Early Lung Cancer Risk Prediction using Ensemble Machine Learning Models with SHAP for Explainability

Submitted by

Nosrat Jahan Mithila

ID: 213-35-768

Department of Software Engineering

Daffodil International University

Supervised by

Dr. Md. Fazla Elahe

Assistant Professor & Associate Head

Department of Software Engineering

Daffodil International University

This thesis paper has been submitted in fulfillment of the requirements for the degree of

Bachelor of Science in Software Engineering

Summer 2025

APPROVAL

This thesis titled on "**Early Lung Cancer Risk Prediction Using Ensemble Machine Learning Models with SHAP for Explainability**", submitted by **Nosrat Jahan Mithila (ID: 213-35-768)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

Fazla Elahe

Chairman

Dr. Md. Fazla Elahe
Assistant Professor & Associate Head
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Marzia

Internal Examiner 1

Dr. Marzia Ahmed
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

13.09.2025

Internal Examiner 2

Dr. Shabnom Mustary
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

09.25

External Examiner

Mohammod Abul Kashem
Professor
Department of Computer Science and Engineering
Dhaka University of Engineering & Technology, Gazipur.

SUPERVISOR's DECLARATION

I hereby declare that I have checked this thesis and in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Bachelor of Science.

(Supervisor's Signature)

Full Name : Dr. Md. Fazla Elahe

Position : Assistant Professor & Associate Head

Date : 15 September 2025

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Daffodil International University or any other institution.

Mithila.

(Student's Signature)

Full Name : Nosrat Jahan Mithila

ID Number : 213-35-768

Date : 15 September 2025

ACKNOWLEDGMENT

I want to sincerely thank Almighty Allah and my parents for their unwavering support during my pursuit of a bachelor's degree and the finishing of this thesis. First and foremost, I would like to express my gratitude to Dr. Md. Fazla Elahe, my supervisor and the assistant professor and associate head of Daffodil International University's software engineering department. His unstinting support and priceless advice have been essential to finishing my study, "Early Lung Cancer Risk Prediction using Ensemble Machine Learning Models with SHAP for Explainability" I would like to express my sincere gratitude to all of my regarded teachers who have guided me during my academic career. Their commitment and expertise have had a significant impact on my education. I am pleased with Daffodil International University for offering the tools and assistance I required to complete my research. Dr. Md. Fazla Elahe's oversight and assistance have been particularly crucial. Finally, I want to express my gratitude to all of my Parents, DIU friends, seniors and classmates for their friendship, cooperation, and support. I was able to reach this goal because of their encouragement.

ABSTRACT

The world is still concerned about lung and pulmonary tissue cancer, which is one of the main reasons for cancer death. This is primarily because of late-stage detection and incorrect diagnosis. The time of diagnosis directly affects treatment success and survival rates. This study aims to ascertain whether it is practical to advance machine learning for early lung cancer risk prediction using K Nearest Neighbors (KNN), Decision Trees, Support Vector Machines, and Logistic Regression models. A stacking ensemble model has been developed for this purpose, which combines multiple forecaster models to increase accuracy. When the model was tested on two separate datasets, it achieved accuracy scores of 99.9% and 98% on Dataset-1 and Dataset-2, respectively, surpassing the other models decisively. Additionally, there was also great predictive success. Predictive performance was transformed by imputing the models with deep learning models that rely on SHAP (SHapley Additive Explanations) in order to improve model transparency and identify risk predictors. The models also proved that and reinforces the ensemble models capabilities on accuracy and also model transparency, serving as a supportive resource in the clinical and clinical settings in order to improve lung cancer actionable decision and diagnosis.

Keywords — Lung Cancer, Risk Prediction, Ensemble Learning, Stacking Model, Machine Learning, Explainable AI, SHAP, Early Diagnosis

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
AUROC	Area Under the Receiver Operating Characteristic Curve
CT	Computed Tomography
EDA	Exploratory Data Analysis
FN	False Negative
FP	False Positive
KNN	K-Nearest Neighbors
ML	Machine Learning
NSCLC	Non-Small Cell Lung Cancer
RF	Random Forest
ROC	Receiver Operating Characteristic
SCLC	Small Cell Lung Cancer
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
TN	True Negative
TP	True Positive

LIST OF APPENDICES

CHAPTER 1

1.0 Introduction:

When aberrant lung cells proliferate uncontrollably and develop into a tumor, lung cancer results. Primary lung cancer is the term used when the cancer first appears in the lungs. On the other hand, it is referred to as additional or metastatic lung cancer if it starts in another area of the body and moves to the lungs. Because it kills so many people every year, lung cancer is one of the most lethal cancers worldwide. Small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC) are the two primary forms of lung cancer. The more typical kind of lung cancer, NSCLC, accounts for around 85% of all cases. Among its subtypes are big cell carcinoma, squamous cell carcinoma, and adenocarcinoma. The other kind, SCLC, spreads more fast and is more challenging to cure. Smoking, secondhand smoke inhalation, exposure to toxic gases or chemicals, advanced age, lung conditions, and family history can all contribute to lung cancer. Early disease detection is frequently challenging because symptoms typically manifest later [1].

Lung cancer remains a global health burden and causes cancer deaths. An estimated 2.2 million people were reported to have been newly infected worldwide in 2020, with approximately 1.8 million deaths also being reported during the same period as per World Health Organisation (WHO). The disease is particularly common in areas where smoking and air pollution are prevalent, resulting in lung cancer being one of the top three most diagnosed cancer types. As seen in the present study, middle-aged and elderly men suffer most frequently, which may be attributed to the cumulative insult of smoking and continued exposure to deleterious environmental agents. Crucially, the prevalence of lung cancer is such that many cases are diagnosed only at an advanced stage. As a result, late diagnosis reduces treatment options and increases the chance of death, making lung cancer a significant health challenge that requires urgent attention for early detection and better management [2].

By offering timely medical intervention, lowering treatment expenses, and raising survival rates, early lung cancer prediction can significantly improve patient outcomes. Patients' quality of life is improved when lung cancer is discovered early on because less invasive treatments and improved management techniques are possible. Early diagnosis may be revolutionized by the use of predictive analytics that are based on readily available patient data, such as symptoms, known risk factors (such as lifestyle), and medical history, especially in settings with limited resources [3].

Usually performed, tests for detecting lung cancer such as imaging (X-rays, CT scans), and tissue biopsies, can be quite expensive. A professional’s skillset, from specialized equipment to interaction with technical gadgets and even result formulation, is intricate. These evaluations appear, and as such are, within the reach and expensive to the majority, especially in poorer countries. Also, this type of prognosis can extend to days, and even to weeks, which is a big loss of time when it comes to commencing the prescription. These cumulative factors create a challenge as a defined treatment approach can pose challenges as well as limitations. Sticking to the conventional methods of diagnosis tends to procrastinate time-sensitive parameters such as result articulation, analysis, and the overall opposing wellness of the patient [4].

At this juncture, machine learning (ML) stands poised as one of the powerful tools to generate predictions in the healthcare sector, primarily driven by the need to analyze complex and large datasets. The proprietary algorithms are able to find and predict risks and diseases by accessing and analyzing the patient’s data without any pattern and interrelation. The medical field has embraced tools, including regression, decision trees, and support vector machines, to build highly predictive models on the basis of their strengths in handling structured data. In spite of this, ML in healthcare continues to face challenges with the complex interactions of non-linear and diverse patient populations across multiple risk factors. The lack of transparency of models is major hurdle, as the lack of proven logic in their so-calledThe "black box" aspect makes it challenging for clinicians to trust their reasoning. In addition, the intricacies and multi-dimensional nature of risk factors associated with lung cancer cannot be adequately represented by the use of single models, thereby illustrating the need for advanced solutions. [5].

By combining many base learners, ensemble machine learning models such as Decision Tree, Support Vector Classifier, and K-Nearest Neighbors have been developed to increase prediction accuracy and reduce overfitting. Additionally, these models are renowned for their durability, scalability, and capacity to identify ever-more intricate patterns in big datasets. Interpretability is one of the main issues facing ensemble models. In order to embrace and use AI-driven medical systems, clinicians and other stakeholders must comprehend the logic underlying the predictions. Although SHAP (SHapley Additive exPlanations) is an explainable AI method that is independent of models, that helps to bridge this gap by assigning each feature a score of contribution to a prediction, complex model outputs still need to be explained. This study aims to ensure predictive accuracy and predictive certainty through fused SHAP with ensemble models to provide accurate insights that enhance clinical decision making and build trust in AI-assisted diagnostics.

1.1 Contribution

In the present work, we primarily aimed to develop a reliable and explainable system for early lung cancer risk prediction based on ensemble machine learning methods. We included SHAP (SHapley Additive exPlanations) in our framework to make the model prediction interpretable by human readable explanation for more transparent and trustful medical decisions. We further found several significant risk factors of lung cancer by SHAP, and they might be valuable elements for clinical evaluation as well as early stage screening of lung cancer. Our dissemination of the proposal model and equated it to the traditional use of Machine Learning showed the ensemble model to be more robust and accurate and precise metrics almost the same. We tried to expand the data to test the model to ensure that our approach is not solely reliant on one data collection policy. We also tried to balance research and value our study offers to the field. We are suggesting usage of a practical and transparent AI tool which is likely to support practical usage of EHDs during early monitoring of diseases.

CHAPTER 2

2.0 Literature Review

To forecast lung cancer, several studies have used machine learning techniques. Al-Jamimi et al. (2025) built a model using advanced techniques like RFE-SVM and XGBoost. Their model got 100% accuracy on the datasets they tested, which shows it worked very well. But they used small datasets [7]. Sumon et al. (2024) made a stacking machine learning model using metabolomics data to predict small cell lung cancer. Their model reached 85.03% accuracy. It gave good results, but the performance could be improved more [8]. Sinjanka and Kaur (2024) used the Random Forest algorithm to detect lung cancer early. Their model had 97.9% accuracy, which is quite high. But the model's predictions were not explainable [9].

Chen and Wu (2025) used a Deep Q Network (DQN) to predict lung cancer risk in older people. Their model worked well, with AUROC scores between 0.937 and 0.953. However, their research focused only on elderly patients, so the model may not be useful for younger people. These studies show that machine learning, especially ensemble models, can help predict lung cancer. But many of the models were tested on small or specific types of data [10]. Our study is conducted to address these issues and improve the physicians' trust in and memorization of results, we propose an ensemble model running on multiple datasets by utilizing SHAP to interpret predictions. Flyckt (2024) created a model for lung cancer diagnosis with regular blood analysis and smoking habit using dynamic ensemble selection (DES). The model had achieved an ROC-AUC of 0.77 but with high false-positive rate (36.22%), it inappropriately diagnosed many normal individuals as cancerous [11].

