Computational Journal of Mathematical and Statistical Sciences
4(1), 139–161
DOI:10.21608/cjmss.2024.309277.1068
https://cjmss.journals.ekb.eg/

Research article

Leveraging Data Mining for Inference and Prediction in Lung Cancer Research

Md Nurul Raihen[1]*, Shakera Begum[2], Sultana Akter[3], Md Nazmul Sardar[4]

[1] Department of Mathematics and Computer Science, Fontbonne University, Saint Louis, 63105, MO, USA

[2] Department of Statistics, Western Michigan University, Kalamazoo, 49006, MI, USA; shakerabgm@gmail.com

[3] Institute for Data Science and Informatics, University of Missouri Columbia, 65211, MO, USA; sa4kf&umsystem.edu

[4] Senior Officer, Product Development at Radiant Nutraceuticals Ltd, Dhaka, 1000, Bangladesh

* **Correspondence:** nraihen@fontbonne.edu

Abstract: Lung cancer is the second most common cancer worldwide, with an estimated 2.21 million new diagnoses and 1.8 million deaths in 2020, according to WHO. Successful lung cancer treatment, early detection, and diagnosis improve survival rates. This study included 270 lung cancer patients and 39 with no lung cancer patients. Logistic regression will be used to analyze the association between variables for inference and Linear Discriminant Analysis, Quadratic Discriminant Analysis, Logistic Regression Analysis, k Nearest Neighborhood, Decision tree, Bagging, Random Forest, and Support Vector Machine used to predict the likelihood of an individual developing lung cancer based on factors. In terms of accuracy, 5 fold cross validation showed higher accuracy than the validation set approach where the Logistic Regression Model had the highest accuracy of 93.54%, followed by the Linear Discriminant Analysis with an accuracy of 92.09%, the Support Vector Machine with an accuracy of 91.29%, Bagging and Random Forest with an accuracy of 90.90 and 91.23 respectively, the Quadratic Discriminant Analysis with an accuracy of 89.97, Decision Tree with an accuracy of 89.97, the Knn-10 model with an accuracy of 17.74%, and lastly KNN-5 Model with an accuracy of 16.12%. The logistic regression model identified key associations between lung cancer and factors such as Allergy, Peer pressure, Swallowing difficulty, Smoking, Chronic disease, Alcohol consumption, yellow fingers, Fatigue, and Coughing. The accuracy rankings varied between 5-fold cross-validation and validation set approaches. Notably, the logistic regression model consistently demonstrated superior performance, achieving an accuracy rate of 93.54%.

Keywords: Lung Cancer; logistic regression; Classification; Accuracy; Validation; Data Mining
Mathematics Subject Classification: 62H30; 62J12; 11R29
Received: 30 Agust 2024; **Revised:** 26 October 2024; **Accepted:** 13 November 2024; **Online:** 21 November 2024.

1. Introduction

Lung cancer poses a significant global threat, ranking among the leading causes of cancer-related deaths. The detection of lung cancer symptoms is challenging, often occurring in advanced stages, leading to a higher mortality rate compared to other cancer types. The disease progresses through various stages, starting with minor tissue involvement and spreading through metastasis across different lung regions. Characterized by uncontrolled cell growth, lung cancer claimed approximately 12,203 lives in 2016, with a higher toll on males (7,130) than females (5,073). Its annual death toll surpasses the combined fatalities of prostate, ovarian, and breast cancers. While smokers face the highest risk, non-smokers can also succumb to lung cancer [1]. The survival rate after a five-year diagnosis hover around 15%, emphasizing the need for effective prediction tools. Data mining techniques, leveraging diagnostic and treatment attributes, present a promising avenue for estimating mortality risk due to lung cancer. The World Health Organization reported over 7.6 million new lung cancer cases annually, projecting a staggering 17 million cases globally by 2030. In 2005, the United States witnessed 1,362,825 new cancer cases, with 571,590 lung cancer patients [2]. As we continue to battle this disease, the fields of healthcare and medical research are embracing advanced technologies to aid in its understanding and management. Data mining tools have emerged as powerful allies in this fight, enabling researchers to uncover hidden insights, infer valuable knowledge, and make accurate predictions. In this study, we explored the dual roles of data mining in lung cancer research: making inferences designed to address specific individual risk factors and generating predictions that can estimate an individual's risk with a high degree of accuracy and reliability

1.1. Research objectives

The primary aim of this study is to leverage advanced data mining techniques to enhance the prediction and understanding of lung cancer risks. Specifically, the research seeks to:

- To find the significant factors and symptoms or associations with the presence of lung cancer among individuals in the dataset.
- To explore the best predictive model to estimate the probability of an individual developing lung cancer based on their demographic and health-related variables, such as age, gender, smoking status, presence of chronic disease, and symptoms (e.g., coughing, shortness of breath, chest pain).
- Improve the accuracy and reliability of predictive models through rigorous validation methods and advanced analytical techniques, aiming to support healthcare professionals in making informed diagnostic decisions.

The remainder of this paper is organized as follows: Section 2 reviews existing literature, contextualizing this study within the broader field of lung cancer research and data mining applications. Section 3 provides an overview of the dataset, including its source, structure, and variables. Section 4 outlines the methodology, detailing data preprocessing steps, analytical techniques, and model evaluation criteria. Section 5 presents the results and discussion, focusing on model performance and subgroup analysis to understand demographic influences. Section 6 concludes the study, summarizing the key findings and their significance. Section 7 highlights the practical implications of integrating predictive models into healthcare systems. Section 8 discusses the limitations of the study, addressing constraints such as sample size and missing covariates. Finally, Section 9 provides directions for future research, including potential enhancements in data collection and model refinement.

2. Related Work

Medical professionals and researchers have long been concerned regarding lung cancer. The majority of recent lung cancer studies have relied on AI. While some research has concentrated on developing methods for diagnosing lung cancer, other studies have sought to identify the disease at an early stage. A model for almost detection and accurate diagnosis of the disease was proposed by Krishnaiah V, Narsimha G, and Subhash Chandra N [3]. This model will aid the doctor in preserving the patient's life. The likelihood of individuals acquiring lung cancer can be predicted using generic symptoms such as age, sex, wheezing, shortness of breath, and pain in the shoulder, chest, or arm.

According to research by Thangaraju, Karthikeyan, and Barkavi [4], smoking is one of the leading causes of lung cancer. The likelihood of acquiring lung cancer increases in direct proportion to the length of time and quantity of cigarettes smoked. Although it most commonly affects people in their 65s and 70s, lung cancer can strike at any age. Cancer of the lung can also develop in young people who have never smoked. Using eleven distinct criteria, Ramachandran and colleagues developed a data-mining-based early detection system for lung cancer [5]. They ran tests on a database with 746 samples, but they didn't say where they got the database. Another group that employed data mining approaches to predict lung cancer risk factors in 2014 was Sowmiya [6]. In order to classify and cluster data, they employed Bayes Trees and Decision Tables. A total of 303 samples were used for the trials.

Research using more modern machine learning techniques, such as decision trees, has shown to be more reliable than the previous methods [7, 8, 9, 10, 11, 12, 13]. Non-Small-Cell Lung Cancer (NSCLC) prognostic models based on neural networks were introduced by Hanai and others [14, 15, 16]. Based on 125 NSCLC patients and 17 possible input risk factors, they constructed their models. Kattan and Bach presented research on the multi-factoral determinants of smokers' lung cancer risk [17]. The researchers assessed the impact of various variables on the lung cancer risk level. While only 0.8% of 51-year-old women who smoked a pack of cigarettes daily for 28 years developed lung cancer, 15% of 68-year-old men who smoked two packs per day for 50 years and still smoked did.

A hybrid neuro-fuzzy system was developed by Manikandan and colleagues to predict the occurrence of lung cancer using eleven symptoms [8]. They drew 163 samples from a larger pool of 271 people, including 221 with medical issues and 50 healthy controls. The symptoms that can be used for the prediction of lung cancer were defined by Arulananth and Bharathi [18, 19]. In order to identify cancer, they separated diagnostic criteria from symptoms. The diagnostic symptoms were specified according to age, sex, smoking status, family history of cancer, radiation exposure, radon exposure, chemical topics exposed, and air pollution. Meanwhile, they identified cancer signs as anorexia, chronic cough, hemoptysis, chest pain, loss of weight, exhaustion, chronic inflammation of the lungs, wheezing, trouble swallowing, and chronic inflammation of the throat.

