How to Cite:
Ramu, B., Vakiti, S., & Asuri, V. (2022). Analysis of lung cancer predictionusing machine learning algorithms. *International Journal of Health Sciences*, *6*(S3), 8623–8634. https://doi.org/10.53730/ijhs.v6nS3.8053

Analysis of lung cancer predictionusing machine learning algorithms

B. Ramu
Assistant Professor, Geethanjali College of Engineering and Technology
*Corresponding author email: ramu0604.ece@gcet.edu.in

Srija Vakiti
UG Student, Dept. of ECE, Geethanjali College of Engineering and Technology

Vaishnavi Asuri
UG Student, Dept. of ECE, Geethanjali College of Engineering and Technology

***Abstract*---**With the advancement of artificial intelligence, Machine learning approaches have played a crucial role in identifying, predicting, and diagnosing malignant illnesses due to their speed and accuracy. In the healthcare industry, machine learning algorithms like Logistic Regression, KNN Classifier, Support Vector Machine, Decision Tree, Nave Bayes, Random Forest, AdaBoost, Stacking Classifier, and Voting Classifier have contributed to the analysis and prediction of lung cancer prognosis [1]. This paper discusses the comparative study of all nine algorithms mentioned above. As a result, instead of referring to many publications, this document will assist scholars in swiftly scanning the project results.

***Keywords*---**KNN classifier, decision tree, Adaboost, feature scaling, correlation matrix.

Introduction

One of the significant contributors to cancer-related deaths worldwide is lung cancer. During its early stages, lung cancer is notoriously difficult to analyze and detect as it does cause any pain and symptoms (in some cases). Lung cancer diagnosed patients may suffer cough, chest pain, shortness of breath, wheezing, hemoptysis, i.e., coughing up blood, Pancoast syndrome (shoulder pain), hoarseness (paralysis of vocal cords), weight loss, weakness, and fatigue. Therefore, early identification, prediction, and diagnosis of lung cancer are critical as they speed up and simplify the recovery process. With the advancement of artificial intelligence, Machine learning approaches have improved the progression and treatment of malignant illnesses due to their speed and accuracy. The

current algorithms used for the analysis and prognosis are Naïve Bayes, Support Vector Machine (SVM), Logistic Regression and Artificial Neural Network (ANN) [2]. These methods are less accurate and have less probability of prediction [8].

An Overview of Study

The implementation of code begins with a study of existing data to help understand the categories and their statistics better. This is followed by feature selection to aid dimensional reduction. Various ML algorithms are then selected based on compatibility with the processed data. Upon executing the said models, a comparative analysis is performed to choose the best one.