Guan and Du (2024) employed XGBoost and metabolic information for early diagnosis of lung cancer, with an accuracy of 75.29%. However, their model only integrated very limited demographic information such as age and sex, which might make it less generalizable [12]. Dritsas and Trigka (2023) presented another study related to RotF of which achieved an accuracy of 97.1% on a public dataset. Regardless of their model's success, it was unexplainable and it also did not indicate to them which features were most predictive. These studies highlight the need for, and the scarcity of, models that are both accurate and interpretable when learning from patient's data from multiple origins. Our contribution in this regard is to apply an ensemble learning model augmented with the SHAP technique, to render the predictions more accurate and explainable to the medical practitioners.

2.1 Research Gap:

One such open question is analysis of risk assessment for catching lung cancer early using machine learning and data mining. Most of the current models only adopt one algorithm, as logistic regression or decision trees, and are unable to effectively embody complicated lung cancer risks. This limits the possible accuracies and reliabilities of predictions. Ensemble learning algorithms that integrate several models may be able to enhance performance, but have not yet been widely applied. Another major issue is the lack of explainability in most machine learning models. Many of them function as black boxes, making it hard for doctors to understand or trust the results. It is this limitation that prevents further integration of these models in the healthcare systems. SHAP is an innovative technique in explaining model outcomes but is still underutilized in lung cancer prediction studies. In addition, many models rely on clinical resources such as CT scans, which are not common in the resource-poor settings. There is a demand for models that rely on easy, unintrusive information such as questionnaires and case history. Fulfilling these gaps through the use of ensemble models augmented with SHAP for explainability would result in lung cancer prediction systems that are more accurate, interpretable, and easier to use which would assist in early prognosis and better healthcare outcomes.

2.2 Research Objectives:

This study seeks to construct the first machine learning model capable of early-stage lung cancer prediction with the highest reliability and precision possible. Initial steps involve identifying the most critical determinants of lung cancer risk such as symptoms, habits, and other medical conditions. To boost the application performance, we develop a novel machine learning model combining multiple probe algorithms associated with predictive modeling. In this instance, we implemented SHAP to improve the model's explainability, which aids in understanding how the model arrives at a conclusion.It is crucial to establish trust in the health care, and this has become more important than ever. Furthermore, the performance of the proposed model is tested and compared with other existing computational models on several benchmark datasets. This guarantees the performance and practicability of the model in practice. For higher accuracy, the model presented is subjected to a strict test standard leading to better results validity. In a crux, the aim of this study is to bridge the gaps in lung cancer risk prediction through model sophistication and aid in providing risk estimates to medical practitioners.

CHAPTER 3

3.0 Methodology

Figure 1 demonstrates all steps in our proposed model. The first step begins by data cleaning and then focuses on deleting duplicates and other complicated aspects such as the conversion of vocabulary into numerical values. In order to ascertain effective results, the SMOTE technique is utilized to maintain an even distribution of cancerous and non-cancerous instances. Upon completing the model, we split the data into separate testing and training sets to ascertain the model's efficacy.

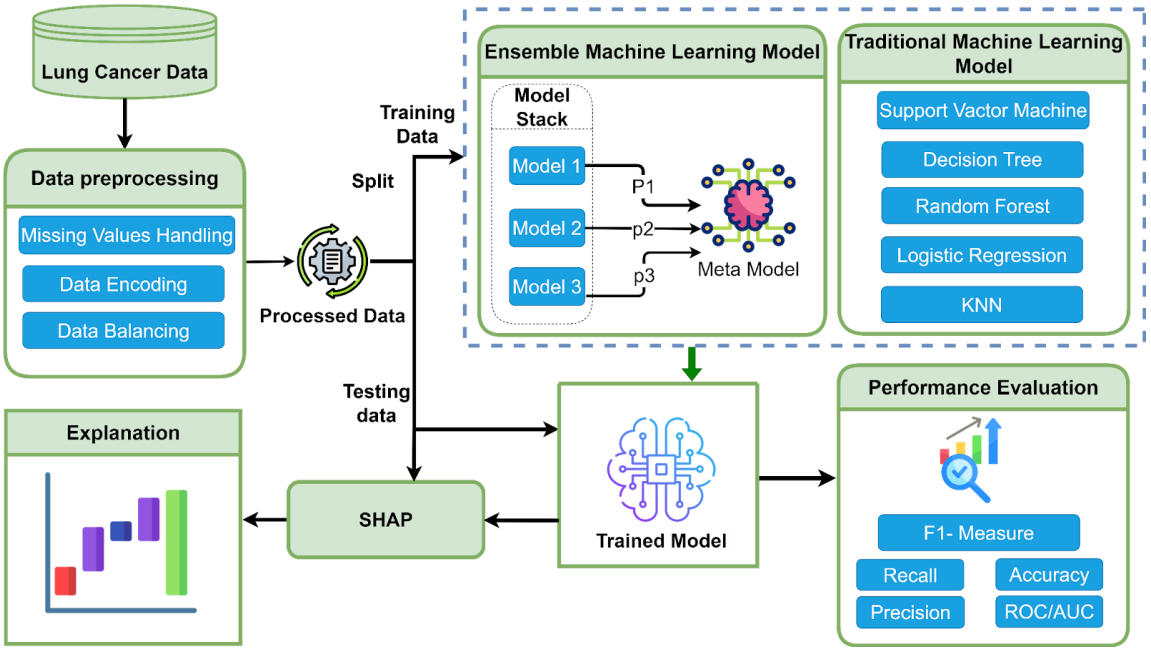


Figure 1: Methodology for Lung Cancer Risk Prediction.

Among the many machine learning methods we employ is Support Vector Machine (SVM), Decision Tree, Random Forest, Logistic Regression, and K-Nearest Neighbors (KNN), to develop distinct forecasting strategies within the framework of model training. We create a stacked ensemble model to supplement the advantages of all models by using a base model in order to improve prediction performance even more. To improve interpretability, SHAP (SHapley Additive Explanations) is utilized, which illustrates the impact of each characteristic on the prediction outcomes. To ensure that all models are reliable and robust on unseen data, we lastly evaluate their performance using classification measures such ROC-AUC, F1-score, recall, accuracy, and precision.

3.1 Data Set Overview:

3.1.0 Data Set-1:

The first dataset was foundational for the prediction of lung cancer risk, and as such, Dataset-1 is a backbone for this analysis and used datasets that are, as of October 2023, publicly available. Dataset-1 contains 1,001 records, spanning 1,000 lung cancer patients, and was collected from 1,000 individuals. There are a total of 26 attributes that encompass a range of patient data, illustrating the demographics like age and sex as well as environmental factors such as pollution and occupation-specific hazards, and lifestyle factors like alcohol and smoking. Furthermore, data contains diverse clinical manifestations such as chest pain, fatigue, breathlessness, and hemoptysis. It includes data related to genetic susceptibility, chronic lung diseases, asthma, and even obesity. All this data makes a comprehensive analysis of the possible factors associated with lung cancer feasible. The last attribute, “Level”, is the target variable. Table 3.1.0 column And features, describes the entire dataset and their features.

3.1.1 Data Set-2

Dataset-2 serves as an extra resource for estimating the risk of lung cancer in this investigation. The dataset, which includes data from 309 people, is openly accessible. There are sixteen features in all in the dataset. One of them is the goal feature that determines if a person has lung cancer, whereas the other fifteen are input features. Basic personal and health-related data are among the entry features. Age, gender, smoking status, yellow fingers (a possible smoking sign), anxiety, peer pressure, chronic illness, fatigue, allergies, wheezing, alcohol use, coughing, shortness of breath, difficulty swallowing, and chest discomfort are a few of these. While the other features are nominal, with values classified as "Yes" or "No" (or equivalent labels), age is a numerical feature. Each feature provides useful clues about potential risk factors for lung cancer. For example, smoking and yellow fingers can reflect tobacco use, while symptoms like coughing or chest pain may signal respiratory issues. The final feature, Lung Cancer, is the target variable and indicates whether the person has been diagnosed with the disease. This dataset helps in understanding the link between lifestyle, symptoms, and lung cancer, and it is used for model training, balancing, feature analysis, and performance evaluation in the proposed methodology.

Table 3.1.0: Description of Features and Data Types in Dataset 1 for Lung Cancer Risk Prediction

Feature Name	Description	Data Type
Age	Age of the individual (in years)	Numerical
Gender	Gender of the patient (1 = Male, 2 = Female)	Nominal
Air Pollution	Exposure level to air pollution (1–10 scale)	Ordinal
Alcohol Use	Frequency of alcohol consumption (1–10 scale)	Ordinal
Dust Allergy	Severity of dust allergy (1–10 scale)	Ordinal
Occupational Hazards	Exposure to harmful work environment (1–10 scale)	Ordinal
Genetic Risk	Genetic/family history risk (1–10 scale)	Ordinal
Chronic Lung Disease	Presence of chronic lung disease (1–10 scale)	Ordinal
Balanced Diet	Frequency of balanced diet (1–10 scale)	Ordinal
Obesity	Obesity level (1–10 scale)	Ordinal
Smoking	Smoking frequency/severity (1–10 scale)	Ordinal
Passive Smoker	Exposure to second-hand smoke (1–10 scale)	Ordinal
Chest Pain	Severity of chest pain (1–10 scale)	Ordinal
Coughing of Blood	Frequency of coughing blood (1–10 scale)	Ordinal
Fatigue	Tiredness or fatigue level (1–10 scale)	Ordinal
Weight Loss	Level of weight loss (1–10 scale)	Ordinal
Shortness of Breath	Difficulty in breathing (1–10 scale)	Ordinal
Wheezing	Presence of wheezing (1–10 scale)	Ordinal
Swallowing Difficulty	Trouble in swallowing (1–10 scale)	Ordinal
Clubbing of Finger Nails	Swelling/clubbing of nails (1-9 scale)	Ordinal
Frequent Cold	Frequency of catching cold	Ordinal
Dry Cough	Frequency of dry cough	Ordinal
Snoring	Severity of snoring (1-7 scale)	Ordinal
Level (Target Class)	Lung cancer risk (Low, Medium, High)	Categorical

Table 3.1.1: Description of Features and Data Types in Dataset 2 for Lung Cancer Risk Prediction

Feature Name	Description	Data Type
Age	Age of the patient (in years)	Numerical
Gender	Gender of the patient (M = Male, F = Female)	Nominal
Smoking	Whether the patient smokes (2 = Yes, 1 = No)	Ordinal
Yellow Fingers	Presence of yellowing fingers due to smoking (2 = Yes, 1 = No)	Ordinal
Anxiety	Presence of anxiety symptoms (2 = Yes, 1 = No)	Ordinal
Peer Pressure	Influence of peer pressure on lifestyle (2 = Yes, 1 = No)	Ordinal
Chronic Disease	Presence of chronic illness (2 = Yes, 1 = No)	Ordinal
Fatigue	Experience of fatigue (2 = Yes, 1 = No)	Ordinal
Allergy	Presence of allergic conditions (2 = Yes, 1 = No)	Ordinal
Wheezing	Experience of wheezing (2 = Yes, 1 = No)	Ordinal
Alcohol	Consumption of alcohol (2 = Yes, 1 = No)	Ordinal
Coughing	Experience of coughing (2 = Yes, 1 = No)	Ordinal
Shortness of Breath	Difficulty in breathing (2 = Yes, 1 = No)	Ordinal
Swallowing Difficulty	Trouble in swallowing (2 = Yes, 1 = No)	Ordinal
Chest Pain	Experience of chest pain (2 = Yes, 1 = No)	Ordinal
Lung Cancer	Final diagnosis (Yes = Lung Cancer, No = No Lung Cancer)	Binary

3.2 Data Preprocessing:

3.2..1 Dataset-1 Preprocessing

3.2.1.1 Removal of Irrelevant Columns

Removing irrelevant or non-informative columns helps to reduce noise and computational complexity, enabling the model to focus on meaningful features. In Dataset-1, columns such as Index and Patient Id were dropped because they do not contribute to lung cancer prediction. This streamlined the dataset for better model performance.

