International Journal of Scientific Research and Engineering Development-– Volume 4 Issue 3, May -June 2021

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Detection of Lung Cancer Using Machine Learning Algorithms

Meghashree M*,Navyashree A M**, Merin Meleet***
*(Department of Information Science and Engineering, RashtriyaVidyalaya College of Engineering,Bengaluru
Email: meghashreem.is18@rvce.edu.in)
**(Department of Information Science and Engineering, RashtriyaVidyalaya College of Engineering, Bengaluru
Email:navyashreeam.is17@rvce.edu.in)
***(Department of Information Science and Engineering, RashtriyaVidyalaya College of Engineering, Bengaluru
Email:merinmeleet@rvce.edu.in)

--************************----------------------------------

Abstract:

Carcinoma refers to the growth of malignant cells in the lungs. As a result of the rising rate of cancer incidence, both men and women's mortality rates have increased. Lung cancer is a condition in which the cells of the lungs grow out of control. Although lung cancer cannot be avoided, the risk of developing it can be reduced. As a result, early detection of lung cancer is critical for patients. The number of persons diagnosed with lung cancer is directly proportionate to the number of people who regularly smoke. Lung cancer prediction was investigated using classification techniques such as Naive Bayes, SVM, Decision Tree, Random Forest, K-NN, and Logistic Regression. The study main objective is to assess how effective six alternative classification algorithms are at detecting lung cancer.

Keywords ***—Machine Learning, Lung cancer, Decision Tree, Logistic Regression, Naive Bayes, Support Vector Machine, Random Forest, K-NN.***

--************************----------------------------------

I. INTRODUCTION

Lung cancer is the most prevalent cause of cancer-related mortality. The windpipe, the main airways, and lungs are all places where cancer can start. The reason is the uncontrollable multiplication and dissemination of specific cells from the lungs. Cancer is more likely to be diagnosed in those who have emphysema or a history of chest difficulties.Early stage cancer refers to lung cancer that has not progressed to the surrounding tissue or other regions of the body, whereas advanced cancer has spread to the surrounding tissue or other areas of the body. A greater awareness of risk factors can help with cancer prevention. Only by detecting the yearly stage can we treat lung cancer. Early detection using machine learning techniques is critical for improving survival rates, and if we can make the diagnosis process more efficient and effective for radiologists as a result, we will be well on our way to our goal of better early detection.

The kaggle datasets were used in this lung cancer study. Using the k-fold cross validation technique, given datasets are separated into training and test data. After that, classification models such as logistic regression, support vector machine, nave bayes, decision tree, random forest, and k-nearest neighbor are created using the training data. To determine how accurate classification models are, test data is used. The goal is to compare the accuracy rates of each categorization model and determine which one is best for further investigation. The following is how the paper is organized:

Section II will talk about the literature survey. Section III will talk about the overview of study. Section IV will talk about the classification algorithms. Section V will talk about the results and discussions. Section VI will talk about the future work and the conclusion is drawn.

II. LITERATURE SURVEY

S. S. Raoof, M. A. Jabbar and S. A. Fathima[1] has conducted research on the progression and treatment of malignant illnesses, To get reliable findings, machine learning techniques were utilized. To analyze and forecast lung cancer, the healthcare sector has used machine learning techniques such as Naive Bayes, Support Vector Machine (SVM), Logistic Regression, and Artificial Neural Network (ANN). The causes of lung cancer are thoroughly investigated, as well as the use of a machine learning algorithms, with an emphasis on their relative advantages and disadvantages. Deep learning techniques could be utilized to improve and extend the Lung Cancer System's prediction and diagnosis, boosting lung cancer detection and prediction accuracy.

A. Yadav and R. Badre[2] published paper on both people and machines utilise medical imaging techniques like computed tomography to detect tumours (CT). Due to the enormous amount of CT scan pictures, radiologists found it challenging to establish a quick and correct diagnosis. As a result, the need for lung cancer computer-assisted diagnosis (CAD) has increased dramatically. For evaluating lung cancer using medical pictures and other parameters, many ML or DL-based models or systems have been presented. Medical imaging techniques are extremely useful for detecting tumours in both manual and automated settings. However, the majority of the models had issues with datasets, resulting in lower accuracy.Thus, for computer-aided systems to train their models, Medical Imaging Technique and image availability with annotation are critical.