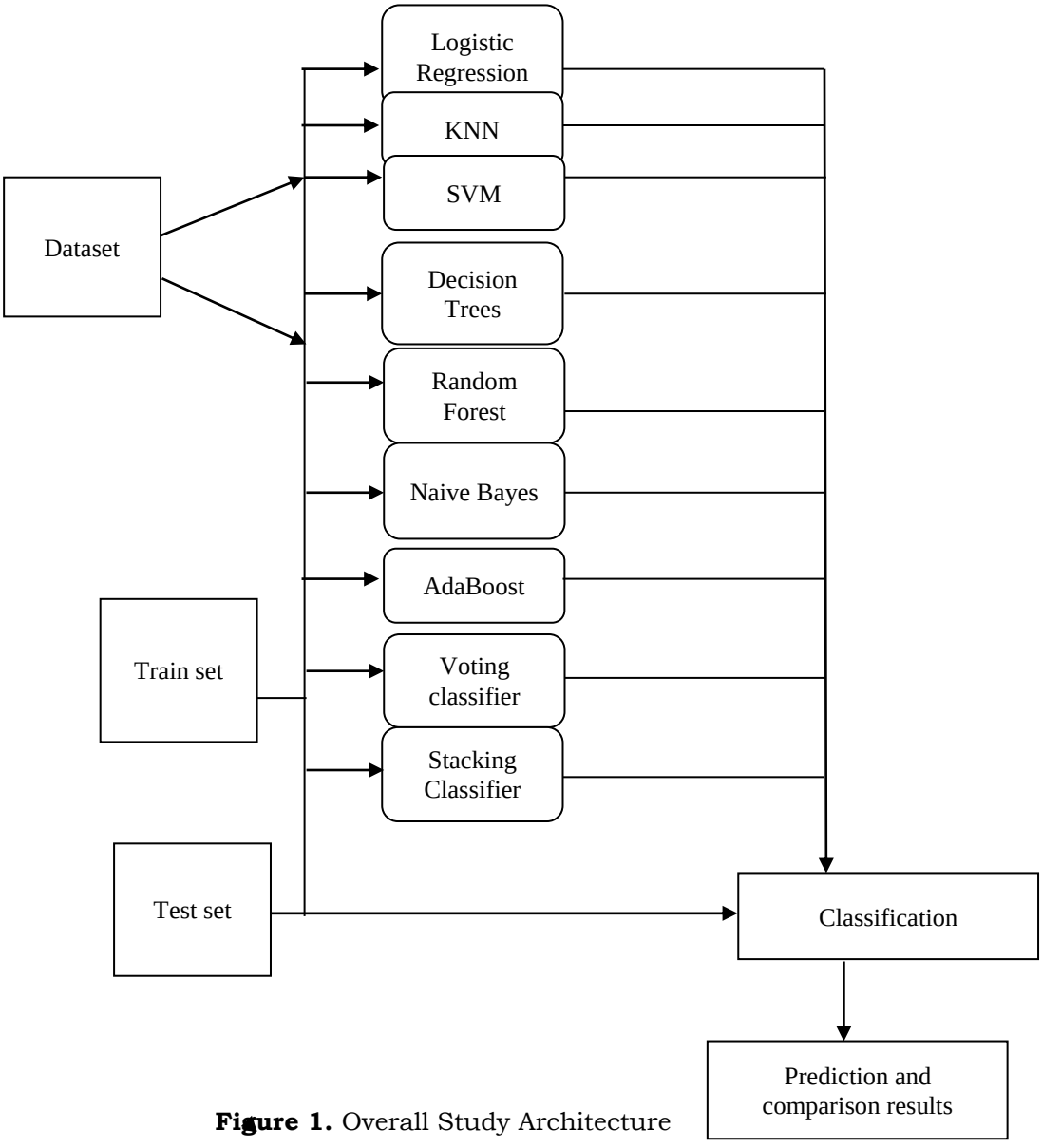

Figure 1. Overall Study Architecture

Related Past Works

It is crucial to understand today's most widely used technologies before looking for new solutions and methods to solve the same problems. After a thorough examination, we found that most Lung Cancer Prediction using Machine Learning models uses Image Classification, and very few use symptom-based analysis. However, creating a model that follows a classification approach based on assessing symptoms is the easiest way to bring awareness to the general public. It leaves much scope to create a User-Interface such as a web application to predict the presence or absence of Lung Cancer with dynamic data consisting of simple Yes/No questions.

[1] The paper primarily focuses on the processing and classification of data based on the Computed Tomography CT scans of patients. A conventional and widely used approach is mainly used to detect lung cancer at various stages of development rather than using symptoms as a basis for classification.

[2] This paper uses images processing as well. However, unlike the previous approach, which eliminates unnecessary areas of interest to find the cancerous part of the scanned image, this method uses a reference image that "masks" the test image. Thus models are trained using reference images that compare closely to the test data.

[3] This paper revolves around comparisons and exploration of techniques used to identify lung cancer based on Machine Learning. The techniques discussed in the paper are Support Vector Machines, Decision Trees, Convolutional Neural Networks, Random Forest classifier, Naive Bayes Classifier, Linear Regression and Linear Discriminate Analysis.

[4] This paper's central point is the Classification of Computed Tomography images using the Random Forest Classifier. The results are then examined under various conditions to maximize the output efficiency obtained. This dataset is taken from Lung Image Database Consortium (LIDC). Hence, this paper provides an in-depth analysis of Random Forest Classification and its application in Lung Cancer data classification.

[5], [6], [7] These papers helped with the quality analysis of machine learning models implemented. Precision, Recall, F1-Score, Sensitivity, Confusion Matrix and Loss function curves are discussed in-depth in these papers.