3.2.1.2 Target Variable Encoding

Machine learning models require numerical inputs, so categorical target variables must be converted into numeric codes through label encoding. "Level," the goal variable indicating lung cancer risk as High, Medium, or Low, was encoded as 2, 1, and 0 respectively. This numeric transformation allowed the model to interpret the risk categories effectively.

3.2.1.3 Categorical Feature Encoding

Categorical input features must be converted into numeric form to be used by most machine learning algorithms, commonly through label encoding. Nominal features such as gender were label encoded; for instance, Male was encoded as 0 and Female as 1. This encoding enabled the inclusion of gender and other categorical variables as meaningful inputs.

3.2.1.4 Class Balancing

Models may be biased toward majority classes as a result of imbalanced datasets. To balance the dataset, methods such as the fake Minority Oversampling Technique (SMOTE) create fake data for minority classes. Lung cancer risk groups were unbalanced in Dataset 1. By using SMOTE to oversample minority classes, a balanced dataset was produced, enhancing the accuracy and fairness of the model.

3.2.2 Dataset-2 Preprocessing

3.2.2.1 Nominal Feature Encoding

Nominal or binary categorical features need to be converted into numeric values for model processing, commonly mapping categories like Yes/No to numbers. In Dataset-2, nominal features such as smoking, anxiety, and fatigue were encoded with Yes as 2 and No as 1. This ensured compatibility with machine learning models.

3.2.2.2 Numerical Feature Normalization

Numerical features with different scales should be normalized to a common scale, often between 0 and 1, to prevent features with large values from dominating the learning process. The age feature in Dataset-2 was normalized, standardizing its scale to improve model convergence and performance.

3.2.2.3 Target Variable Encoding

Binary classification targets need to be encoded numerically (Yes = 1, No = 0) for classification algorithms. The lung cancer diagnosis target was encoded as 1 for presence and 0 for absence, facilitating binary classification.

3.2.2.4 Class Balancing

Model generalization is enhanced and bias towards the majority class is avoided by distributing the amount of samples evenly among the classes. The proportion of positive and negative lung cancer cases was unbalanced in Dataset 2. In order to balance the dataset and improve prediction, SMOTE was used to oversample the minority class.

3.3 Exploratory Data Analysis

Figure 3.3.1: heatmap shows the relationship between different features in Dataset-1 and their connection to lung cancer risk. The colors indicate how strongly two features are related: dark blue means a strong positive relationship, light yellow means a weak or negative relationship. For example, features like air pollution, alcohol use, occupational hazards, and smoking have strong positive links with lung cancer risk (Level). This helps us understand which factors are important for predicting lung cancer.

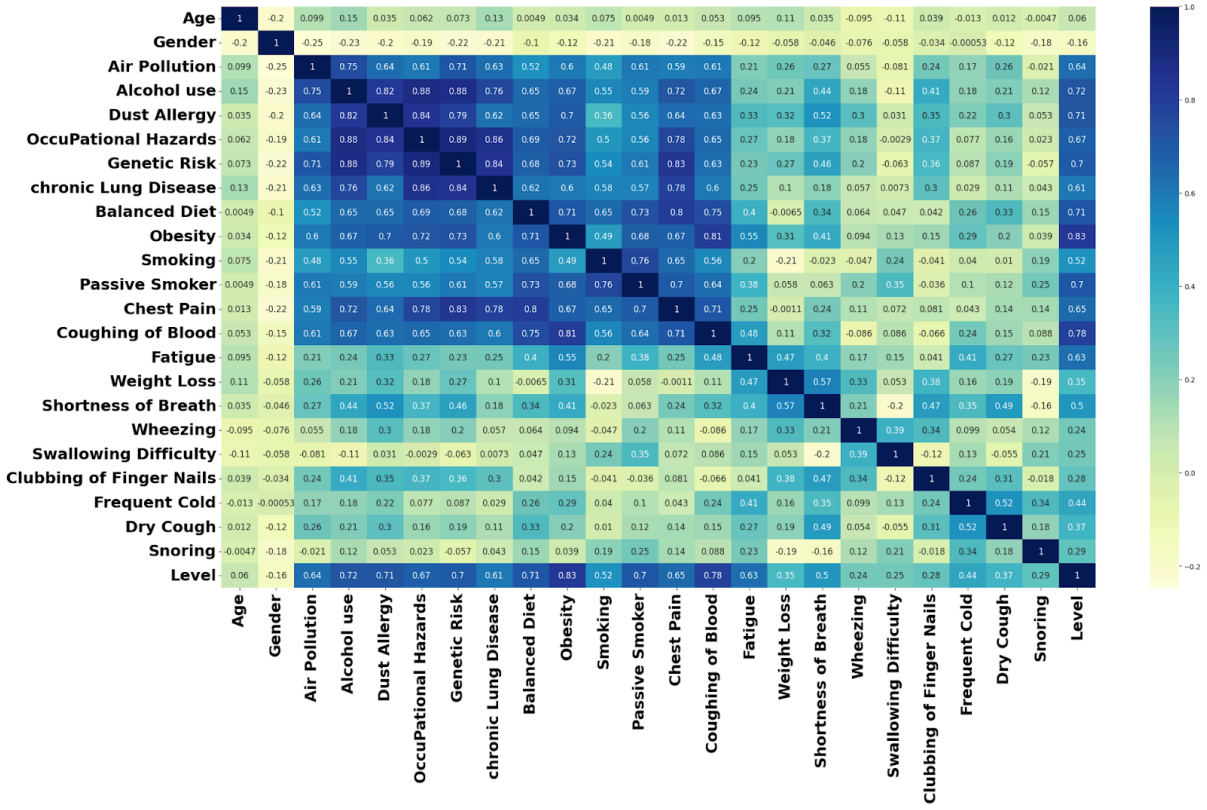


Figure 3.3.1: Correlation Between Health Features and Lung Cancer Risk (Dataset-1)

Figure 3.3.2: The bar chart shows the distribution of lung cancer risk levels among patients based on gender. Among males, there is a noticeably higher number of patients in the medium and high-risk categories, especially at the highest risk level. In contrast, female patients are more concentrated in the lower risk levels, with the number of patients decreasing as the risk level increases. This visualization highlights a clear pattern where male patients are more frequently represented in the higher lung cancer risk groups, whereas females tend to be associated with lower risk categories.

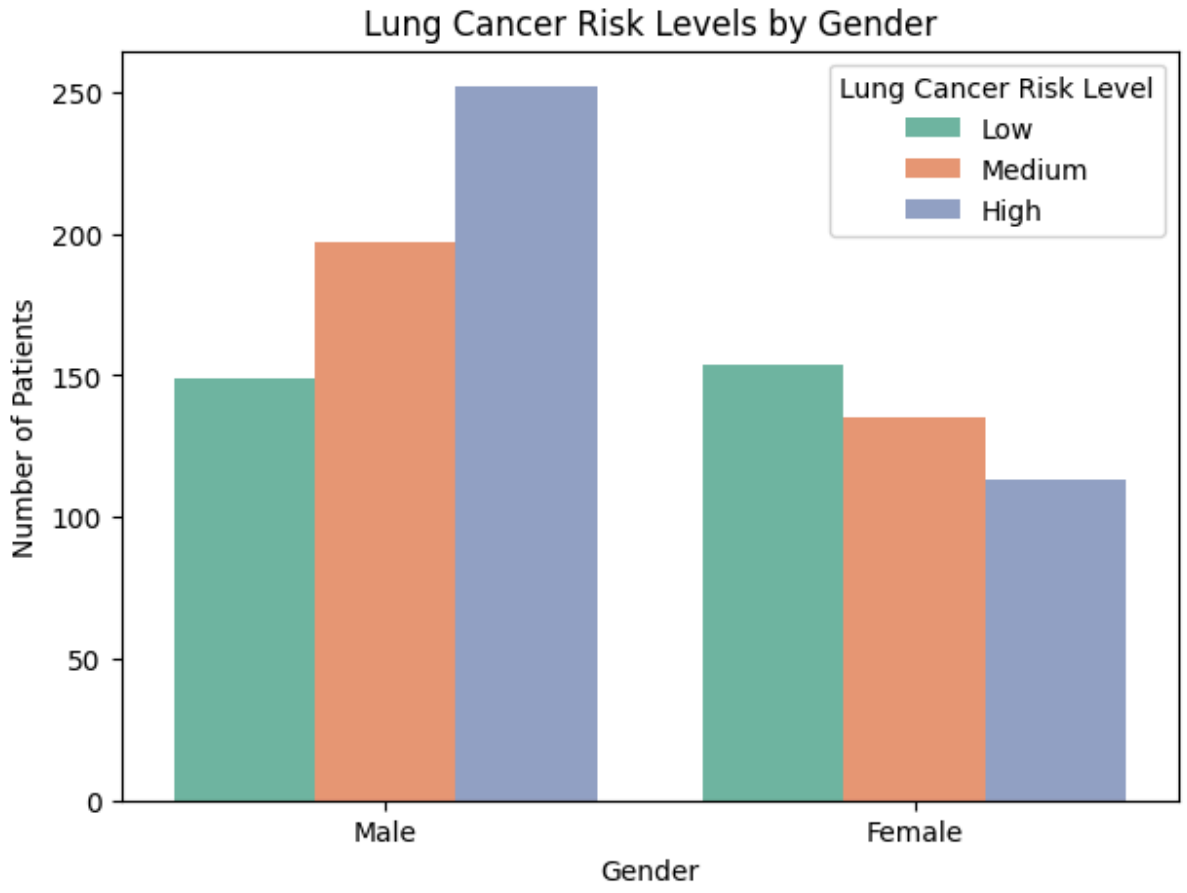


Figure 3.3.2: Lung Cancer Risk Levels by Gender (Dataset-1)

Figure 3.3.3 shows the number of male and female patients who were found to be at risk of lung cancer based on Dataset-2. The chart shows that both men and women have cases marked as "YES" (lung cancer risk) and "NO" (no risk), but the number of high-risk cases is slightly higher in males compared to females.