Senthil and Ayshwaya defined the risk degree of lung cancer based on risk factors using neural networks and evolutionary algorithms in 2018 [10]. Due to the lack of specific symptoms and the small sample size (32 samples), these algorithms were applied to the UCI Global Lung Cancer Database. In 2018, Markaki and colleagues developed a smoking-symptom based clinical risk prediction algorithm for lung cancer [13]. In addition to age, sex, weight, height, number of years smoked, quantity of cigarettes smoked daily, hours spent in contaminated places, frequency of coughing, and number of years since no-smoking was instituted, these factors also played a role. When it came time to classify large databases, other research made good use of sophisticated machine learning methods such random forests and random trees [20].

Raja Ranjan Baitharu, Subhendu Kumar Pani [21] Research shows that lung cancer, an illness characterized by unchecked cell development in lung tissues, is the leading cause of mortality for both men and women. Among the several steps involved in KDD (knowledge discovery in databases), data classification stands out. Quite a few uses might be made of it. Data sets utilized for learning have a significant impact on classifier performance. Improved understanding of the models, faster learning, and improved predictive or descriptive accuracy are the results of this [22]. It also reduces the processing time required to generate the models.

Using lung cancer data in various settings, a comparative study of data categorization accuracy is presented [23]. Some, however, have used image processing methods developed for use in radiation therapy to diagnose lung cancer [24]. Other studies [25, 26, 27] looked at how well the US Military Health System could predict the deaths of patients with NSCLC. In order to construct a reliable model for predicting the likelihood of lung cancer, Cassidy found that additional variables beyond age and smoking were desirable [24].

In order to obtain information regarding the percentage of people who are afflicted by lung cancer and the percentage of fatalities that are related to lung cancer by the year 2024, we conducted research on the history of lung cancer and developed some graphs illustrating the data. In Figure 1, which demonstrates the gender disparities in lung cancer prevalence that are present between males and females in a number of different nations, a global perspective on the impact of this disease is presented. This figure also provides a global perspective on the impact of this disease.

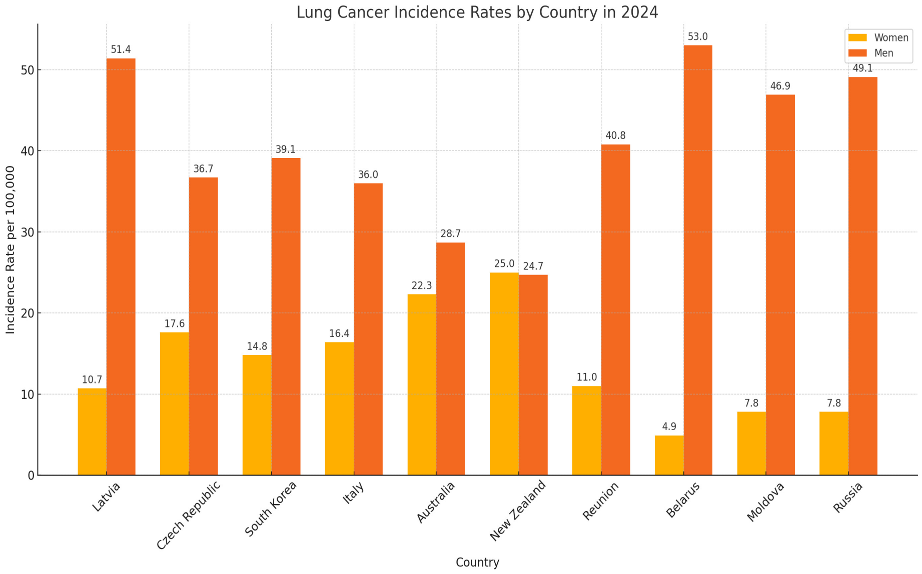

Figure 1. Incidence Rates of Lung Cancer Among Men and Women in 2024, Organized by Country

The lung cancer incidence rates by nation for men and women in 2024 are shown in Figure 1. According to the statistics, different nations have somewhat different incidence rates; Latvia and Belarus have especially high rates for men.

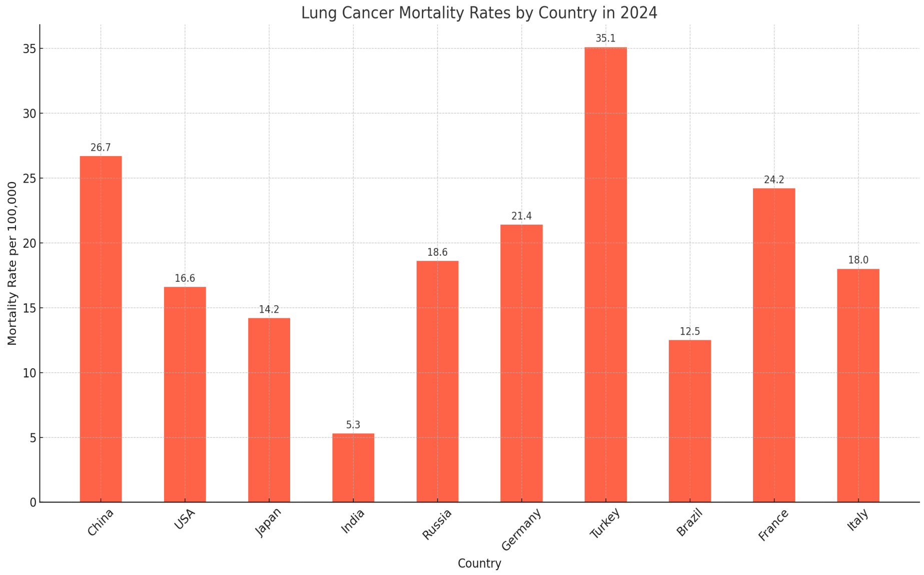

Figure 2. Graphical Representation of the Lung Cancer Mortality Rates by Country in 2024

Figure 2 illustrates the mortality rates due to lung cancer across different countries in the year 2024. The data indicates notable variations in death rates between different nations, with Turkey and China having particularly high rates. Furthermore, we incorporate an extra study that examines an issue with a significant impact on lung cancer. We conducted a study on the rate of change following the onset of the COVID-19 pandemic, specifically focusing on lung cancer. Our findings revealed a significant and dramatic shift in the incidence of lung cancer cases after the pandemic.

We constructed a graph in Figure 3 that included contextual information regarding the factors contributing to the rise in lung cancer rates after the Covid pandemic.

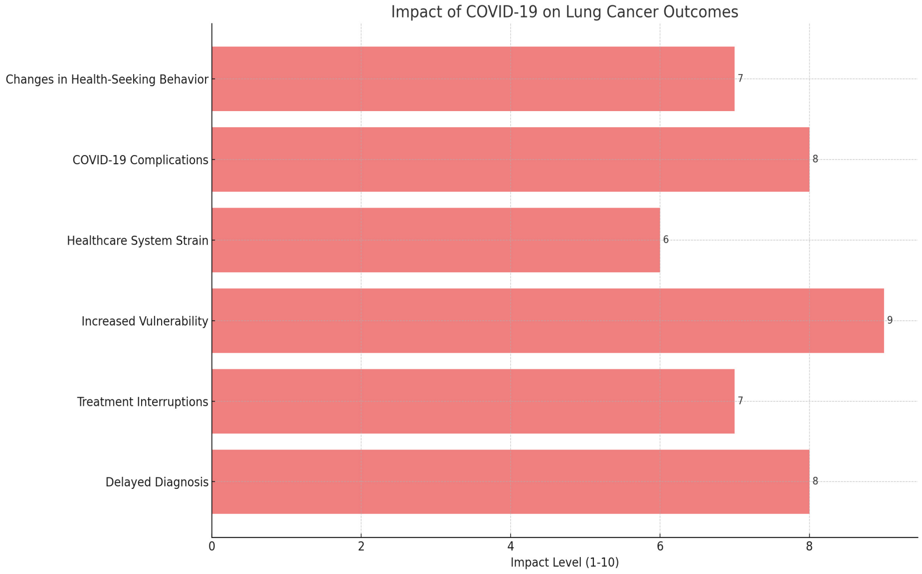

Figure 3. Reasonable Factors Related to the Covid Pandemic and Lung Cancer

The purpose of this research was to develop a methodology for predicting the occurrence of lung cancer using a set of predetermined risk factors. There is also research on the symptoms and how they relate to lung cancer. A robust international prediction tool is constructed by taking into account both domestic and foreign studies and publications [28]. In addition, a global database of 1000 records with 23 variables related to lung cancer is analyzed using machine learning techniques.