[8] Similar to [3], the main aim was to perform a literature study of the existing ML models that help with Lung Cancer Prediction. Specifically, this paper surveys Support Vector Machines, Naive Bayes classifier, Artificial Neural Networks and Logistic Regression methods. It is a quick and efficient way to go over currently used technologies without getting too much into the details.

Materials and Methods

The procedure of implementation can be broadly classified into the following sections:

- Importing Datasets and Libraries
- Data Preprocessing
- Feature Extraction and Selection
- Model Building

Below is a brief overview of each process involved.

Importing datasets and libraries

The dataset used for this project was taken from Kaggle. The execution platform used was Jupyter Notebooks since it provides cell to cell execution that makes error correction faster and easier. Since the project works based on Machine Learning models, libraries such as Numpy, Pandas, Sklearn, Seaborn and Matplotlib were imported. It consists of fifteen factors contributing to Lung Cancer taken across three hundred and nine variables. Following are the initial fifteen factors present in the raw data file:Yellow Fingers, Anxiety, Peer Pressure, Chronic Disease, Fatigue, Allergy, Wheezing, Alcohol consumption, Coughing, Shortness of Breath, Swallowing difficulty, Chest pain.

Table 1
Tabular view of symptoms recorded from 6 patients (P1-P6)in terms of numbers (1=No and 2=Yes)

	P1	P2	P3	P4	P5	P6
Gender	M	M	F	M	F	F
Age	69	74	59	63	63	75
Smoking	1	2	1	2	1	1
Yellow Fingers	2	1	1	2	2	2
Anxiety	2	1	1	2	1	1
Peer Pressure	1	1	2	1	1	1
Chronic Disease	1	2	1	1	1	2
Fatigue	2	2	2	1	1	2
Allergy	1	2	1	1	1	2
Wheezing	2	1	2	1	2	2
Alcohol Consumption	2	1	1	2	1	1
Coughing	2	1	2	1	2	2
Shortness of Breath	2	2	2	1	2	2
Swallowing Difficulty	2	2	1	2	1	1
Chest Pain	2	2	2	2	1	1
Lung Cancer	YES	YES	NO	NO	NO	YES

Data Pre-Processing

To ensure the maximum efficiency of ML models and prevent bias in the results, missing values must be compensated for. The dataset under consideration consists of no missing values. The current size of the dataset is three hundred and nine rows(variables) and sixteen columns(including a column for the presence or absence of lung cancer). Gender and Lung Cancer columns had to be label-encoded since they were not in the integer data type. All features are mapped for correlation to eliminate unnecessary features.

Feature Extraction and Selection

As shown in the above correlation matrix, certain features have null to negative correlations. Thus, they were exempt from further processing and training of models. They are Age, Gender, Shortness of Breath and Smoking. One surprising conclusion made here is that smoking- one of the widely presumed causes of cancer- is less of a contributor when compared to others, such as Alcohol consumption and peer pressure.