This suggests that gender may have some influence on lung cancer risk, but it is not the only factor. Other health features must also be considered for accurate prediction.

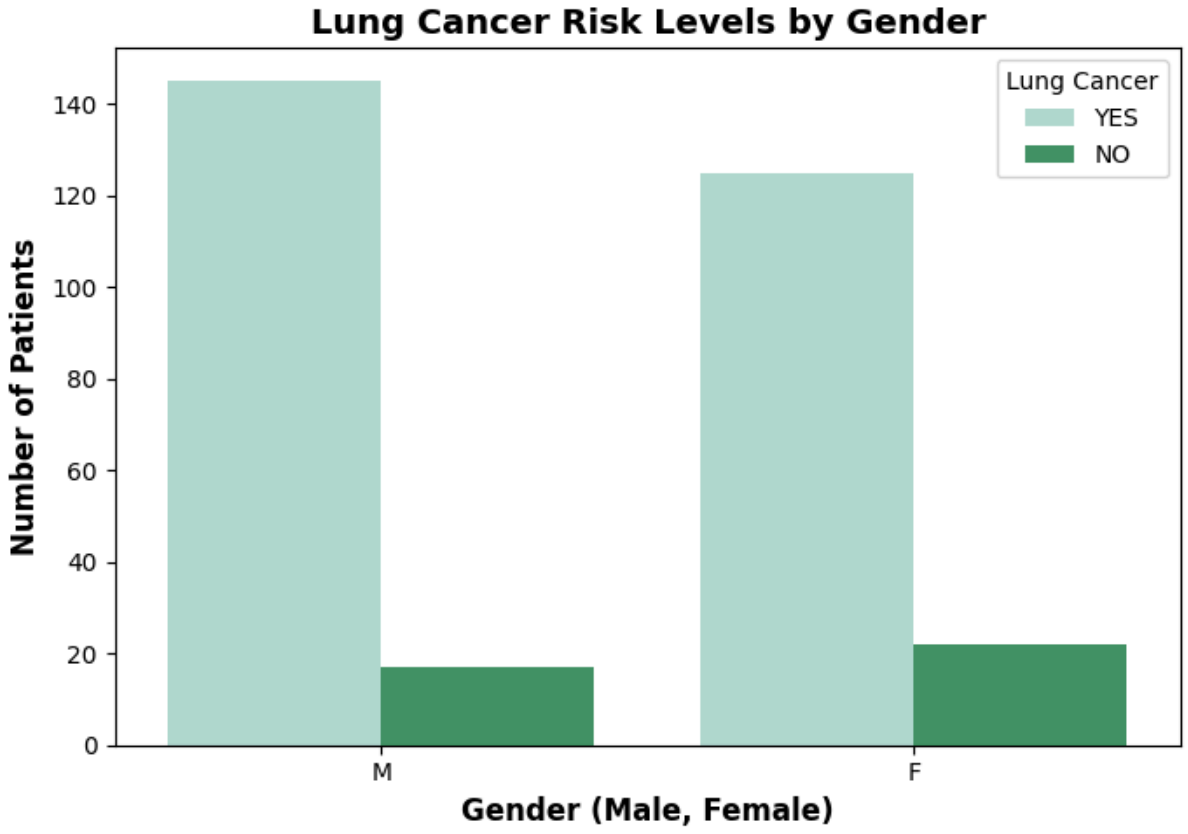


Figure 3.3.3: Lung Cancer Risk Levels by Gender (Dataset-2)

Figure 3.3.3 shows how different health features are related to lung cancer risk. The values in the heatmap are correlation scores, which range from -1 to 1. A value closer to 1 means a strong positive relationship, while a value near -1 means a strong negative one. A value near 0 means little or no relationship.From the figure, we see that features like anxiety (0.14), yellow fingers (0.18), peer pressure (0.11), allergy (0.15), wheezing (0.25), and chest pain (0.19) have a noticeable positive correlation with lung cancer. These features are more common in people who are at higher risk. However, none of the features show a very strong correlation individually, which means that lung cancer risk depends on a mix of factors rather than one single cause.This analysis helps us understand which features might be more useful when predicting lung cancer using machine learning models.

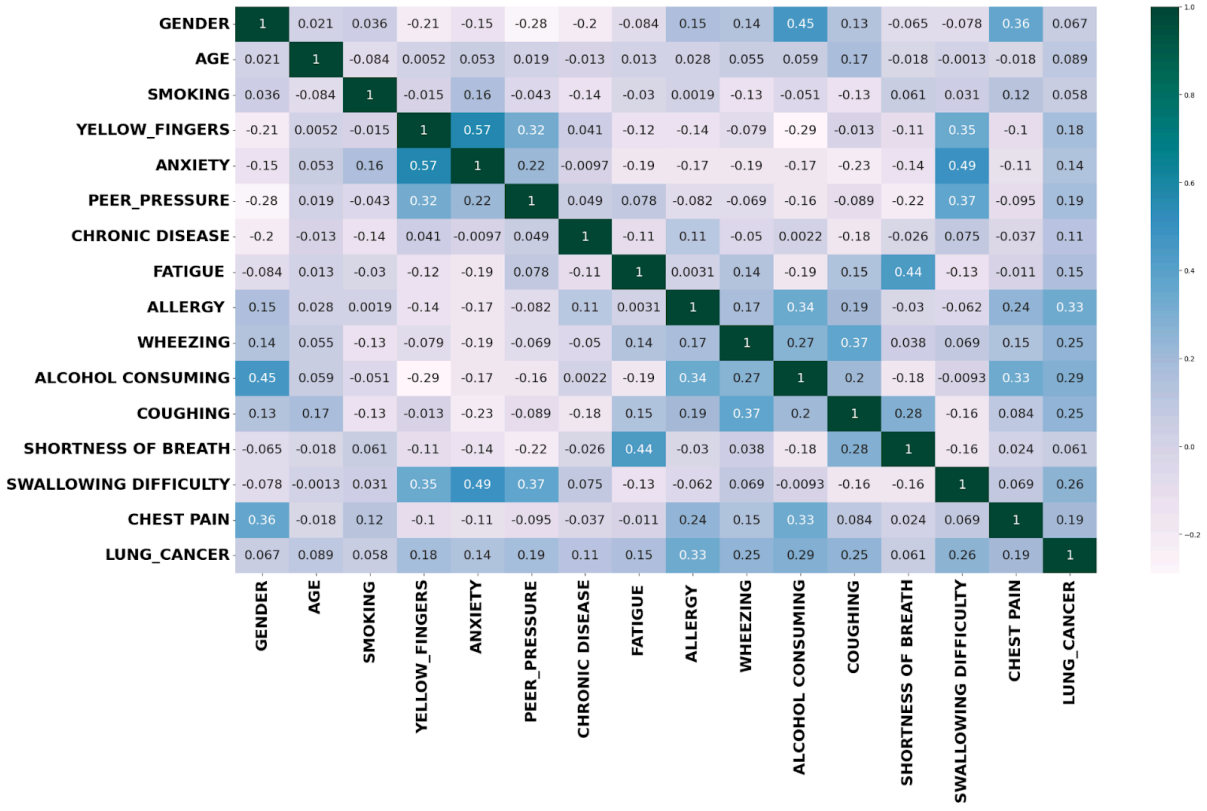


Figure 3.3.3: Correlation Between Health Features and Lung Cancer Risk (Dataset-2)

3.4 Machine Learning Models

This study used many supervised machine learning approaches to evaluate the risk of early lung cancer. SVM, Random Forest, Decision Trees, K-Nearest Neighbors (KNN), Logistic Regression, and an Ensemble Learning model are some of the models that are employed. These algorithms were selected in order to evaluate the performance of individual models to a combined strategy and to investigate both linear and non-linear decision limits. To provide a thorough assessment of prediction accuracy and dependability, each model offers a unique approach to learning from data.

3.4.1 Support Vector Classifier:

Discouragement Classification in "data science" is cutting edge technology. It uses a vector classifier to split datasets into two categories. It calculates the so-called optimal hyperplane that minimizes the distance to the two groups. It uses some clever trick to accomplish the separation of enormously complicated intertwining data. Distance in Dataset-1 was characterized by age, sex, smoking, alcohol, pollution, and symptoms like fatigue, dyspnea, and angina. These attributes trained the model to estimate the probability of having lung cancer. Dataset-2 consisted of age, gender, smoking, anxiety, and a sick coughing fit, chest pain, and a number of 'co-morbid' conditions. The model was able to suss the data and come out with a conclusion – does the patient have lung cancer? It also bears mentioning that the Support Vector Classifier has a unique and rigorous method of filtering out data patterns and making predictions using a decision function. This function can be written as:

$$f(x) = sign(w \cdot x + b)$$

Here:

- w (weight vector)
- x (input feature vector)
- b (bias term)

The function determines which side of the decision boundary a data point lies on is false. If $f(x)$ is positive, the data is assigned to one class. If it is negative, it is assigned to the other class. The model attempts to find the best values for w and b such that the classes are separated with the largest possible gap or margin between them. In doing this, SVC solves an optimization problem. For SVC, more complex or nonlinear data requires the use of the

kernel trick. This permits the model to classify data more easily in a higher-dimensional space. This mathematical technique permits the model to improve its predictions by learning the best boundary between classes.

3.4.2 Decision Tree Classifier

A Decision Tree is an example of machine learning that classifies and predicts outcomes of a problem. It resembles a flowchart in which data sets are divided into branches as answers are derived from questions about the features. In the example shown, every internal node represents a test on a feature, every branch depicts test outcomes, and every leaf provides the ultimate class label. The tree begins from the root and continues to subdivide the data into smaller portions until it comes to a decision. The objective is to divide the data in a manner so that the groups are as pure as possible which implies that every group contains mostly one class. In Dataset-1, the features apportioned to the Decision Tree were age, gender, smoking, air pollution, and the symptoms of coughing with blood as well as shortness of breath to decide on lung cancer risk level. In Dataset-2, the tree interacted with features like anxiety, peer pressure, fatigue, and chest pain in order to predict lung cancer. A Decision Tree decides where to split the data using measures like Information Gain or Gini Index. The most common one is Information Gain. It is built on a concept called Entropy.