3. Data Overview

In this section, we provide a detailed examination of the dataset employed in this study, including its origin, structure, and the specific attributes it encompasses. We discuss the sourcing of the data from Kaggle, the composition of the dataset featuring observations from lung cancer patients and controls, and the variables involved which are crucial for subsequent analyses. This overview is essential for understanding the dataset's capacity to support our research objectives, setting the groundwork for the comprehensive methodological approaches described in Section 4.

3.1. Data Description

The dataset employed in this study, referred to as ”Lung Cancer,” was sourced from Kaggle, provided by the online lung cancer prediction system. It encompasses 309 observations across 16 distinct variables. The dataset captures a range of demographic and health-related characteristics crucial for our analysis. We detail the processes of data cleaning, including handling missing values and normalizing entries, to ensure data integrity. Ethical considerations were strictly followed, with all data anonymized to protect subject privacy. This dataset forms the empirical foundation for our subsequent modeling and analysis, aimed at identifying predictive factors for lung cancer..Detailed information and access to the dataset are available at the following URL https://www.kaggle.com/code/hasibalmuzdadid/lung-cancer-analysis-accuracy-96-4.

3.2. Variables Description

In this subsection, we provide a detailed description of the variables included in the ”Lung Cancer” dataset. Each variable is defined in terms of its role within our analytical framework, specifying whether it is treated as dependent or independent in our models. This clarity is crucial for understanding the interactions between variables and their impact on the predictive accuracy of our models. The characteristics of these variables, including their categorization and measurement scales, are outlined to aid in the comprehension of the dataset’s structure and the rationale behind our methodological choices.

Dependent variable:

Lung cancer (Yes or No).

Independent Variables:

Age, gender, Smoking, yellow fingers, Anxiety, peer pressure, Chronic Disease, Fatigue, Allergy, Wheezing, Alcohol Consumption, Coughing, Shortness of breath, Swallowing Difficulty, Chest pain.

All variables are categorical except age as shown in Table 1.

Table 1. Lung Cancer Data Distribution Frequency Table

Independent Variable	**Numeric Variable**	
Age		
Categorical	**YES (level=1)**	**NO (level=2)**
Gender	162 (Male)	147 (Female)
Smoking	135	174
Yellow Fingers	133	176
Anxiety	155	154
Peer pressure	154	155
Chronic disease	153	156
Fatigue	101	208
Allergy	137	172
Wheezing	137	172
Alcohol Consumption	130	179
Coughing	162	147
Shortness of Breath	111	198
Swallowing Difficulty	164	145
Chest Pain	137	172
Dependent Variable	**YES**	**NO**
Lung Cancer	270	39

Figure 4 provides a frequency distribution of the independent variables, grouped by their correlation with lung cancer presence. This visualization aids in identifying key factors that could potentially influence lung cancer risk.

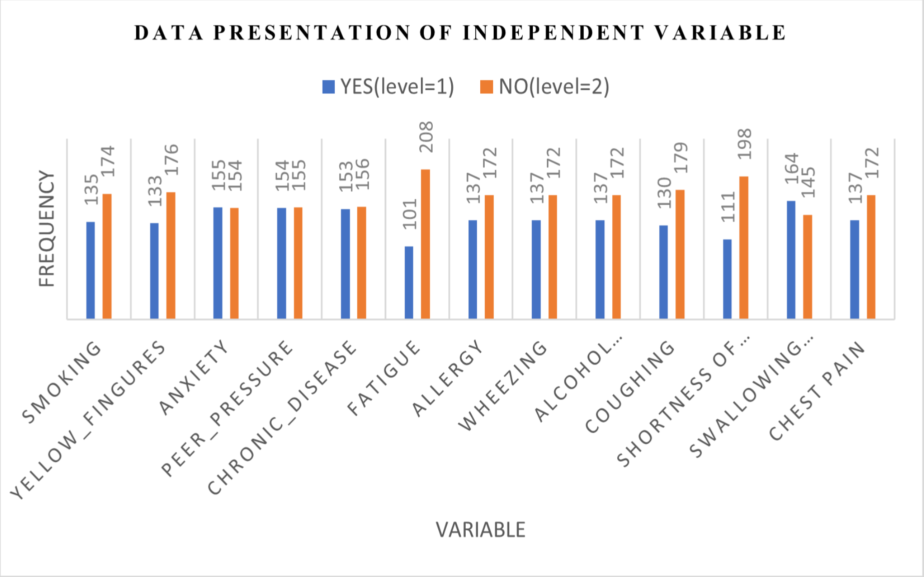

Figure 4. Frequency table of independent variables Grouped by Lung Cancer (Yes and No)

The description of the data is shown in Figure 5. This pie chart visualizes the distribution of the dependent variable—lung cancer status—within our dataset. It effectively illustrates the proportion of subjects diagnosed with lung cancer compared to those without, providing a clear visual representation of the case-control balance in our study. This figure aids in understanding the dataset's composition, which is pivotal for interpreting the predictive analyses that follow.

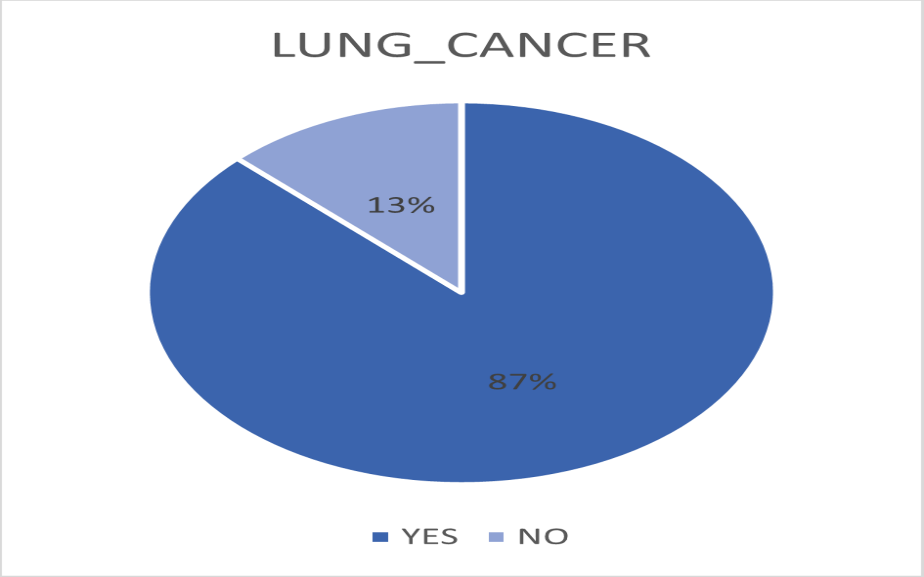

Figure 5. Data presentation of Dependent Variable in Pie Chart

4. Methodology

This section delineates the comprehensive methodologies employed in our study to assess and predict lung cancer risk using data mining techniques. We begin by describing the data collection process, including the sources and characteristics of the dataset utilized. Subsequent subsections detail the analytical approaches adopted, including the selection and justification of various predictive models, the variable selection techniques employed to refine these models, and the statistical methods used to evaluate their performance. This structured approach ensures the robustness and validity of our findings, aiming to optimize both the accuracy and reliability of lung cancer predictions.