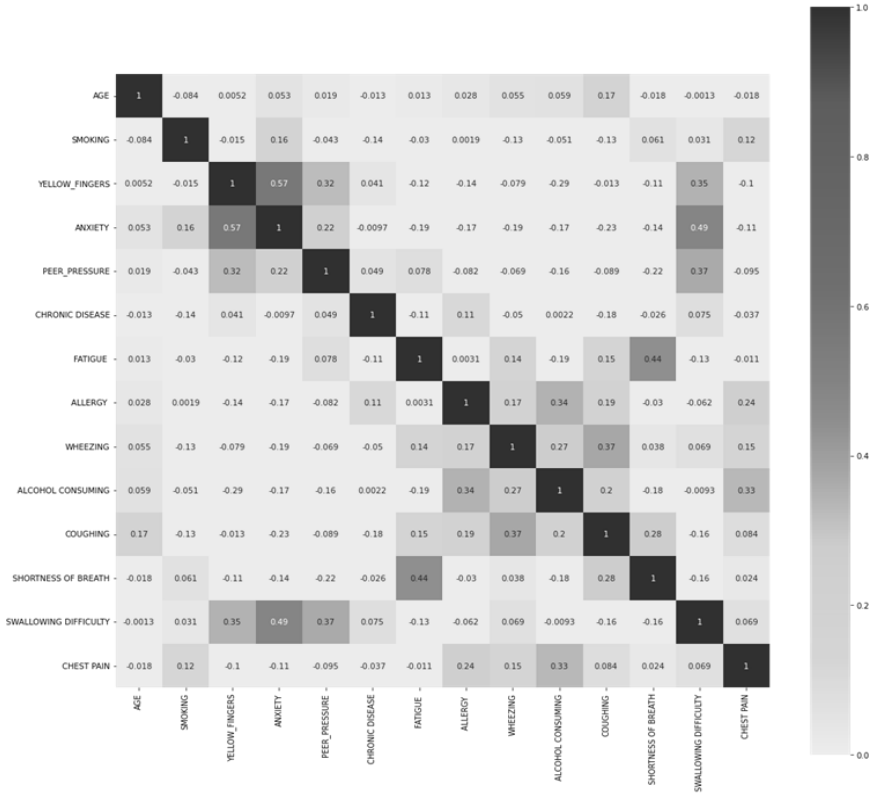

Figure 2. Correlation matrix for the features mentioned in section 2.1

The most prominent contributors to Lung Cancer were anxiety, the presence of Chronic disease and Yellow fingers. After eliminating the aforementioned features

and the Lung Cancer column, the dataset was reduced to Three hundred and nine rows and eleven columns.

Model Building

The following nine Machine Learning models were used to predict the outcome: Logistic Regression, KNN Classifier, Support Vector Machine, Naïve Bayes, Random Forest, AdaBoost, Stacking Classifier Voting Classifier, and Decision Tree.

- Logistic Regression: It is a classification algorithm used to predict binary outcomes for a given set of independent variables. This algorithm sets a threshold value. Any value above the threshold value is taken as outcome 1, and any value below the threshold value is taken as 0. This algorithm is applied in text editing and credit scoring. Its disadvantages include being bound to discrete outputs only and a mandatory requirement of multicollinearity between its independent variables.
- K-Neighbors Classifier: This stands for K-Nearest Neighbour Classifier. This algorithm classifies a data point based on how its neighbours are classified. This algorithm stores all available cases and classifies all new cases based on the measure of similarity. The 'k' is a parameter that refers to the number of nearest neighbours included in the majority voting process (done to classify the new case). This 'k' value is usually √n (n= number of data points). If the result is even (x), the final k value is x+1 or x-1. This algorithm is applied in facial recognition and recommendation systems in Amazon, Hulu, Netflix etc. Its disadvantages include not working well with large datasets and sensitivity to noisy/missing data.
- Support Vector Machine: This ML algorithm looks at data and sorts it into 1 of 2 specific categories by mapping them through a hyperplane (a line that serves as the classification border). This algorithm then finds the best hyperplane (the plane that has the max distance from the support vectors–extreme points in the datasets–on both sides) of all and declares it the outcome. [3] This algorithm is applied in image classification and handwriting detection. Its disadvantages include high noise sensitivity and slowness with respect to large datasets.
- Naïve Bayes: This algorithm is one of the simplest ML algorithms. "Naïve" is used because the algorithm assumes by default that all the features entered as input to the model are independent of each other. It then makes frequency tables of all the independent variables and constructs likelihood tables for all of them. Outcomes are calculated using these likelihood tables. It is based on the Bayes theorem and tells us about the probability of a feature based on prior knowledge about the conditions that are related/might lead to the feature. This algorithm is applied in spam filtering and multi-class predictions. Its disadvantages include bad estimation and unrealistic assumption that all factors are independent.
- Random Forest: This algorithm consists of several decision trees during the training phase. The prediction of each tree is taken and based on majority votes the final output is predicted. This algorithm is immune to overfitting due to many decision trees, which give low bias and low variance. It is used in real-life situations where run-time performance is essential [4]. This

algorithm is applied in credit card fraud detection, product recommendation and price optimization. Its disadvantages include lower prediction accuracy for complex problems compared to gradient-boosted trees and less interpretable compared to a single decision tree. It also requires significant storage memory after training.