D. Reddy, E. N. Hemanth Kumar, D. Reddy and M. P[3] has conducted research, Text-based symptoms and Machine Learning (ML) approaches can be used to predict lung cancer stages. Using machine learning techniques, this study improves the model for predicting lung cancer stages. To enhance overall prediction accuracy, the suggested model integrates K-NN, Decision Trees, and Neural Networks models with the bagging ensemble technique. When compared to individual algorithms, the suggested model's predicted results are more accurate. When the bootstrap aggregating technique is applied, individual models with the greatest accuracy scores perform better. In the future, the proposed technology could be utilized to anticipate various chronic diseases in healthcare and other fields.

S. R. Jena, T. George and N. Ponraj[4] this paper discusses about early cancer detection can aid in the complete cure of the disease. The most common lung cancer screening methods are widely accepted in the literature. To improve the effectiveness of cancer detection, a variety of approaches have been developed. Many applications are used in cancer detection, including support vector machines, neural networks, and image processing techniques. Early detection of lung cancer is complicated depending on the nature of tumour cells, which is characterized by the most of cells being covered by one another. This study discovered a number of strategies for detecting lung cancer in its early stages. Manual sample evaluation is time-consuming and inaccurate, necessitating the use of a highly skilled professional to avoid diagnostic errors.

K. Roy *et al*[6], The objective of this study is to use a combination of biological image processing technologies and data knowledge discovery to produce precision and specific value for early lung cancer detection. The Region Of Interest has been segmented and the pictures of the lungs from a CT scan have been preprocessed (ROI). To discover the unique characteristics, the Random Forest technique is employed. Using an SVM (Support Vector Machine) Classifier, the SURF (Speeded Up Robust Characteristics) approach was utilized to extract characteristics such as entropy, co-relation, energy, and variance from the Saliency Enhanced pictures. A picture's categorization decides whether it is beneficial or harmful (cancer). The method's performance is evaluated using a minimal goal vs. a number of functions evaluation plot. Throughout the process, the random forest technique and SVM classification were utilized. Using SVM classification, the best possible result is produced. With 74.2% sensitivity, 66.3% recall, and 77.6% specificity, total effectiveness is 94.5%.

N. Khosravan and U. Bagci[9] has conducted research on to develop guidelines for early diagnosis and treatment of lung cancer, many types and volume assessments of abnormalities are used. Segmentation is a hot topic in the world of computer-assisted diagnosis systems because it is critical to our understanding of nodule shape. The researchers intended to test if combining FP nodule reduction with nodule segmentation learning will help CAD systems perform better in both tasks. They back up their claim with 3D deep multi-task CNN that can tackle both issues at the same time. The algorithm achieved a segmentation accuracy of 91% and an FP reduction score of approximately 92% on the LUNA16 dataset. Over two baselines, showed increases in segmentation and FP reduction tasks, demonstrating the hypothesis. According to the findings, using the multi-task learning technique to train these two tasks at the same time increases system performance on both. Furthermore, a semi-supervised technique was demonstrated to overcome the paucity of labelled data in the 3D segmentation task.

III. OVERVIEW OF STUDY

In this study, the Kaggle datasets for lung cancer were used. Using the k-fold cross validation technique, given datasets are first separated into training and test data. The gathered training data is then used to develop classification models using techniques such as Logistic Regression, SVM, Nave Bayes, Decision Tree, Random Forest, and K-NN. The training data is used to create classification models, while the test data is used to assess model accuracy. After calculating the accuracy rates of all of the classification models we used, we made a decision.

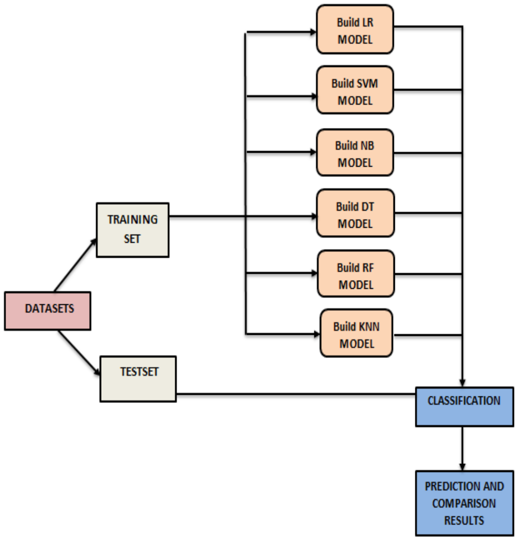

Fig 1. Overall Architecture

IV. CLASSIFICATION ALGORITHMS

A. Logistic Regression

The logistic regression approach can be used to address binary classification issues. The feature vector, also known as the logistic function, is an S-shaped curve that may assign any exact figure to a number between 0 and 1, but never exactly inside those bounds.Despite being a feature vector of inputs, the model is restricted by the default class model data. Here, begin by creating a model instance and assigning default values to it. The inverse of the normalisation strength is specified in