4.1. Data Preprocessing

The dataset underwent rigorous preprocessing to ensure its suitability for analysis. Data Cleaning involved handling missing values through multiple imputation, which preserves the integrity of data by replacing missing data with estimated values based on other available data. Data Transformation was necessary to standardize the range of continuous input variables to prevent variables with larger scales from dominating the model's prediction. This included normalization techniques to scale the variables to a unit scale (mean = 0 and variance = 1). Additionally, categorical variables were encoded using one-hot encoding to transform them into a format that could be provided to the ML algorithms to do a better job in prediction.

4.2. Model Selection

The selection of models for this study was strategically based on their suitability to handle specific characteristics of the lung cancer dataset, focusing on predictability, interpretability, and computational efficiency. Logistic Regression was chosen for its ability to provide probabilities for outcomes and its clarity in interpretation, which are indispensable for clinical decision-making. Decision Trees were selected due to their capacity to model complex, non-linear relationships and provide transparent decision rules that can be easily communicated and understood by clinical practitioners. Support Vector Machines (SVM) were utilized for their robust performance in high-dimensional spaces, typical of complex medical datasets. This diverse array of models allows for a comprehensive approach to the prediction task, each bringing unique strengths to tackle the varied facets of the analysis required in predicting lung cancer risk.

To enhance model performance, we applied hyperparameter tuning using grid search. For example, in the random forest model, we optimized the number of trees and the maximum depth of each tree, while in the SVM model, we fine-tuned the kernel function and regularization parameters. Each model was evaluated using cross-validation to ensure generalizability, and performance metrics such as accuracy, precision, recall, and AUC-ROC were used to assess their predictive capabilities.

We fit the logistic regression model to find the association between lung cancer and the corresponding explanatory variable. Since there are 15 explanatory variables, first we applied the three variable selection methods Forward, backward, and mixed stepwise methods.

4.3. Empirical Model of Logistic Regression

The logistic regression model can be formulated as follows:

logit(p) = $\beta0 + \beta1*$Age + $\beta2$Gender + $\beta3$Smoking + $\beta4$yellow fingers + $\beta5$Anxiety +$\beta_6$peer pressure +β_8 Chronic Disease +β_9 Fatigue +β_9 Allergy +β_{10} Wheezing + $\beta11$ Alcohol Consumption + $\beta12$ Coughing + $\beta13$ Shortness of breath + $\beta14$ Swallowing Difficulty + β_{15} Chest pain

In this model, **logit(p)** represents the log odds of having lung cancer. Age, gender, Smoking, yellow fingers, Anxiety, peer pressure, Chronic Disease, Fatigue, Allergy, Wheezing, Alcohol Consumption, Coughing, Shortness of breath, Swallowing Difficulty, Chest pain (2 = present, 1 = absent) or categorical variables representing the presence or absence of specific factors or symptoms. β0, β1, β2, ..., β_{15} are the coefficients associated with each independent variable.

4.4. Variable Selection Method

Variable selection is a procedure used to iteratively add or remove predictor variables from a model based on their statistical significance or contribution to the model's performance. The method involves a series of steps where variables are added or removed one at a time, guided by predefined criteria, until a stopping condition is met. The stepwise variable selection method typically includes three variations: forward selection, backward elimination, and mixed (forward-backward) selection. Here's a description of each variation:

4.4.1. Forward Selection:

The forward selection method starts with an empty model and iteratively adds variables that provide the most significant improvement in the model's performance. At each step, all candidate variables not yet included in the model are evaluated based on predetermined criterion, such as p-values, likelihood ratio tests, or information criteria (e.g., AIC, BIC). The variable with the highest significance is added to the model, and the process continues until no additional variables meet the inclusion criterion.

4.4.2. Backward Elimination:

In contrast to forward selection, backward elimination begins with a model containing all predictor variables and subsequently removes variables that are found to be least significant or contribute the least to the model. At each step, the variable with the highest p-value or the smallest contribution to the model (based on a specified criterion) is eliminated, and the model is refit. The process continues until no remaining variables meet the elimination criterion.

4.4.3. Mixed Selection:

Mixed selection combines forward selection and backward elimination. It starts with an empty model and iteratively adds variables that meet an inclusion criterion as in forward selection. However, after adding a variable, the method also checks if any variables already in the model become insignificant and should be removed, following the backward elimination procedure. This bidirectional process continues until no additional variables meet the inclusion or elimination criteria.

4.5. Model Evaluation and Classification:

To assess the efficacy of our predictive models, we employed a suite of evaluation metrics designed to measure both the accuracy and the reliability of the predictions. Accuracy, a straightforward metric, was calculated to gauge the overall correctness of the models across the dataset. Precision, F-1 Score, and recall were particularly focused upon to evaluate the models' effectiveness in identifying true positive cases of lung cancer, a critical aspect in medical diagnostics. Furthermore, the Area Under the

Receiver Operating Characteristic (AUC-ROC) curve was used as a comprehensive measure, providing insights into the models' performance at various threshold settings, which is vital for balancing sensitivity and specificity in clinical applications. These metrics collectively ensure that our models are not only accurate but also practical and reliable for use in predicting lung cancer, thereby supporting healthcare professionals in making informed diagnostic decisions.

4.5.1. Linear Discriminant Analysis (LDA):

LDA is a statistical method commonly used for classification and dimensionality reduction. It's a supervised learning algorithm that is particularly useful when the goal is to separate two or more classes based on their features. The right pane demonstrates the outcomes for preparing and testing. It likewise shows the quantity of effectively grouped and misclassified samples.

4.5.2. Quadratic Discriminant Analysis (QDA):

QDA is another supervised learning algorithm, similar to Linear Discriminant Analysis (LDA), but with some differences in its assumptions. QDA, like LDA, is used for classification and dimensionality reduction. However, unlike LDA, QDA does not assume that all classes share the same covariance matrix. Instead, it allows for different covariance matrices for each class.

4.5.3. Logistic Regression Analysis:

To represent the probability of a given class or event appearing, statisticians use the strategic model (logistic model). For numerical purposes, a binary logistic model is used when the dependant variable can take on two alternative values, in this case, yes or no, for the lung cancer study.

4.5.4. K-Nearest Neighbour:

Information gathering, and regression are both assisted by the nonparametric technique known as k-nearest neighbors estimation (K-NN). In both instances, the information pertains to the k nearest segment space getting ready models. When using k-NN for either requests or backslides, the result will change accordingly. We employ k=5 and k=10 in our investigation.

4.5.5. Decision Tree:

A decision tree is a supervised machine-learning algorithm used for both classification and regression tasks. It's a tree-like model where an internal node represents a feature (or attribute), the branch represents a decision rule, and each leaf node represents the outcome or the target variable. For a classification tree, we predict that each observation belongs to the most commonly occurring class of training observations in the region to which it belongs. Decision trees are popular because they are easy to understand, interpret, and visualize. In this study, there were 15terminal nodes and used 9 variables. The Tree diagram for the lung cancer study is shown as Figure 6

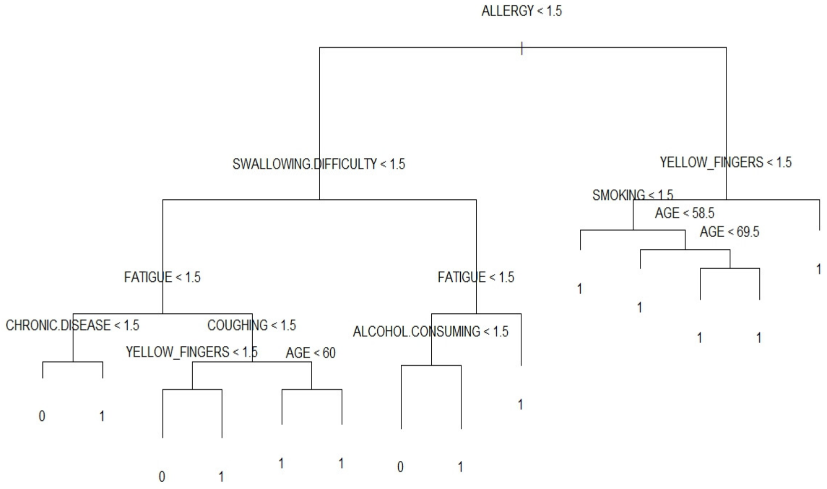

Figure 6. Tree diagram for the lung cancer study

4.5.6. Bagging:

Bagging, or bootstrap aggregation, is an ensemble learning technique for reducing variance in noisy datasets. Bagged data comes from a training set that is randomly selected with replacement, which means that any data point might be chosen multiple times. Following the generation of several data samples, these tentative models are trained separately. The type of task, such as regression or classification, determines whether the average or majority of these predictions produce a more precise estimate.