- AdaBoost Classifier: This classifier stands for Adaptive Boosting. It is an ensemble/meta-learning method created to increase the efficiency of binary classifiers. It is a sequential learning algorithm where many models are generated sequentially (one after the other), and the successive models learn the mistakes of the previous models. Here the dependency of the models on their predecessors is high and is exploited. This algorithm is applied in predicting customer churn, classifying the topics a customer is interested in etc. Its disadvantages include outlier sensitivity and slower execution compared to XGBoost.
- Stacking Classifier: This is an ensemble learning technique (a hybrid of basic classifiers with other meta-classifiers). 2 or 3 levels of classifiers exist. Level 2 classifiers analyse level 1 output predictions and then give the final output. The stacking classifier used in our project consists of 3 algorithms (logistic regression, random forest, SVM). This algorithm is applied to improve the accuracy of pre-existing algorithms. Its disadvantages include higher cost of deployment and unnecessary complexity for small datasets.
- Voting Classifier: This classifier is an ensemble learning-based algorithm where basic classifiers and meta-classifiers are combined. It aggregates the total findings given to it by the 3 classifiers and then produces an output based on the highest majority, aka the output that gets majority votes. The voting classifier used in our project consists of 3 algorithms (logistic regression, random forest, Gaussian Naive Bayes). This algorithm is applied to improve the accuracy of pre-existing algorithms. Its disadvantages also include higher cost of deployment and unnecessary complexity for small datasets.
- Decision Tree Classifier (highest accuracy): This algorithm maps the data into a tree-like graph of decisions and their consequences, making it excellent for predicting future outcomes. It uses a set of logical if-then conditions to classify problems. This algorithm is applied in fraudulent statement detection, healthcare management etc. Its disadvantages include high training time, expensiveness and sensitivity to data changes (no matter how small they may be).

Each model was trained with eighty percent of the data and tested with the rest twenty percent. The aim was to compare all the models and select the one with the most accuracy.

Observations

- The decision tree classifier outperformed all the other ML models by reaching an accuracy of 93.54%
- It was followed closely by the Random Forest classifier and Voting classifier, which achieved 91.93% accuracy.
- These models performed moderately with an accuracy of 90.322% - Logistic Regression, KNN classifier, SVM classifier and AdaBoost classifier.

- The least accuracy was around 87%, given by Naive Bayes and Stacking Classifier

Table 2
Obtained accuracy of all 9 algorithms mentioned in section 2.4

Model	Accuracy(%)
Logistic Regression	90.32
KNN Classifier	90.32
Support Vector Machine (SVM)	90.32
Naïve Bayes	87.09
Random Forest	91.93
AdaBoost	90.32
Stacking Classifier	87.7
Voting Classifier	91.93
Decision Tree.	93.54