Entropy as:

$$Entropy(S) = -\sum p_i \, log2 \, (p_i)$$

where p_i is each class's percentage in data set S. Following the split, the tree selects the feature that exhibits the largest decrease in uncertainty and provides the maximum Information Gain. Information Gain is computed using:

$$Gain(S, A) = Entropy(S) - \sum \frac{|Sv|}{|S|} \times Entropy(Sv)$$

where A is the feature used for splitting, and Sv is the subset of data after splitting by feature A. This splitting continues until the tree reaches a stopping condition, such as maximum depth or minimum samples, and then makes the final classification.

3.4.3 Random Forest Classifier

A well-liked approach for regression and classification is Random Forest. Building numerous decision trees and integrating their output is how it operates. Every tree employs random features to create splits after being trained on a randomly selected portion of the data. The model is more accurate and less prone to overfit thanks to this randomness. To estimate the risk of lung cancer in **Dataset-1**, Random Forest used characteristics including age, gender, smoking, air pollution, and symptoms like exhaustion and blood in the cough. Features like anxiety, peer pressure, chest pain, and exhaustion were employed in Dataset-2 for the same objective. Every tree is voted on to determine the final prediction, and the class with the most votes wins. This technique lowers errors from individual trees and increases accuracy.

In terms of mathematics, the ultimate forecast is:

$$y\hat{} = majority\_vote(h1(x), h2(x),..., hn(x))$$

Where each $h_i(x)$ is a tree's prediction. The class that appears most often is chosen as the output.

3.4.4 Logistic Regression

For classification tasks, the straightforward but effective algorithm known as logistic regression is employed. It operates by determining a correlation between the probability of a particular class and the input features. Rather than providing precise values, it classifies based on a probability between 0 and 1. To predict the risk of lung cancer in Dataset-1, Logistic Regression employed characteristics like age, gender, smoking, alcohol consumption, and symptoms like coughing and chest pain. It worked with characteristics like anxiety, exhaustion, dyspnea, and peer pressure in Dataset-2. It uses the sigmoid function to turn the result into a probability:

$$P(y = 1|x) = \frac{1}{1+e^{-(w \cdot x+b)}}$$

Here, w is the weight vector, x is the input data, and b is the bias. If the probability is greater than 0.5, the input is classified as having a high risk; otherwise, it is considered low risk. This model is simple, fast, and works well when the data is linearly separable.

3.4.5 K-Nearest Neighbors (KNN)

A straightforward machine learning approach for classification is called K-Nearest Neighbors. A new input's "k" nearest data points, or neighbors, are examined, and the most common class among them is assigned. It does not build a model during training instead, it stores the data and makes decisions at prediction time. In **Dataset-1**, KNN checked nearby data points using features like age, smoking, shortness of breath, and coughing of blood to predict the risk level for lung cancer. Within Dataset-2, it used features such as chest pain, anxiety, fatigue, and peer pressure to classify whether the person is at risk.

The algorithm calculates the **distance** (often using Euclidean distance) between the new data and existing points:

$$d = \sqrt{(x1 - x2)^2 + (y1 - y2)^2 + \dots}$$

Then it picks the 'k' nearest neighbors and chooses the class that appears the most among them. KNN is easy to use and often gives good results, especially when the data is not too large and features are properly scaled.

3.4.6 Ensemble Machine Learning Model

In this study, a Stacking Ensemble Machine Learning model was employed to increase the overall prediction accuracy of lung cancer risk. The stacking technique generates a more reliable final prediction by integrating the strengths of multiple independent models. In particular, three different classifiers were used as foundation learners: Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN). Each of these models was first trained separately using the training data in order to identify unique patterns and interactions. Each of their individual predictions were sent to the meta-classifier, a second-level model that in this case was also a KNN . After training, the stacked model was evaluated using the test data. Common metrics such as F1-score, recall, accuracy, and precision were used to assess the model's performance.

3.5 Performance Evaluation

For measuring how well the machine learning models perform, some common classification metrics were employed. These are accuracy, precision, recall and F1-score. Accuracy is the aggregate correctness of the model whereas how many of the cases that were measured as positive are actually correct is denoted by precision. Recall is a measure of the recall with which model could contain actual positive cases and F1 - score gives you the balance between precision and recall.

3.5.1 Accuracy

Accuracy assesses the performance of a machine learning model. It shows the proportion of accurate predictions the model made.

Formula:

$$Accuracy = \frac{Total\ Number\ of\ Predictions}{Number\ of\ Correct\ Predictions}$$

If the accuracy is high, it means the model is correctly predicting most cases. However, in datasets with imbalanced classes, accuracy alone may not be a reliable metric.

3.5.2 Confusion Matrix

The Confusion Matrix offers a thorough summary of the model's prediction results by showing the proportion of cases that were correctly or incorrectly predicted. This table has four values:

- True positive: The patient's lung cancer was accurately predicted by the program.
- False Positive (FP): The patient's does not have lung cancer, despite the model's prediction.
- True Negative (TN): The patient's lack of lung cancer was accurately predicted by the model.
- False Negative (FN): The patients has cancer even though the model said they wouldn't.

This matrix is essential since it allows us to calculate other performance measures including accuracy, recall, and F1-score.

3.5.3 Precision

Precision indicates the proportion of instances that are truly positive, such as lung cancer, out of those that are anticipated to be positive. Stated differently, it assesses the model's accuracy in forecasting a favorable outcome.

Formula:

$$Precision = TP + FPTP$$

The model produces fewer incorrect predictions when its accuracy is high.

3.5.4 Recall

The number of real positive instances (lung cancer patients) that the model was able to identify is called recall, often referred to as sensitivity or true positive rate.

Formula:

$$Recall = TP + FNTP$$

When a model has a high recall, it is able to identify the majority of real positive cases and miss less.

3.5.5 F1-Score

The F1-score is the harmonic mean of recall and accuracy. It is helpful when you want to strike a compromise between recall and precision, particularly when the data is unbalanced (for instance, there are more healthy patients than sick ones).

Formula:

$$F1 = 2 \times \frac{1}{\frac{1}{Precision} + \frac{1}{Recall}}$$

An F1-score closer to 1 indicates that the model performs well in both detecting positives and avoiding false alarms.

CHAPTER 4

4.0 Results:

The performance of many machine learning models for early lung cancer risk prediction on two datasets is shown in this section. K-Nearest Neighbors, SVM, Decision Trees, Logistic Regression, and our suggested Stacking Ensemble Machine Learning Model are among the models that were put to the test. By using each base model's unique characteristics, the stacking ensemble enhances overall prediction accuracy. This method helps achieve better results than any single model alone, which is important for accurate early diagnosis.

Table 4.1: Model Performance Summary on Dataset-1 and Dataset-2

Model	Dataset	Accuracy	Precision	Recall	F1-score
SVM	D1	98.5%	98.6%	98.2%	98.4%
	D2	95.4%	95.6%	95.2%	95.3%
Decision Tree	D1	99.8%	99.7%	99.8%	99.9%
	D2	95%	96%	95%	95%
Random Forest	D1	99.9%	99.9%	99.8%	99.8%
	D2	96%	96%	95%	95%
Logistic Regression	D1	99%	99%	98%	98%
	D2	96%	95%	95.3%	95%
KNN	D1	99.5%	99.5%	99.4%	9.4%
	D2	95.4%	96%	95.4%	95.4%
Ensemble Model (Proposed Model)	D1	99.9%	99.9%	99.9%	99.8%
	D2	98%	98%	98%	98%

The following graphics are provided to show how well the suggested stacking ensemble model performs:

The confusion matrix for the Stacking Ensemble model on Dataset-1 is displayed in **Figure 4.1**. The matrix shows that all three classes were accurately and flawlessly categorized by the model: 55 class 0 samples, 63 class 1 samples, and 82 class 2 samples were all accurately predicted. This demonstrates the model's high accuracy and capacity to differentiate between various risk categories for lung cancer in this dataset.

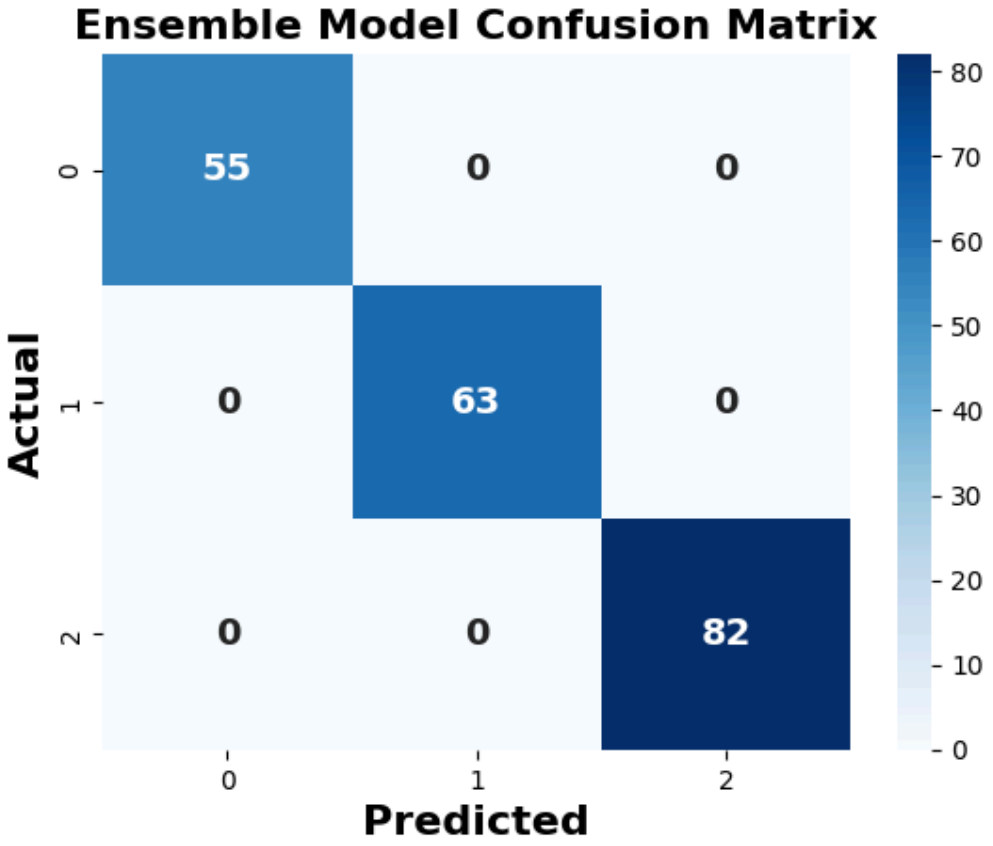


Figure 4.1: Dataset-1's Stacking Ensemble model's confusion matrix.