Figure 7 illustrates the important independent variables identified by the bagging approach for lung cancer data, ranked based on their Mean Decrease Gini values. Variables such as Age, Allergy, and Fatigue exhibit the highest importance, suggesting their strong influence on model predictions. Other factors, including Chronic Disease, Alcohol Consuming, and Yellow Fingers, also contribute significantly, emphasizing their relevance in assessing lung cancer risk. This figure highlights the model's ability to prioritize variables that are critical for accurate classification, providing insights into key predictors of lung cancer.

4.5.7. Random Forest:

Random forests represent an improvement over bagged trees by means of a modification that decorrelates the trees. This is accomplished by a modest alteration. In addition, this brings the variance down even further when we average the trees. The process of building a number of decision trees on bootstrapped training samples is similar to that of bagging. But while these decision trees are being constructed, whenever a split in a tree is being considered, a random selection of m predictors is taken as split candidates from the entire collection of p predictors. This is done in order to ensure reliability. At each split, a new selection of m predictors is selected, and in most cases, we choose $m = \sqrt{p}$. This means that the number of predictors that are taken into consideration at each split is approximately equal to the square root of the total number of predictors. In the course of our research, we took into

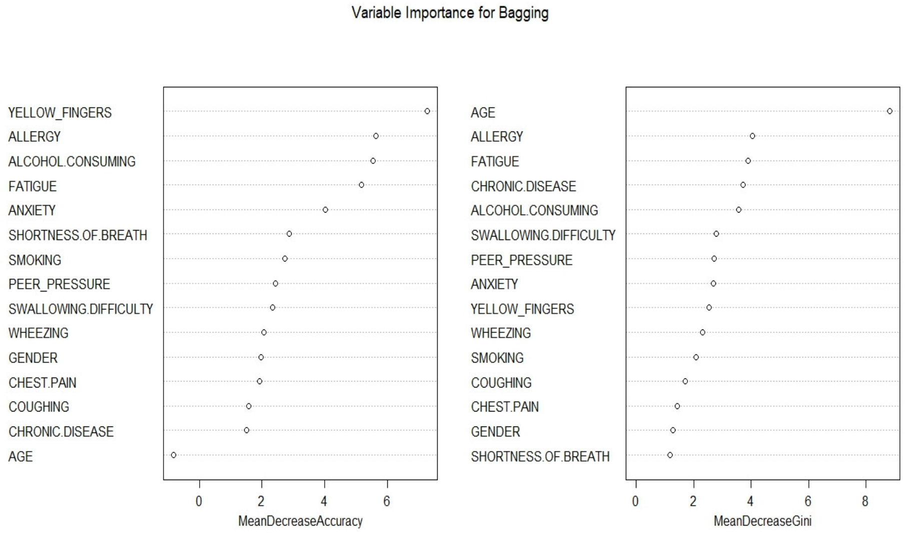

Figure 7. Important Independent Variables for Bagging Approach for Lung Cancer data

consideration a total of four predictors, denoted as $m = \sqrt{15} = 4$.

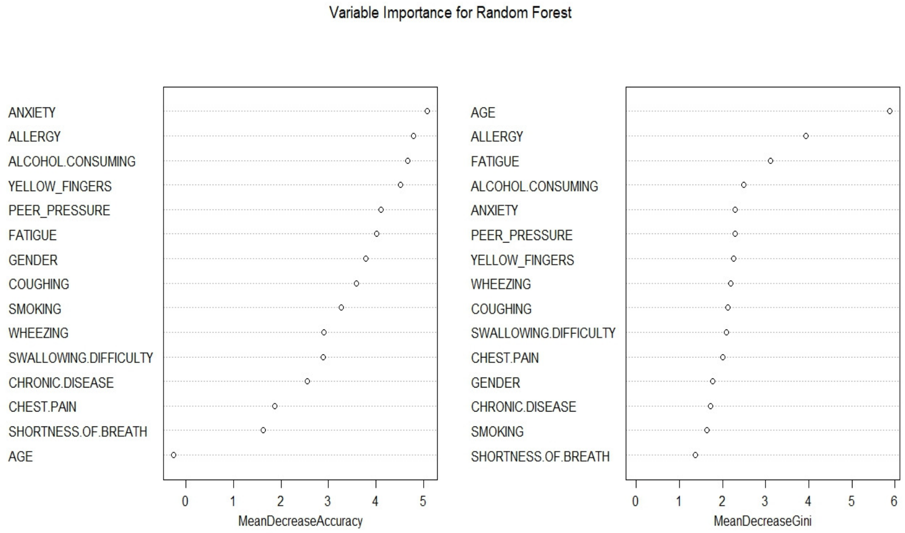

Figure 8. Important Independent Variables for Random Forest Model for Lung Cancer data

Figure 8 highlights the important independent variables identified by the Random Forest model for lung cancer prediction. The variables are ranked by their Mean Decrease Gini scores, which indicate their relative importance in the model's classification decisions. Key variables such as Age, Allergy, and Fatigue are shown to have the highest importance, emphasizing their significant contribution to the predictive accuracy of the model. This visualization reinforces the model's ability to prioritize influential predictors in assessing lung cancer risk.

The most influential variables for lung cancer prediction, ranked by their Mean Decrease Gini values, are summarized in Table 2. Key variables, including Age, Peer Pressure, and Allergy, consistently emerge as critical predictors across both models, highlighting their significant role in enhancing model accuracy. This comparison underscores the alignment between the two ensemble methods in identifying high-impact predictors, offering robust insights into the key factors influencing lung cancer risk.

Table 2. Important variable from Bagging & Random Forest Model

Important Variable for Bagging Model		**Important Variable for Random ForestModel**	
Variable	MeanDecreaseGini	Variable	MeanDecreaseGini
AGE	8.345291	AGE	8.432555
PEER_PRESSURE	6.1758932	ALLERGY	5.478739
ALLERGY	5.7757434	PEER_PRESSURE	4.017965
WHEEZING	3.9144435	YELLOW_FINGERS	3.991659
COUGHING	3.2749532	ALCOHOL.CONSUMING	3.745326
YELLOW_FINGERS	2.8969782	COUGHING	3.2118
ALCOHOL.CONSUMING	2.8480417	SWALLOWING.DIFFICULTY	3.135628
SWALLOWING.DIFFICULTY	1.9182844	FATIGUE	3.05841
CHEST.PAIN	1.5382041	WHEEZING	2.79263
ANXIETY	1.300547	ANXIETY	2.638624
FATIGUE	1.1478747	CHRONIC.DISEASE	2.375772
CHRONIC.DISEASE	1.0523739	CHEST.PAIN	2.18453
GENDER	0.8665353	GENDER	2.086803
SMOKING	0.8283375	SMOKING	1.940928
SHORTNESS.OF.BREATH	0.4741762	SHORTNESS.OF.BREATH	1.83699

4.5.8. Support Vector Machine:

For prediction, regression, and classification the most prominent method employed is SVM. It classifies the input data set by introducing a boundary called a hyperplane that separates the dataset into two parts. The favorable asset of SVM is, that SVM is a data-driven approach and is feasible without a hypothetical scheme that produces an accurate classification. Particularly when the size of the sample is small. SVMs are broadly used for classification when the datasets are biomarkers, to predict and diagnose cancer, neurological, and cardiology diseases.

4.6. Analytical Framework of Analysis:

This subsection delineates the analytical framework adopted for this study, detailing the structured approach used to transform raw data into actionable insights. We outline the sequence of analytical techniques—from data preprocessing and variable selection to the application of sophisticated statistical models. This framework is designed to ensure that the analysis is both robust and replicable, providing clear insight into how data-driven predictions are made and the rationale behind the selection of specific methodologies for optimizing model performance. This systematic approach not only supports the validity of our findings but also enhances the reproducibility of the research, contributing to ongoing efforts in the field of predictive analytics in healthcare. The analytical framework of the analysis has shown below in Figure 9.