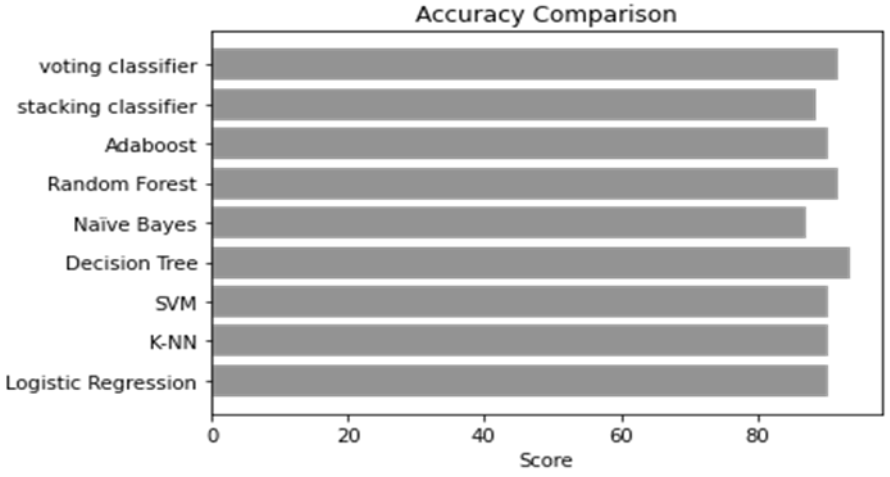

Figure 3. Accuracy comparison graph of all 9 algorithms

Quality Metrics

For any model, accuracy is not a measure enough to check the efficiency. Several other metrics are used to determine where the misclassifications are occurring [5]. Precision, Recall, F1-Score, Sensitivity and Specificity were calculated for this project [6]. Here are the mathematical formulae for the quality metrics used where TN= True Negatives, TP= True Positives, FN= False Negatives and FP= False Positives.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall/Sensitivity} = \frac{TP}{TP+FN}$$

$$\text{Specificity} = \frac{TN}{TN+FP}$$

F1-Score= $2*\frac{Precision * Recall}{Precision + Recall}$

Table 3
Comparison of 9 algorithms in terms of precision, recall and F1-Score

Model	Accuracy(%)	Case	Precision	Recall	F1-Score
Logistic Regression	90.32	0	0.83	0.50	0.62
		1	0.91	0.98	0.94
KNN Classifier	90.32	0	0.75	0.60	0.67
		1	0.93	0.96	0.94
SVM Classifier	90.32	0	0.83	0.50	0.62
		1	0.91	0.98	0.94
Naïve Bayes	87.09	0	0.62	0.50	0.56
		1	0.91	0.94	0.92
Random Forest	91.93	0	0.78	0.70	0.74
		1	0.94	0.96	0.95
AdaBoost Classifier	90.32	0	0.83	0.50	0.62
		1	0.91	0.98	0.94
Stacking Classifier	87.70	0	0.80	0.40	0.53
		1	0.89	0.98	0.94
Voting Classifier	91.93	0	0.86	0.60	0.71
		1	0.93	0.98	0.95
Decision Tree (highest accuracy)	93.54	0	0.80	0.80	0.80
		1	0.96	0.96	0.96

The above figure shows that the quality metrics were calculated for both the cases (Presence of Lung Cancer= 1; Absence of Lung Cancer= 0) using the confusion matrix [7]. Thus, we can conclude that the decision tree, with the best values for Precision, Recall and F1-Score, is the most efficient model. Graphically, these values can be represented as follows.

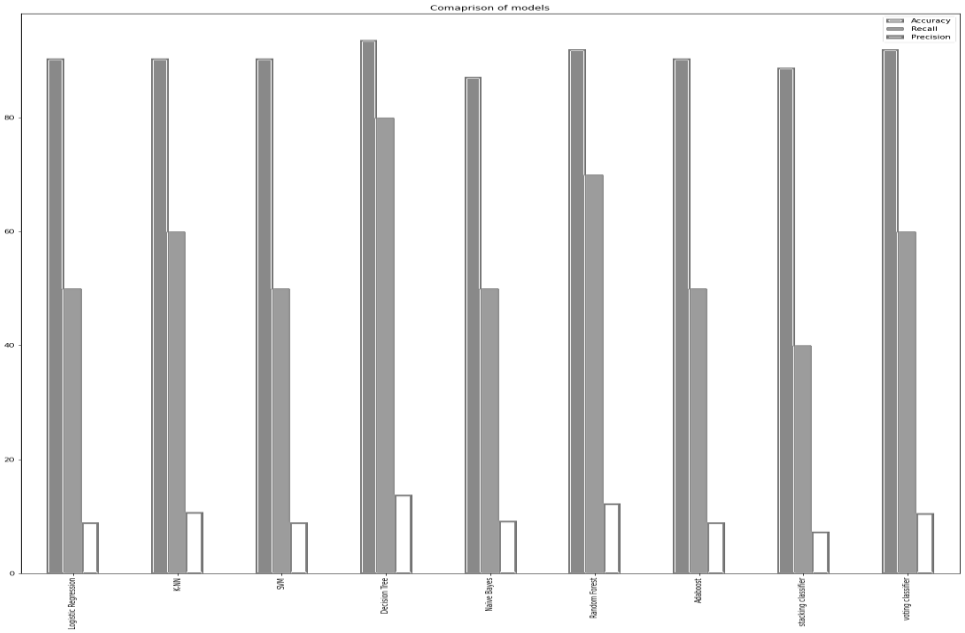

Figure 4. Graphical representation of Table3

Results

Decision trees have many advantages over other models, making them a winner.