10. With the training data, we trained the logistic regression model, and then we used it on the test data.

B. Support Vector Machine (SVM)

Support Vector Machines that accepts actual values rather than binary integers as outputs. It is a great tool for solving prediction issues [1], [6]. It reduces inaccuracy by increasing the hyper-plane margin. In order to enhance performance, it produces a sub-class from training data called support vectors and seeks to minimize the gap between actual and predicted data.

C. Naïve Bayes

Naive Bayes is good at multi-class prediction. When the requirement of independence is met, it outperforms alternative models such as logistic regression and requires less training data [6]. We used GaussianNB from sklearn to perform the Naive Bayes method after reading the data, generating the feature vectors X and target vector Y, and separating the dataset into a training set (X train, Y train), and a test set (Y test, Y predict). True or false denotes whether the classifier successfully identified the class, whereas in the confusion matrix 'positive' or 'negative' denotes whether the classifier accurately predicted the target class, in this, 'positive' corresponds to 'cancerous cells,' as this is the cancer we want to detect.

D. Decision Tree

The decision tree method is a supervised learning technique. The decision tree is depicted in the tree structure. In this study, decision tree accepts input based on predetermined criteria and returns a true or false result. The values of each property are compared to determine the node's values. The approach is to divide the dataset into smaller parts while concurrently creating a decision tree linked with these data that finally ends at a single leaf or end node when the data subset cannot be split further. The subset's ultimate designation is chosen. In this case, continuous cross validation is used to choose a tree depth of up to 10 levels to in the expectation of best accuracy.

E. Random Forest

By combining multiple decision trees, RF (Random Forest) ensembles a forest of trees. A single decision tree, according to the reasoning, might provide either a basic or a highly specific model. More stability is provided by the random forest. This indicates that the noise in the input data set has no effect on RF (Random forest). A new case is pushed down the tree for prediction. The label of the terminal node where it stops is then assigned. All of the trees in the assembly repeat this procedure, and the label with the most incidences is reported as the prediction. Here, the number of trees in the forest is counted in 100.

F. K-Nearest Neighbours (K-NN)

The K-NN method saves all available data and classifies a new data point based on its similarity to the existing data. This implies that fresh data may be quickly categorized into a well-defined category using the K-NN method. In the feature vectors X and target vector Y, and separating the dataset into a training set (X train, Y train), and a test set (Y test, Y predict). True or false denotes whether the classifier successfully identified the class, whereas in the confusion matrix 'positive' or 'negative' denotes whether the classifier accurately predicted the target class, in this example, 'positive' corresponds to 'cancerous cells,' as this is the cancer we want to detect and true negative is detects that healthy people are not sick. The confusion matrix, based on the positive and negative values, denotes whether the classifier accurately predict the target class.

V. RESULTS AND DISCUSSIONS

Each of the six algorithms' testing results are shown below. The number of right answers divided by the total number of predictions is used to measure classification accuracy. A certain outcome has an impact on these variables. The terms TP (True Positive) and TN (No Event) refer to correctly anticipated event values and no event values, respectively (True Negative). The acronyms FP (False Positive) and FN (False Negative) denote wrongly anticipated event values and no mistakenly predicted event values, respectively.

Algorithms	Precision	Recall	F1-Score	Support
Logistic Regression	0.94	1.00	0.97	30
SVM	0.86	0.74	0.80	30
Naïve Bayes	0.94	1.00	0.97	30
Decision Tree	0.94	0.86	0.90	30
Random Forest	0.94	0.92	0.93	30
K-NN	0.86	0.82	0.83	30

Table 1. Classification Report

It compares and contrasts six machine learning algorithms for determining post-operative survival in lung cancers using predictive data mining methods. On the aforementioned algorithms, a stratified 10-fold cross-validation comparison study was conducted, and accuracy was calculated for each classifier. The accuracy rates of each classifier are compared. Quantitative comparisons of classifier performance are made. The Logistic Regression and Naïve Bayes as same and giving more accuracy in the project as shown in the table. On the lung cancer dataset, different outcomes are produced for each classifier.