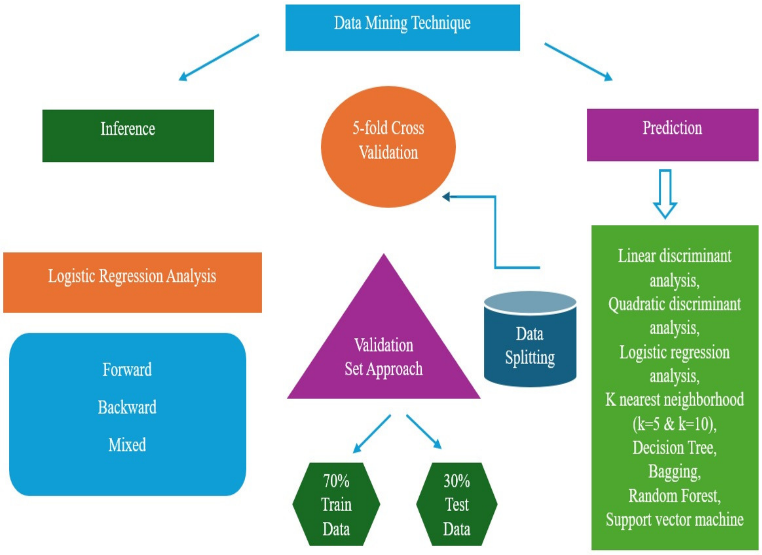

Figure 9. Illustration of the analysis process, from data collection to result interpretation

5. Results and Discussion

In this research, a cohort of 309 individuals was involved, comprising 270 with lung cancer and 39 without lung cancer. The demographic distribution included 162 males and 147 females. We applied the logistic Regression analysis to measure the association of predictor variables with lung cancer for inference. We conducted Linear Discrimination Analysis, Quadratic Discriminant Analysis, Logistic Regression Analysis, KNN, Decision Tree, Bagging, Random Forest, and Support Vector Machine for prediction. In addition to that, to explore whether model accuracy varied across different demographic groups, a subgroup analysis was conducted based on age, gender, and smoking status. The results revealed that the model's accuracy was higher among older individuals and those with a smoking history. In contrast, the accuracy decreased slightly for younger participants and non-smokers. This indicates that demographic factors may influence the model's predictive performance, suggesting the need for further model refinement to improve its effectiveness across all subgroups.

5.1. Logistic Regression Analysis for Inference

Before fitting the final logistic regression model, we apply the forward, backward, and stepwise variable selection methods as we can find a subset of variables that are important for lung cancer.

Table 3 represents the significant coefficient for three variable selection methods. Although the backward and stepwise variable selection method gives the same result and lower AIC, we use the mixed method as it is a bidirectional variable selection model, we proceed to fit the mixed model because of the lower AIC = 116.6, among the three methods.

5.2. Result of Coefficient for Logistic Regression Analysis

Table 4 represents the result of the final logistic regression model. We observe that Allergy, Peer pressure, Swallowing Difficulty, Smoking, Chronic Disease, Alcohol Consumption yellow fingers, Fa-

Table 3. Result of Backward, Forward, and Mixed Variable Selection Method

Backward		Forward		Mixed	
AIC	116.6	AIC	123.6	AIC	116.6
Variable	Coefficient	Variable	Variable	Variable	Coefficient
Intercept	27.323	Intercept	30.656	Intercept	27.323
Smoking	-1.454	Smoking	-1.776	Smoking	-1.454
Yellow_fingers	-1.741	Yellow_fingers	-1.376	Yellow_fingers	-1.741
Peer_pressure	-1.874	Peer_pressure	-1.731	Peer_pressure	-1.874
Chronic.disease	-2.695	Chronic.disease	-3.192	Chronic.disease	-2.695
Fatigue	-2.870	Fatigue	-2.870	Fatigue	-2.870
Allergy	-1.834	Allergy	-1.646	Allergy	-1.834
Alcohol.consuming	-1.751	Alcohol.consuming	-1.409	Alcohol.consuming	-1.751
Swallowing.difficulty	-3.427	Swallowing.difficulty	-3.122	Swallowing.difficulty	-3.427
Coughing	-3.065	Coughing	-3.311	Coughing	-3.065
		Gender	0.5261		
		Age	-0.022		
		Anxiety	-0.888		
		Wheezing	-0.966		
		Shortness of Breath	0.729		
		Fatigue	-3.070		

tigue, and Coughing are significant as they have p-values is less than 0.05 but have negative associations with lung cancer. Such as the coefficient of Allergy is -1.834 that means when holding other variables constant, for people who have Allergy, the probability of lung cancer will be lower than for those who do not have Allergy.

5.3. *Classification Analysis for prediction*

To ensure robust model evaluation, we adopted distinct strategies for data partitioning. We employed two approaches for analysis such as k-fold cross-validation approach with k=5 and the validation set approach. Conversely, for a validation set approach, the lung cancer data were randomly divided into (70:30) % of the data set, with (217 cases) assigned to the training dataset and the other half (92 cases) to the test dataset. This diverse methodology in data partitioning aimed to assess the performance and generalizability of the models across different statistical data mining techniques.

Table 5 presents the Confusion matrix for different models representing True Positive Rate and True Negative Rate using the Validation set approach. Examining the confusion matrix, we observe that the sensitivity and specificity of the test are maximized at 0.93 and 0.71 for Random Forest. This implies that when applied to a group of 100 individuals with lung cancer, the test accurately identifies 95 of them as positive, on the other hand, the specificity of the test is optimal at 0.71 this signifies that when the test is administered to a group of 100 individuals without lung cancer, it correctly identifies 71 of them as negative.

Table 4. Result of Coefficient for Logistic Regression Analysis

Variable	Estimate	Std. Error	Z-value	Pr(>|z|)
(Intercept)	27.32	4.653	5.871	0.000***
Allergy	-1.834	0.723	-2.534	0.011*
Peer_pressure	-1.874	0.637	-2.942	0.003**
Swallowing.difficulty	-3.427	0.979	-3.498	0.000***
Smoking	-1.453	0.653	-2.224	0.026*
Chronic.disease	-2.695	0.761	-3.537	0.000***
Alcohol.consuming	-1.751	0.711	-2.461	0.013*
Yellow_fingers	-1.741	0.639	-2.722	0.006**
Fatigue	-2.871	0.671	-4.272	0.000***
Coughing	-3.065	0.836	-3.663	0.000***

Table 5. Confusion matrices for different models representing True Positive Rate and True Negative Rate using Validation set approach

Model	LDA			QDA			Logistic		
TestLung_Cancer	Positive	Negative	Total	Positive	Negative	Total	Positive	Negative	Total
Yes	73	6	79	73	6	79	73	7	80
No	4	9	13	4	9	13	4	8	12
Total	77	15	92	77	15	92	77	15	92
TPR	0.92			0.92			0.91		
TNR	0.69			0.69			0.67		
Model	KNN(k=5)			KNN(k=10)			Decision Tree		
TestLung_Cancer	Positive	Negative	Total	Positive	Negative	Total	Positive	Negative	Total
Yes	2	2	4	1	1	2	72	6	78
No	76	12	88	77	13	90	5	9	14
Total	78	14	92	78	14	92	77	15	92
TPR	0.5			0.5			0.78		
TNR	0.13			0.14			0.64		
Model	Bagging			Random Forest			Support Vector Machine		
TestLung_Cancer	Positive	Negative	Total	Positive	Negative	Total	Positive	Negative	Total
Yes	73	6	79	73	5	78	73	7	80
No	4	9	13	4	10	14	4	8	12
Total	77	15	92	77	15	92	77	15	92
TPR	0.92			0.93			0.91		
TNR	0.69			0.71			0.66		