Figure 4.2 displays the confusion matrix for the Stacking Ensemble model using Dataset-2. This matrix shows how well the model predicted lung cancer cases compared to the actual results. The model correctly recognized 49 healthy people and 57 people with lung cancer. It made only 2 mistakes by labeling healthy people as having cancer, and it did not wrongly classify any cancer cases as healthy. This means the model is very accurate and reliable in identifying early lung cancer. Such performance helps in detecting lung cancer early, which is important for better treatment outcomes.

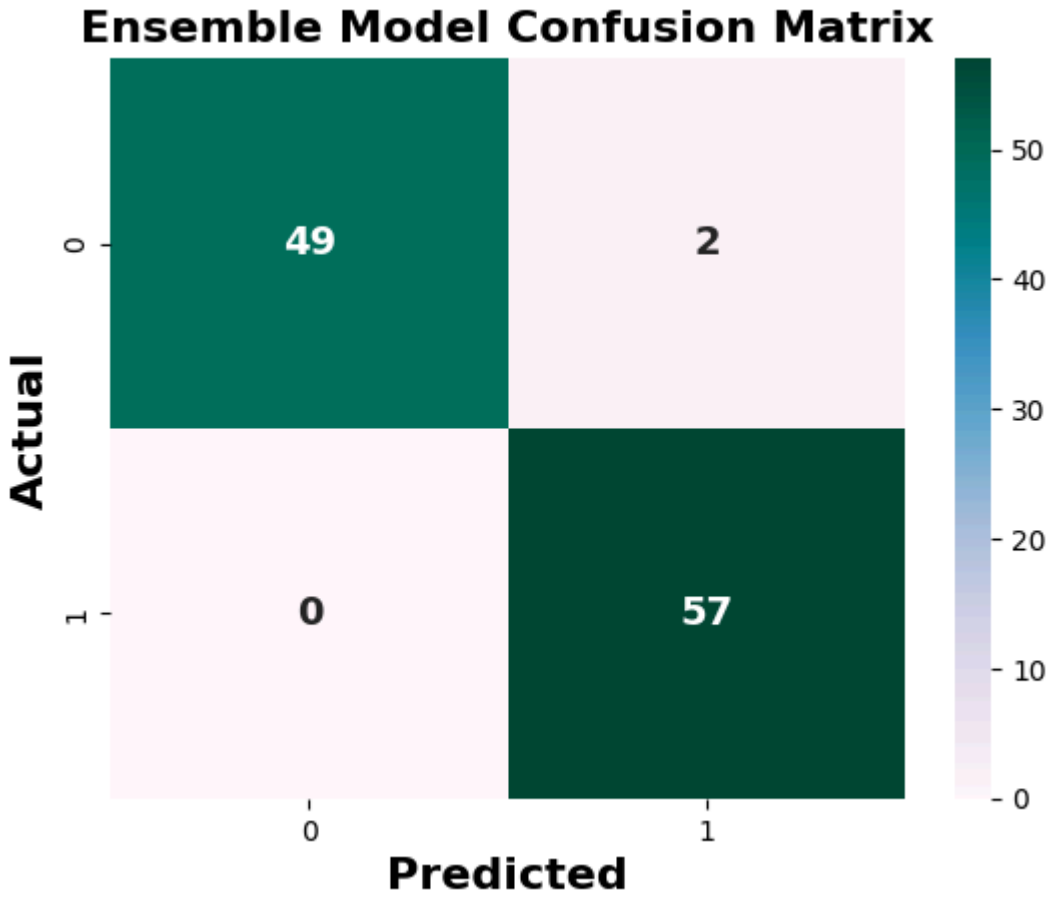


Figure 4.2: Dataset-2's Stacking Ensemble model's confusion matrix

Figure 4.3 shows the ROC (Receiver Operating Characteristic) curves for different machine learning models, including the stacking ensemble and five individual models (SVM, Decision Tree, Random Forest, Logistic Regression, and KNN). The ROC curve helps us understand how well each model separates the lung cancer risk classes. All models in this figure achieved an AUC (Area Under the Curve) score of **1.00**, which means they performed extremely well in this test. A higher AUC means better model performance. The stacking model also performed equally well, showing that combining models can be just as effective or even more stable for predicting lung cancer risk.

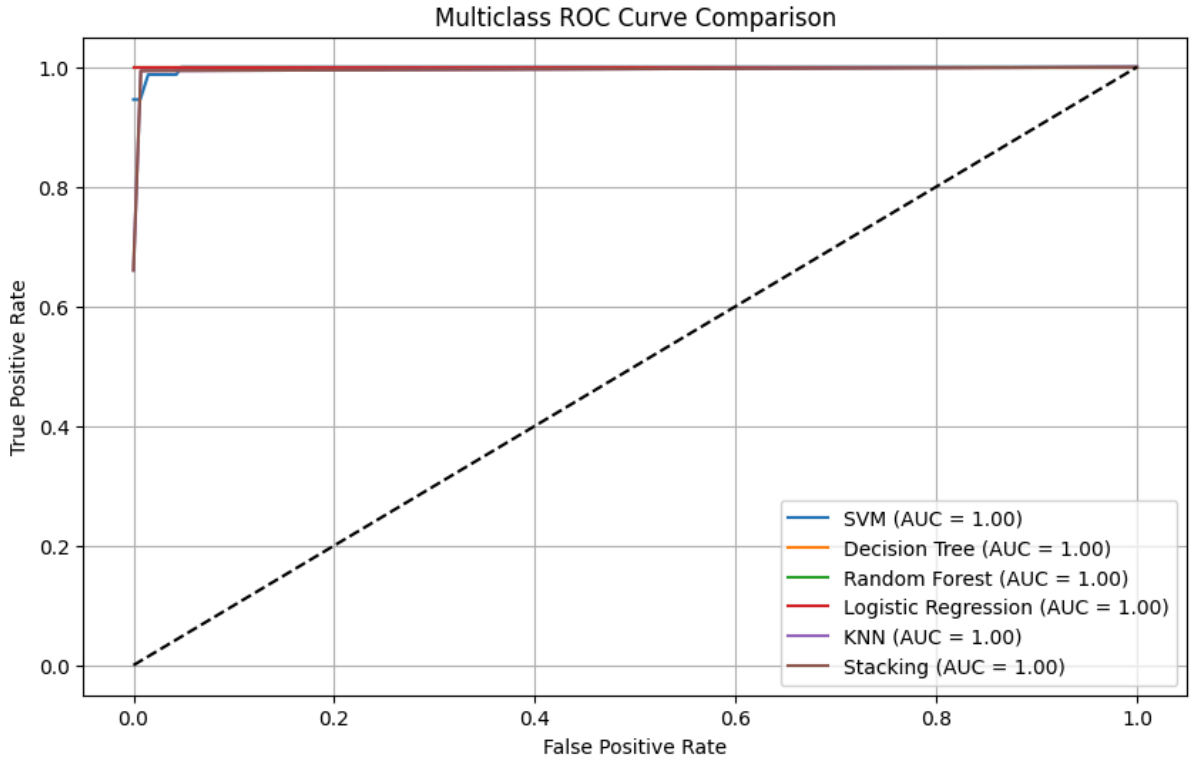


Figure 4.3: ROC curves of the Stacking Ensemble and selected individual models on Dataset-1.

The accuracy of several machine learning models evaluated on two datasets is displayed in **Figure 4.4.** The accuracy on Dataset-1 is shown by the blue bars, while the accuracy on Dataset-2 is shown by the light blue bars. With 99.9% on Dataset-1 and 98% on Dataset-2, the suggested Ensemble Model performed the best and most reliably. Among the individual models, Random Forest and Decision Tree demonstrated excellent results, achieving 99.9% and 99.8% accuracy on Dataset-1 and maintaining strong performance on Dataset-2. Logistic Regression and SVM also produced reliable results, while K-Nearest Neighbors (KNN) achieved 99.5% accuracy on Dataset-1 and continued to perform well on Dataset-2.

Overall, this figure highlights that all models deliver strong performance on both datasets, with the Ensemble Model standing out as the most effective and stable across the two datasets.

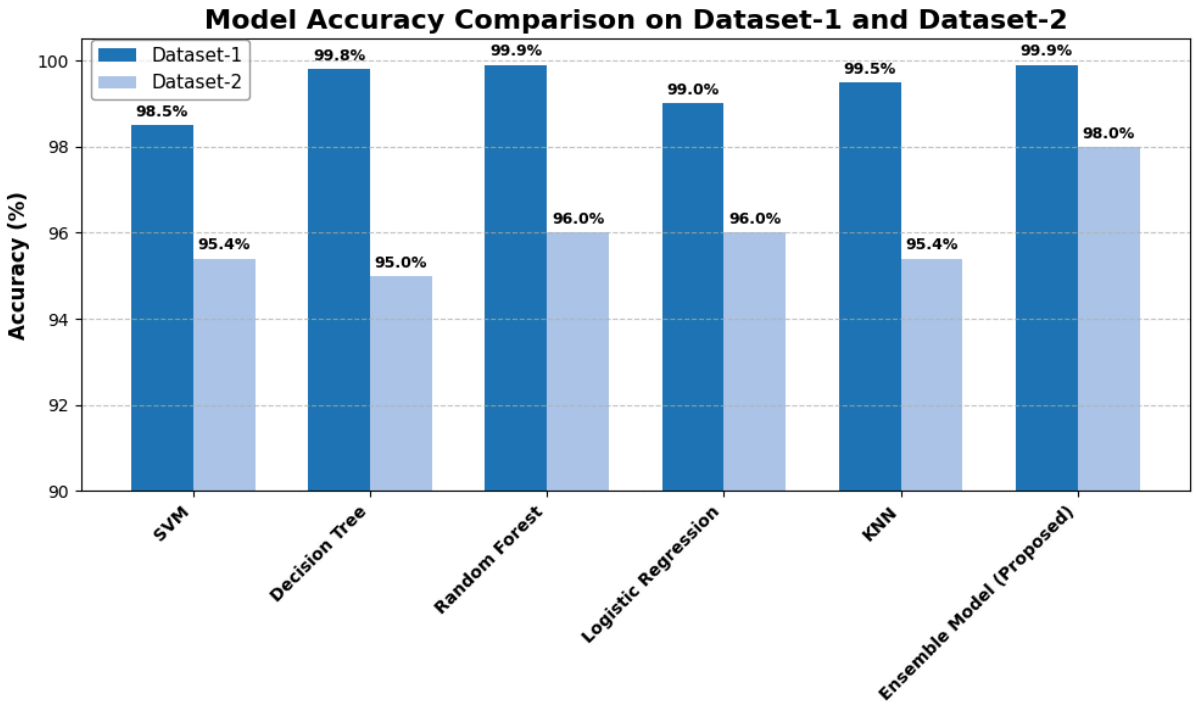


Figure 4.4: Model Accuracy Comparison on Dataset-1 and Dataset-2

Table 4.2 presents a comparison of the proposed ensemble model with previously published studies. The results indicate that the ensemble model achieved the highest performance, recording almost perfect outcomes on Dataset-1 with 99.9% accuracy, precision, and recall, along with a 99.8% F1-score. On Dataset-2, the model also maintained strong performance with consistent values of 98% across all evaluation metrics. In contrast, Dritsas and Trigka (2022), who applied a Rotation Forest on the Kaggle lung cancer dataset, reported an accuracy of 97.1%, while Sinjanka et al. (2024) obtained 97.9% accuracy using a Random Forest model on the same dataset, although precision, recall, and F1-score were not provided. These findings highlight that the proposed ensemble approach demonstrates superior predictive ability compared to earlier methods, particularly on Dataset-1 where the results are near perfect.