The PCA's fundamental idea is to look at how features are related to each other. If there is a lot of correlation among a subset of the features, PCA will try to combine them and represent the data with a lower number of completely uncorrelated features.We can see that the first ten main components keep roughly 96.67% of the dataset's variability despite removing 20 (30-10) features. The remaining 20 features account for less than 5% of the data variability.

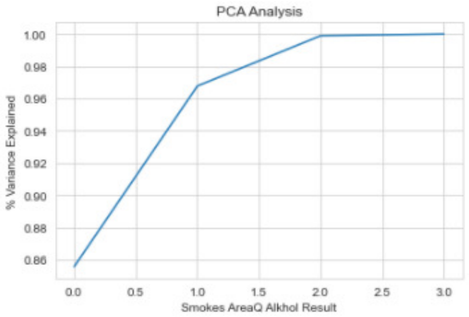

Fig 2. PCA analysis

VI. CONCLUSION AND FUTURE ENHANCEMENTS

A doctor used to have to perform numerous tests to determine whether or not a patient had lung cancer in the past. A patient may be subjected to unnecessary examinations or tests during a diagnosis in order to diagnose lung cancer. In the prediction and classification of medical data, machine learning techniques are now widely used. For comparative analysis, machine learning methods include logistic regression, SVM, decision trees, Random Forest, Nave Bayes, and K-NN. There is a comparison of each classifier's accuracy rates [6]. The performance of classifiers is compared quantitatively. On the lung cancer dataset, different outcomes are produced for each classifier during the performance. When it comes to the correct classification and other metrics, the machine learning algorithms, namely Logistic Regression and Naïve Bayes provide the best results. These algorithms used high-dimensional classification to classify the observation, resulting in the best results. Using these tools, it is possible to identify lung cancer with greater accuracy.

There are many advancements which can be done on this project some are, 1. A number of imaging, such as X-rays, CT scans, MRIs, and PET scans, can be used to improve accuracy. 2. We can strive to develop automated medical image processing techniques that can detect cancer cells before they become visible. Because this disease has a negative economic impact, more research should be conducted to discover knowledge gaps in disease, control and detection methods, which will aid in the development of vaccines and other control measures.

REFERENCES

[1] S. S. Raoof, M. A. Jabbar and S. A. Fathima, "Lung Cancer Prediction using Machine Learning: A Comprehensive Approach," 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), 2020.

[2] A. Yadav and R. Badre, "Lung Carcinoma Detection Techniques: A Survey," 2020 12th International Conference on Computational Intelligence and Communication Networks (CICN), 2020.

[3] D. Reddy, E. N. Hemanth Kumar, D. Reddy and M. P, "Integrated Machine Learning Model for Prediction of Lung Cancer Stages from Textual data using Ensemble Method," 2019 1st International Conference on Advances in Information Technology (ICAIT), 2019.

[4] S. R. Jena, T. George and N. Ponraj, "Texture Analysis Based Feature Extraction and Classification of Lung Cancer," 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2019.

[5] N. S. Nadkarni and S. Borkar, "Detection of Lung Cancer in CT Images using Image Processing," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 2019.

[6] K. Roy et al., "A Comparative study of Lung Cancer detection using supervised neural network," 2019 International Conference on Opto-Electronics and Applied Optics (Optronix), 2019.

[7] J. Alam, S. Alam and A. Hossan, "Multi-Stage Lung Cancer Detection and Prediction Using Multi-class SVM Classifie," 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), 2018.

[8] T. Patel and V. Nayak, "Hybrid Approach for Feature Extraction of Lung Cancer Detection," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018.

[9] N. Khosravan and U. Bagci, "Semi-Supervised Multi-Task Learning for Lung Cancer Diagnosis," 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2018.

[10] Q. Wu and W. Zhao, "Small-Cell Lung Cancer Detection Using a Supervised Machine Learning Algorithm," 2017 International Symposium on Computer Science and Intelligent Controls (ISCSIC), 2017.

[11] M. Vas and A. Dessai, "Lung cancer detection system using lung CT image processing," 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA), 2017.

[12] R. Sammouda, "Segmentation and Analysis of CT Chest Images for Early Lung Cancer Detection," 2016 Global Summit on Computer & Information Technology (GSCIT), 2016.