5.4. Comparison Results of Accuracy Rate of Predicted Model

Table 6, Table 7 and Figure 10 represented the result of the accuracy rate of the different predicted models using the validation set approach and 5-fold cross-validation. The evaluation of model performance, measured in terms of accuracy, revealed notable distinctions among the developed models. In 5-fold cross-validation, the Logistic Regression Model exhibited the highest accuracy at 93.54%, leading the pack. Following closely, the LDA demonstrated an accuracy of 92.56%, while The Random Forest and Support Vector Machine models achieved a commendable accuracy of 91.23 and 91.29 respectively. The bagging models performed well with an accuracy of 90.90%. Subsequently, the QDA attained an accuracy of 89.97%. The Decision Tree models secured an accuracy of 87.33%. Further down the accuracy spectrum, the Knn-10 registered 17.74%, and KNN-5 concluded with an accuracy of 16.12%. During the validation process, the Random Forest model exhibited the highest accuracy, reaching a peak of 90.21%. Following closely, LDA, QDA, and Bagging secured the second-highest accuracy, attaining a score of 89.13%. Logistic regression analysis and Support Vector Machine claimed the third position with an accuracy of 88.04%. Meanwhile, the Decision Tree model achieved an accuracy of 86.95%, and KNN-10 and KNN-5 lagged with accuracies of 15.21% and 14.13%, respectively. This outcome indicates that the logistic regression model successfully classified the lung cancer status for 93.54% of the patients based on the provided predictors. This outcome indicates the model's potential for effectively identifying cases of lung cancer based on the provided predictors.

Table 6. Comparison Results of Accuracy Rate of Predicted Model

	Accuracy Rate (%)	
Data Mining Predictive Model	**Validation Set Approach**	**5-fold Cross Validation**
Logistic Regression Analysis	88.04	93.54
Linear Discriminant Analysis	89.13	92.56
Quadratic Discriminant Analysis	89.13	89.97
K- Nearest Neighbor (k=5)	14.13	16.12
K- Nearest Neighbor (K=10)	15.21	17.74
Decision Tree	86.95	87.33
Bagging	89.13	90.90
Random Forest	90.21	91.23
Support vector Machine	88.04	91.29

Table 7. Comparison Results of performance metrics of Predicted Model

Data Mining Predictive Model	**Precision (%)**	**Recall(%)**	**F1 Score (%)**	**AUC-ROC (%)**
Logistic Regression Analysis	93.33	96.25	91.53	98.25
Linear Discriminant Analysis	98.67	92.5	95.48	95.65
Quadratic Discriminant Analysis	93.90	96.25	95.06	94.67
K- Nearest Neighbor (k=5)	88.50	96.25	92.21	92.12
K- Nearest Neighbor (K=10)	87.64	97.5	92.30	92.20
Decision Tree	91.35	92.5	91.92	91.75
Bagging	90.58	92.5	95.68	92.75
Random Forest	92.68	95.0	93.82	93.85
Support vector Machine	88.76	98.75	93.49	94.75

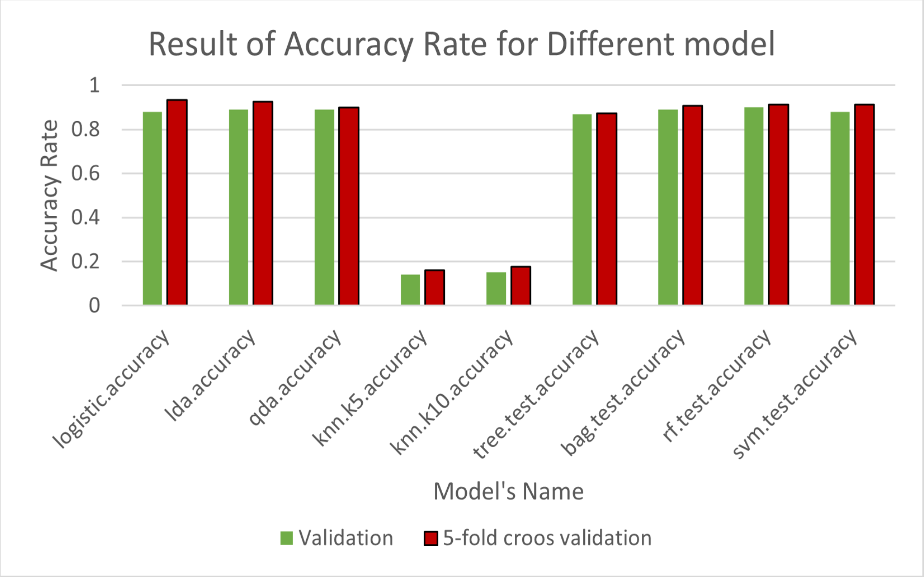

Figure 10. Comparison Results of Accuracy Rate of Predicted Model

The results of this study, incorporating metrics such as accuracy, precision, recall, F1 score, and AUC-ROC, provide a comprehensive evaluation of the predictive models' clinical utility for early lung cancer detection. Logistic Regression and Random Forest, in particular, demonstrate robust performance across these metrics, offering reliable identification of high-risk individuals. The high precision ensures that false positives are minimized, reducing unnecessary diagnostic follow-ups, while strong recall highlights the models' capacity to correctly identify true positive cases, critical for early intervention. The F1 score balances precision and recall, reflecting the models' effectiveness in managing the trade-off between missed diagnoses and over-diagnosis. Moreover, the high AUC-ROC scores indicate that these models maintain strong discriminatory power across various thresholds, which is essential for adapting to different clinical screening protocols.

By integrating these predictive models into clinical workflows, healthcare systems can optimize lung cancer screening efforts. For instance, models with high recall and precision could be used to triage patients for further testing, ensuring that resources are directed toward those most at risk while avoiding the burden of unnecessary testing. Additionally, by providing data-driven insights into individual risk factors, these models support personalized screening and treatment decisions, potentially leading to earlier detection, more timely interventions, and ultimately improved patient outcomes. These results underscore the models' practical value in enhancing the effectiveness of lung cancer screening and treatment strategies.

6. Conclusion

This study endeavors to shed light on the obscure landscape of lung cancer research by employing data mining techniques for both inference and prediction. We conduct a comprehensive comparison of multiple predictive models, which provides valuable insights into their relative performance for lung cancer prediction. Additionally, we incorporate less commonly studied predictors such as *yellow fingers*, *peer pressure*, and *swallowing difficulty*, shedding new light on their roles in lung cancer risk.

We also employ rigorous validation methods, including both a validation set approach and 5-fold cross-validation, ensuring that our results are generalizable. By addressing class imbalance with SMOTE and undersampling techniques, we improve model performance in real-world scenarios.

Our analysis demonstrates the efficacy of various statistical and machine learning models in predicting lung cancer, with logistic regression consistently showing superior accuracy, reaching as high as **93.54%**. This finding underscores the model's potential in clinical diagnostics, suggesting that logistic regression could significantly aid health workers and medical personnel in the early detection and prediction of lung cancer. The study identifies key predictive variables such as smoking status, chronic diseases, and demographic factors, which are instrumental in enhancing the understanding and prediction of lung cancer risks. These insights could be pivotal for public health strategies and clinical interventions, aimed at improving patient outcomes through early diagnosis.

Furthermore, the successful application of data mining in this research highlights the transformative potential of these techniques in medical research, opening avenues for more personalized and timely healthcare solutions. Looking ahead, the inclusion of additional covariates like genetic markers and environmental factors in future studies could further refine the accuracy of predictive models. Longitudinal studies could also provide valuable data on the progression of lung cancer and the evolving efficacy of prediction models over time. By continuously refining these models and integrating new data, the field can move closer to the broader goals of precision medicine and improved patient care.

7. Practical Implications for Healthcare Systems

The integration of predictive models, such as the ones evaluated in this study, into healthcare systems holds significant potential for improving early lung cancer detection and screening. These models could assist clinicians in identifying high-risk individuals earlier, enabling timely interventions that may improve patient outcomes. By automating risk assessment and triaging patients for further screening, these models could reduce the burden on healthcare professionals, streamline diagnostic processes, and optimize resource allocation, particularly in high-demand settings. Moreover, predictive models can be used to personalize treatment plans, tailoring interventions based on the unique risk profiles of individual patients. However, successful implementation would require thorough validation across diverse populations to ensure their reliability and effectiveness in real-world settings.