Table 4.2: Comparison of Model Performance with Previous Studies

Study and Model	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Our Ensemble Model	Dataset-1	99.9	99.9	99.9	99.8
Our Ensemble Model	Dataset-2	98.0	98.0	98.0	98.0
Dritsas & Trigka (2022) – Rotation Forest	Kaggle LC Dataset	97.1	97.1	97.1	97.1
Sinjanka et al. (2024) – Random Forest	Kaggle LC Dataset	97.9	NA	NA	NA

Figure 4.5 The SHAP (SHapley Additive exPlanations) summary plot for **Dataset-1** illustrates the global feature importance and interpretability of the ensemble model. Features like Age, Coughing of Blood, and Passive Smoker show the highest impact across prediction classes. By quantifying each feature's contribution to the model output, SHAP enhances both feature understanding and explainability of lung cancer risk predictions.

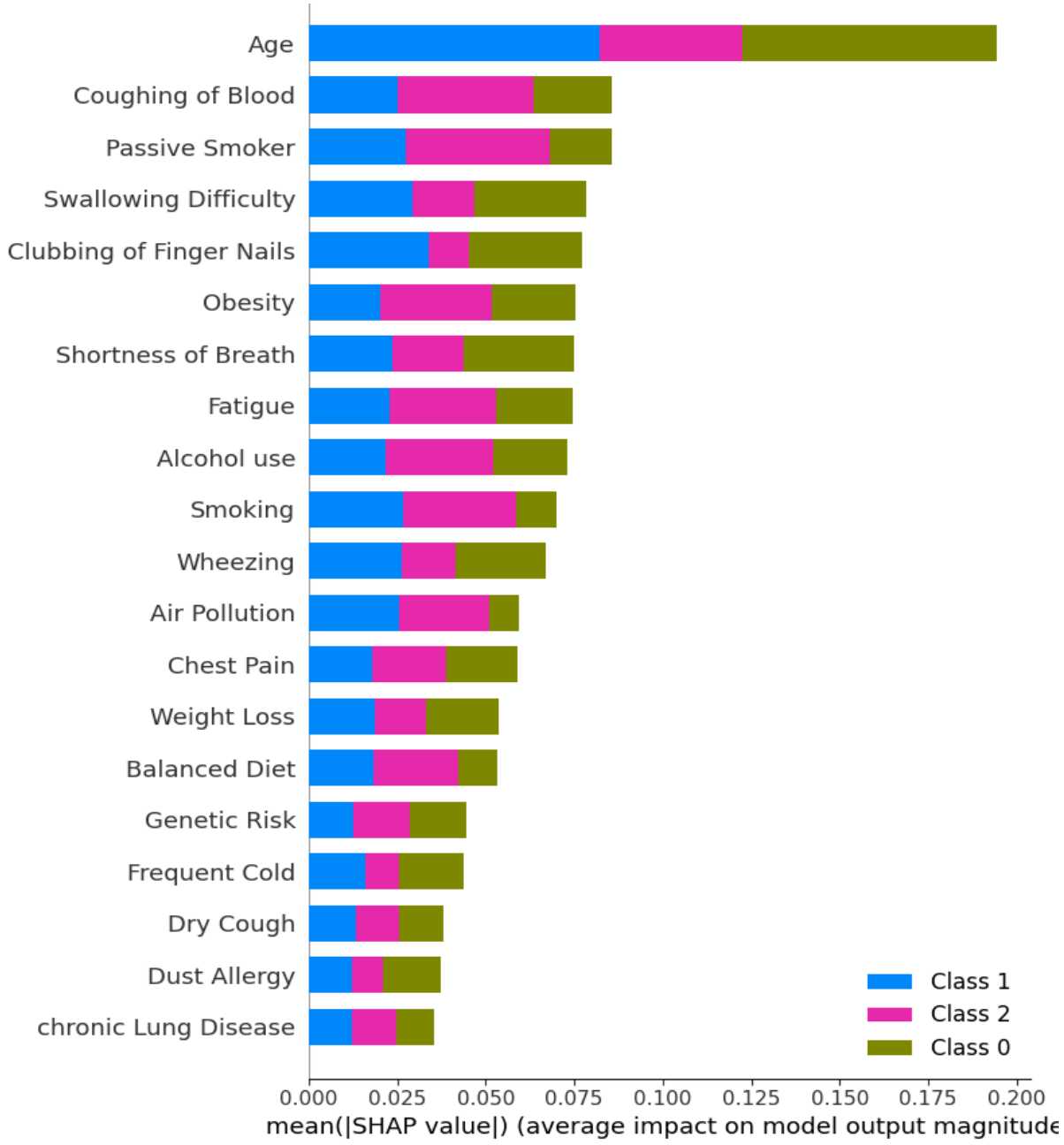


Figure 4.5: SHAP Summary Plot Demonstrating Explainability of the Ensemble Machine Learning Model for Dataset-1

Figure 3.3.4 shows a SHAP summary plot that explains how the ensemble machine learning model predicts early lung cancer risk. Each dot represents a patient, and its position on the plot shows how much a feature increases or decreases the prediction. The color of the dots goes from blue (low values) to red (high values). Features like smoking, yellow fingers, and chronic disease have higher SHAP values, meaning they have a stronger effect on increasing lung cancer risk. This plot makes the model easier to understand by showing which factors influence the predictions the most.

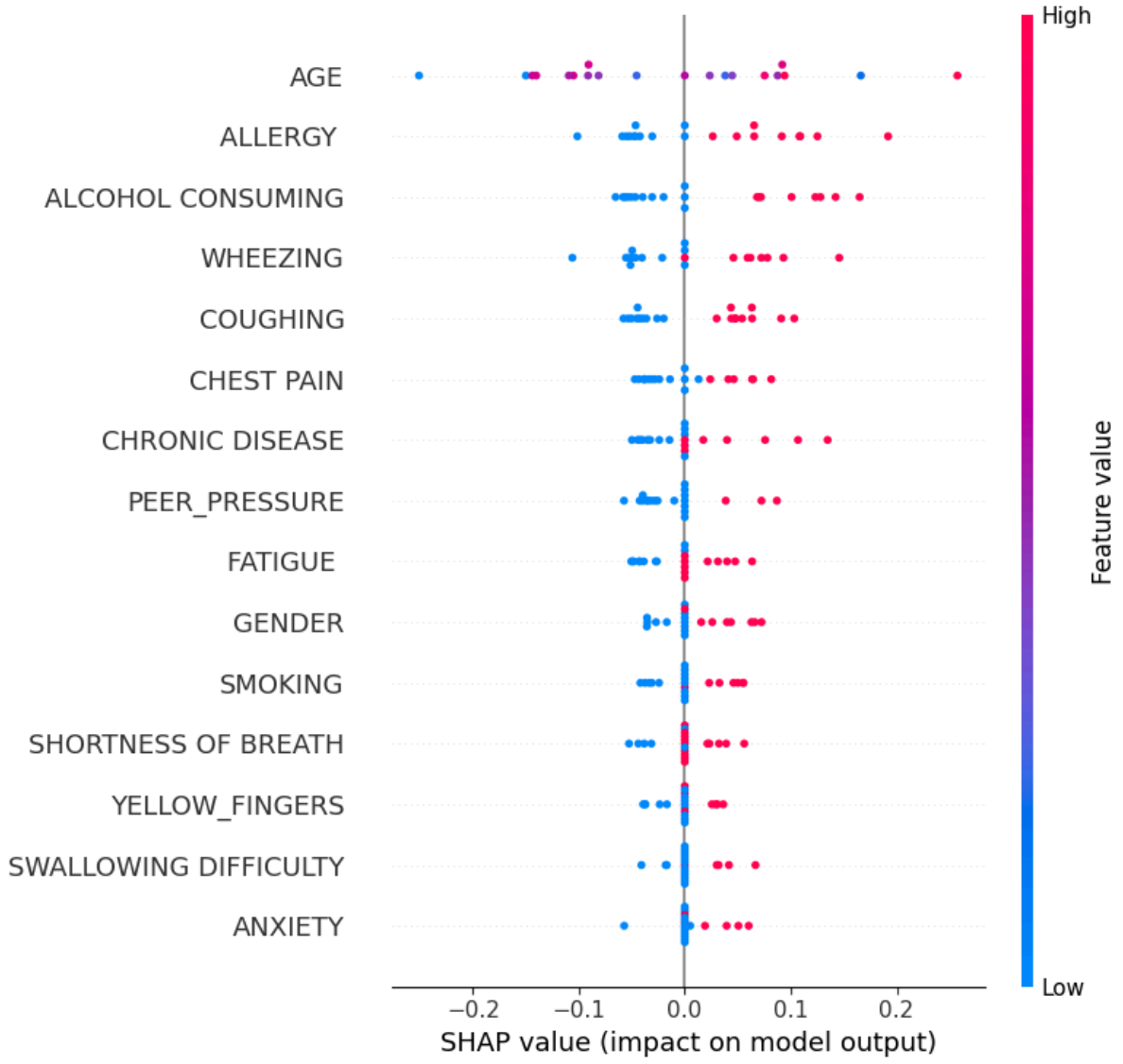


Figure 4.6: SHAP Summary Plot Demonstrating Explainability of the Ensemble Machine Learning Model for Dataset-3

CHAPTER 5

5.0: Result Discussion

In this study, we tested several machine learning models to find out how well they could predict the risk of early lung cancer. We used two different datasets and checked the models based on accuracy, precision, recall, and F1-score. From the results in Table 4.1, we saw that our proposed ensemble model gave the best performance. It reached 99.9% accuracy on Dataset-1 and 98% accuracy on Dataset-2, with equally high values for other metrics. This shows that combining multiple models can improve prediction quality.

Among the individual models, Random Forest and Decision Tree also performed very well, especially on Dataset-1, where both reached nearly 100% accuracy. However, their scores were slightly lower on Dataset-2, which may be due to some differences in the data. Support Vector Machine (SVM) and Logistic Regression gave good and consistent results across both datasets. The K-Nearest Neighbors (KNN) model also showed strong performance, but there may have been a small mistake in its F1-score for Dataset-1, which needs to be corrected.