- The amount of Pre-Processing required here is relatively minor since there is no need to scale data.
- Decision trees are interpretative in a way that makes it very easy to arrive at the outcome by following the branches of the trees.

Although it might seem enough for a model to satisfy these criteria, here are a few aspects of it that leave us a step behind while considering more extensive datasets on a much larger scale

- They take a significant amount of time to train. The computational time for Decision trees was 0.8 seconds. Compared to others, such as KNN and Logistic Regression (both taking about 0.3 seconds), Decision trees seem slower.
- Although missing values do not affect Decision Trees, a slight change in existing values can cause a significant change in the outcome, resulting from the instability of trees.

Table 4
Results after testing the model built using Decision Tree Algorithm

Case	1	2	3	4	5
Yellow Fingers	No	No	No	Yes	No
Anxiety	No	No	Yes	Yes	Yes
Peer Pressure	Yes	No	No	Yes	No
Chronic Disease	Yes	Yes	No	No	No
Fatigue	Yes	Yes	No	No	Yes
Allergy	No	No	Yes	Yes	No
Wheezing	No	No	Yes	No	Yes
Alcohol Consumption	No	No	Yes	Yes	No
Coughing	Yes	Yes	Yes	No	Yes
Swallowing Difficulty	No	No	No	No	No
Chest Pain	No	No	No	No	Yes
Presence of Lung Cancer	Yes	No	Yes	Yes	No

Scope and Conclusion

Overall, this paper provides a comparative analysis that will assist doctors and scholars in swiftly scanning the project results instead of referring to many publications. The prediction and diagnosis of the Lung cancer system can be embellished and extended further by employing deep learning techniques to enhance the accuracy of both identification and prediction of lung cancer. Clarity of all figures is of the utmost importance. If the final version is not prepared in a format or does not include all the details, the publication process will be delayed.

Acknowledgements

This work is supported by the Research Program supported by the Department of Electronics and Communication, Geethanjali College of Engineering and Technology, India.

References

1. R. P.R., R. A. S. Nair and V. G., "A Comparative Study of Lung Cancer Detection using Machine Learning Algorithms," 2019 IEEE International

Conference on Electrical, Computer and Communication Technologies (ICECCT), 2019, pp. 1-4, doi: 10.1109/ICECCT.2019.8869001.
2. Waykule Jyoti, Sushmita S. Patil, Aishwarya V. Budhe, Rutuja R. Chavan. Lung Cancer Detection Using Machine Learning (ISSN-2349-5162)
3. Pradeep K R, Naveen N C. Lung Cancer Survivability Prediction based on Performance Using Classification Techniques of Support Vector Machines, C4.5 and Naive Bayes Algorithms for Healthcare Analytics, Procedia Computer Science, Volume 132, 2018, Pages 412-420, ISSN 1877-0509,
4. D. Jayaraj and S. Sathiamoorthy, "Random Forest based Classification Model for Lung Cancer Prediction on Computer Tomography Images," 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT), 2019, pp. 100-104, doi: 10.1109/ICSSIT46314.2019.8987772.
5. Sokolova, M., Japkowicz, N., Szpakowicz, S. (2006). Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. In: Sattar, A., Kang, Bh. (eds) AI 2006: Advances in Artificial Intelligence. AI 2006. Lecture Notes in Computer Science(), vol 4304. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11941439_114
6. Zhou, J.; Gandomi, A.H.; Chen, F.; Holzinger, A. Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. Electronics 2021, 10, 593. https://doi.org/10.3390/ electronics10050593
7. N. D. Marom, L. Rokach and A. Shmilovici, "Using the confusion matrix for improving ensemble classifiers," 2010 IEEE 26-th Convention of Electrical and Electronics Engineers in Israel, 2010, pp. 000555-000559, doi: 10.1109/EEEI.2010.5662159.
8. 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, pp. 108-115, doi: 10.1109/ICIMIA48430.2020.9074947.