and usefulness of machine learning-based lung cancer prediction models. In the battle against lung cancer, this will help us achieve better patient outcomes in terms of early detection and prognosis.

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 NewTrends 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 .

[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.