8. Limitations of the Study

This study acknowledges certain limitations that may impact the generalizability of its findings.

- The study acknowledges the constraint of a relatively small dataset, comprising 270 lung cancer patients and 39 non-lung cancer patients. This limited sample size may impact the generalizability of the findings to broader populations, necessitating caution in extrapolating the results.

- While the study explores various risk factors associated with lung cancer, the inclusion of additional covariates, such as genetic predispositions or environmental exposures, could further enrich the predictive models. The absence of certain critical covariates may limit the comprehensiveness of the risk assessment.

- The cross-sectional nature of the dataset captures information at a single point in time, preventing the analysis of changes in risk factors or disease progression over time. Future research utilizing longitudinal data could provide deeper insights into how risk factors evolve and impact lung cancer outcomes, further improving predictive accuracy.

9. Future Research

- Future investigations can delve into incorporating advanced biomarkers, genomic data, or molecular signatures to enhance the models' predictive capabilities.
- Conducting longitudinal studies to observe changes in risk factors over time could provide valuable insights into the dynamic nature of lung cancer development. This longitudinal
- perspective may capture evolving factors influencing both the initiation and progression of the disease.
- The practical implementation of data mining predictions in clinical settings should be a focal point of future research. Assessing the feasibility, acceptance, and impact of incorporating these models into routine clinical practice will be vital for their real-world applicability and impact on patient outcomes.

References

1 Global cancer burden growing, amidst mounting need for services (who.int)

2 Cancer (who.int).

3 Krishnaiah, V., Narsimha, G., & Chandra, N. S. (2013). Diagnosis of lung cancer prediction system using data mining classification techniques. *International Journal of Computer Science and Information Technologies*, 4(1), 39-45.

4 Thangaraju, P., Barkavi, G., & Karthikeyan, T. (2014). Mining lung cancer data for smokers and non-smokers by using data mining techniques. *International Journal of Advanced Research in Computer and Communication Engineering*, 3(7), 7622-7626.

5 Ramachandran, P., Girija, N., & Bhuvaneswari, T. (2014). Early detection and prevention of cancer using data mining techniques. *International Journal of Computer Applications*, 97(13).

6 Sowmiya, T., Gopi, M., Begin, M., & Robinson, L. T. (2014). Optimization of lung cancer using modern data mining techniques. *International Journal of Engineering Research*, 3(5), 309-314.

7 Christopher, T., & Banu, J. J. (2016). Study of classification algorithm for lung cancer prediction. *International Journal of Innovative Science, Engineering & Technology*, 3(2), 42-49.

8 Manikandan, T., Bharathi, N., Sathish, M., & Asokan, V. (2017). Hybrid neuro-fuzzy system for prediction of lung diseases based on the observed symptom values. *J Chem Pharm Sci ISSN*, 974, 2115.

9 Raihen, M. N., & Akter, S. (2023). Forecasting Breast Cancer: A Study of Classifying Patients' Post-Surgical Survival Rates with Breast Cancer. *Journal of Mathematics and Statistics Studies*, 4(2), 70-78.

10 Senthil, S., & Ayshwarya, B. (2018). Lung cancer prediction using feed forward back propagation neural networks with optimal features. *International Journal of Applied Engineering Research*, 13(1), 318-325.

11 Durga, S., & Kasturi, K. (2017). Lung disease prediction system using data mining techniques. *J Adv Res in Dynamical and Contr Sys*, 9(5), 62-66.

12 Raihen, M. N., & Akter, S. (2024). Sentiment analysis of passenger feedback on US airlines using machine learning classification methods. *World Journal of Advanced Research and Reviews*, 23(1), 2260-2273.

13 Markaki, M., Tsamardinos, I., Langhammer, A., Lagani, V., Hveem, K., & Røe, O. D. (2018). A validated clinical risk prediction model for lung cancer in smokers of all ages and exposure types: a HUNT study. *EBioMedicine*, 31, 36-46.

14 Raihen, M. N., & Akter, S. (2024). Comparative Assessment of Several Effective Machine Learning Classification Methods for Maternal Health Risk. *Computational Journal of Mathematical and Statistical Sciences*, 3(1), 161-176.

15 Raihen, M. N., & Akter, S. (2024). Prediction modeling using deep learning for the classification of grape-type dried fruits. *International Journal of Mathematics and Computer in Engineering*.

16 Hanai, T., Yatabe, Y., Nakayama, Y., Takahashi, T., Honda, H., Mitsudomi, T., & Kobayashi, T. (2003). Prognostic models in patients with non-small-cell lung cancer using artificial neural networks in comparison with logistic regression. *Cancer science*, 94(5), 473-477.

17 Bach, P. B., Kattan, M. W., Thornquist, M. D., Kris, M. G., Tate, R. C., Barnett, M. J., ... & Begg, C. B. (2003). Variations in lung cancer risk among smokers. *Journal of the National Cancer Institute*, 95(6), 470-478.

18 Bharathi, H., & Arulananth, T. S. (2017). A review of lung cancer prediction system using data mining techniques and self organizing map (SOM). *International Journal of Applied Engineering Research*, 12(10), 2190-2195.

19 Ankrah, B. N., Brew, L., & Acquah, J. (2024). Multi-Class Classification of Genetic Mutation Using Machine Learning Models. *Computational Journal of Mathematical and Statistical Sciences*, 3(2), 280-315.

20 Mishra, A. K., & Ratha, B. K. (2016). Study of random tree and random forest data mining algorithms for microarray data analysis. International *Journal on Advanced Electrical and Computer Engineering*, 3(4), 5-7.

21 Weaver, H., & Coonar, A. S. (2017). Lung cancer: diagnosis, staging and treatment. *Surgery (Oxford)*, 35(5), 247-254.

22 Kandel, M. A., Rizk, F. H., Hongou, L., Zaki, A. M., Khan, H., & El-Kenawy, E. S. M. (2023). Evaluating the Efficacy of Deep Learning Architectures in Predicting Traffic Patterns for Smart City Development. *Full Length Article*, 6(2), 26-6.

23 Enriko, I. K. A., Mahuzza, T. M., Purnama, S. I., & Gunawan, D. (2022, December). Comparative Study of Lung Disease Prediction System Using Top 10 Data Mining Algorithms with Real Clinical Medical Records. In *First Mandalika International Multi-Conference on Science and Engineering 2022, MIMSE 2022 (Informatics and Computer Science)(MIMSE-IC-2022)* (pp. 269-281). Atlantis Press.

24 Cassidy, A., Duffy, S. W., Myles, J. P., Liloglou, T., & Field, J. K. (2007). Lung cancer risk prediction: a tool for early detection. *International journal of cancer*, 120(1), 1-6.

25 Abdollahzadeh, B., Khodadadi, N., Barshandeh, S., Trojovský, P., Gharehchopogh, F. S., El-kenawy, E. S. M., ... & Mirjalili, S. (2024). Puma optimizer (PO): A novel metaheuristic optimization algorithm and its application in machine learning. *Cluster Computing*, 1-49.

26 Raihen, M. N., & Tran, J. (2024). Optimizing reinforcement learning in complex environments using neural networks. *International Journal of Science and Research Archive*, 12(2), 2047-2062.

27 Saii, M., & Mayya, A. (2015). Lung detection and segmentation using marker watershed and laplacian filtering. *International Journal of Biomedical Engineering and Clinical Science*, 1(2), 29-42.

28 Jeong, C. W., Jeong, S. J., Hong, S. K., Lee, S. B., Ku, J. H., Byun, S. S., ... & Lee, S. E. (2012). Nomograms to predict the pathological stage of clinically localized prostate cancer in Korean men: comparison with western predictive tools using decision curve analysis. *International journal of urology*, 19(9), 846-852.

SASAR

Open Access Journal

Disclaimer/Publisher's Note: The content in all publications reflects the views, opinions, and data of the respective individual author(s) and contributor(s), and not those of the scientific association for studies and applied research (SASAR) or the editor(s). SASAR and/or the editor(s) explicitly state that they are not liable for any harm to individuals or property arising from the ideas, methods, instructions, or products mentioned in the content.