To make the model's predictions easier to understand, we used SHAP (SHapley Additive exPlanations). SHAP helped explain which features had the biggest effect on the predictions. Features like age, smoking habits, lung disease history, and family background were the most important. The SHAP plots clearly showed how these features influenced each prediction.

Overall, the ensemble model not only gave the most accurate results but also allowed us to explain its decisions, which is very important in medical cases. This means the model can be useful in real healthcare systems for early lung cancer risk detection.

CHAPTER 6

6.0 Future Work

Although in this work we have demonstrated that ensemble machine learning method can predict lung cancer risk, the deep learning approach and more complex ensembles might be explored to provide better prediction of likelihood of cancer. In addition, using different explainability frameworks a part of SHAP could result in more interpretable explanations for the model behavior and improve the clinical usability of our findings.

6.1 Conclusion

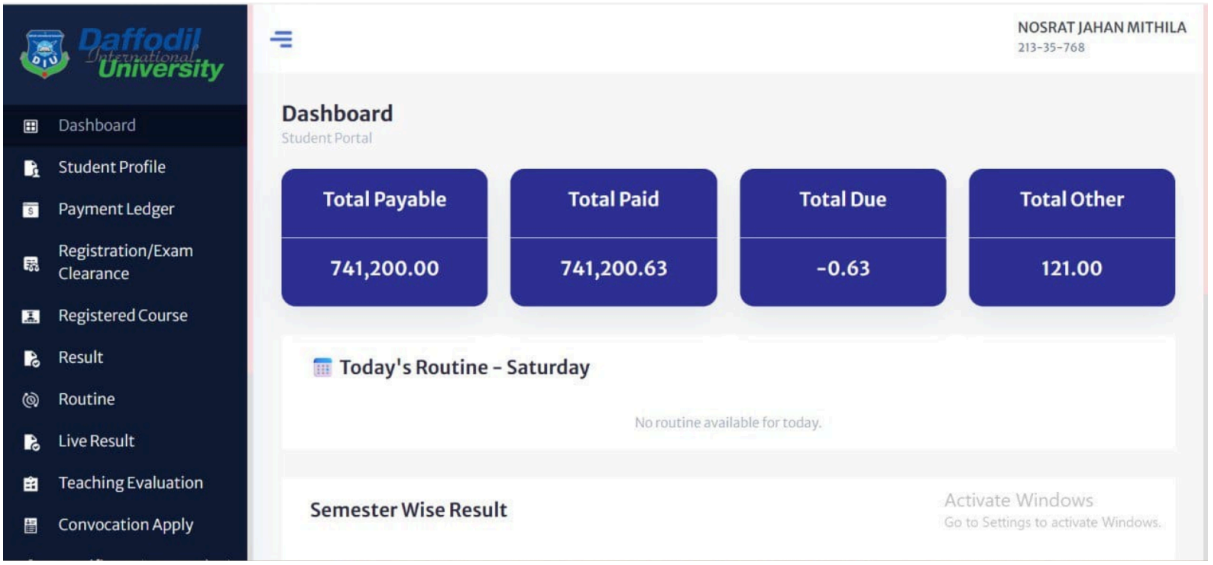
Using ensemble models and SHAP for interpretability, this study offers an explainable machine learning framework for early lung cancer risk prediction. The study shows how AI can help with early detection and healthcare decision-making by comparing several baseline algorithms and proving the superior performance of the suggested ensemble approach. Crucially, SHAP values gave clinicians a clearer understanding of feature contributions, which helped them to trust and comprehend the model's predictions. Our results demonstrate the importance of combining interpretability and predictive accuracy in delicate fields like healthcare. This work highlights the importance of interpretable ensemble models in enhancing early lung cancer risk assessment and establishes the groundwork for future developments in explainable AI-driven healthcare systems.

References:

1. *Qureshi, R., Zou, B.., Alam, T., Wu, J., Lee, V. H., & Yan, H. (2022). Computational methods for the analysis and prediction of egfr-mutated lung cancer drug resistance: Recent advances in drug design, challenges and future prospects. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 20(1), 238-255.*
2. *Melosky, B., Kambartel, K.., Haentschel, M., Bennetts, M., Nickens, D. J., Brinkmann, J., ... & Cappuzzo, F. (2022). Worldwide prevalence of epidermal growth factor receptor mutations in non-small cell lung cancer: a meta-analysis. Molecular Diagnosis & Therapy, 26(1), 7-18.*
3. *Ten Haaf, K., van der Aalst, C. M., de Koning, H.. J., Kaaks, R., & Tammemägi, M. C. (2021). Personalising lung cancer screening: An overview of risk-stratification opportunities and challenges. International journal of cancer, 149(2), 250-263.*
4. *Dutta, A. K.. (2022). Detecting Lung Cancer Using Machine Learning Techniques. Intelligent Automation & Soft Computing, 31(2).*
5. *Mienye, I. D., & Sun, Y.. (2022). A survey of ensemble learning: Concepts, algorithms, applications, and prospects. Ieee Access, 10, 99129-99149.*
6. *Sahlaoui, H., Nayyar, A.., Agoujil, S., & Jaber, M. M. (2021). Predicting and interpreting student performance using ensemble models and shapley additive explanations. IEEe Access, 9, 152688-152703.*
7. *Sumon, M.S.I., Malluhi, M.., Anan, N., AbuHaweeleh, M.N., Krzyslak, H., Vranic, S., Chowdhury, M.E. and Pedersen, S., 2024. Integrative Stacking Machine Learning Model for Small Cell Lung Cancer Prediction Using Metabolomics Profiling. Cancers, 16(24), p.4225.*
8. *Al-Jamimi, H.A., Ayad, S. and El Kheir, A., 2025. Integrating Advanced Techniques: RFE-SVM Feature Engineering and Nelder-Mead Optimized XGBoost for Accurate Lung Cancer Prediction. IEEE Access.*
9. *Sinjanka, Y., Kaur, V., Musa, U.I. and Kaur, K., 2024. ML-based early detection of lung cancer: an integrated and in-depth analytical framework. Discover Artificial Intelligence, 4(1), pp.1-18.*
10. *Chen, S. and Wu, S., 2025. Ensemble machine learning models for lung cancer incidence risk prediction in the elderly: a retrospective longitudinal study. BMC cancer, 25(1), p.126.*
11. *Flyckt, R.N.H., Sjodsholm, L., Henriksen, M.H.B., Brasen, C.L., Ebrahimi, A., Hilberg, O., Hansen, T.F., Wiil, U.K., Jensen, L.H. and Peimankar, A., 2024. Pulmonologists-Level lung cancer detection based on standard blood test results and smoking status using an explainable machine learning approach. Scientific Reports, 14(1), p.30630.*
12. *Guan, X., Du, Y., Ma, R., Teng, N., Ou, S., Zhao, H. and Li, X., 2023. Construction of the XGBoost model for early lung cancer prediction based on metabolic indices. BMC medical informatics and decision making, 23(1), p.107.*
13. *Dritsas, E. and Trigka, M., 2022. Lung cancer risk prediction with machine learning models. Big Data and Cognitive Computing, 6(4), p.139.*
14. *Janik, A., Torrente, M., Costabello, L., Calvo, V., Walsh, B., Camps, C., Mohamed, S.K., Ortega, A.L., Nováček, V., Massutí, B. and Minervini, P., 2023. Machine learning–assisted recurrence prediction for patients with early-stage non–small-cell lung cancer. JCO Clinical Cancer Informatics, 7, p.e2200062.*
15. *Hussain Ali, Y., Chinnaperumal, S., Marappan, R., Raju, S.K., Sadiq, A.T., Farhan, A.K. and Srinivasan, P., 2023. Multi-layered non-local bayes model for lung cancer early diagnosis prediction with the internet of medical things. Bioengineering, 10(2), p.138.*
16. *Gopinath, A., Gowthaman, P., Gopal, L., Walid, M.A.A., Priya, M.M. and Kumar, K.K., 2023, June. Enhanced Lung Cancer Classification and Prediction based on Hybrid Neural Network Approach. In 2023 8th International Conference on Communication and Electronics Systems (ICCES) (pp. 933-938). IEEE.*
17. *Berg, C.D., Schiller, J.H., Boffetta, P., Cai, J., Connolly, C., Kerpel-Fronius, A., Kitts, A.B., Lam, D.C., Mohan, A., Myers, R. and Suri, T., 2023. Air pollution and lung cancer: a review by International Association for the Study of Lung Cancer Early Detection and Screening Committee. Journal of Thoracic Oncology, 18(10), pp.1277-1289.*

18. *Bhuiyan, M.S., Chowdhury, I.K., Haider, M., Jisan, A.H., Jewel, R.M., Shahid, R., Ferdus, M.Z. and Siddiqua, C.U.,, 2024. Advancements in early detection of lung cancer in public health: a comprehensive study utilizing machine learning algorithms and predictive models. Journal of Computer Science and Technology Studies, 6(1), pp.113-121.*
19. *Moubarz, G., Saad-Hussein, A., Shahy, E.M., Mahdy-Abdallah, H.,, Mohammed, A.M., Saleh, I.A., Abo-Zeid, M.A. and Abo-Elfadl, M.T., 2023. Lung cancer risk in workers occupationally exposed to polycyclic aromatic hydrocarbons with emphasis on the role of DNA repair genes. International Archives of Occupational and Environmental Health, 96(2), pp.313-329.*
20. *Thanoon, M.A., Zulkifley, M.A., Mohd Zainuri, M.A.A. and Abdani, S.R., 2023. A review of deep learning techniques for lung cancer screening and diagnosis based on CT images. Diagnostics, 13(16), p.2617.*
21. *Naseer, I., Akram, S., Masood, T., Rashid, M., & Jaffar, A. (2023). Lung cancer classification using modified U-Net based lobe segmentation and nodule detection. IEEE Access, 11, 60279-60291.*
22. *Guan, P., Yu, K., Wei, W., Tan, Y., & Wu, J. (2023). Big data analytics on lung cancer diagnosis framework with deep learning. IEEE/ACM transactions on computational biology and bioinformatics, 21(4), 757-768.*
23. *Mamun, M., Farjana, A., Al Mamun, M., & Ahammed, M. S. (2022, June). Lung cancer prediction model using ensemble learning techniques and a systematic review analysis. In 2022 IEEE World AI IoT Congress (AIIoT) (pp. 187-193). IEEE.*
24. *Chandrasekar, T., Raju, S. K., Ramachandran, M., Patan, R., & Gandomi, A. H. (2022). Lung cancer disease detection using service-oriented architectures and multivariate boosting classifiers. Applied Soft Computing, 122, 108820.*
25. *Bharathy, S., & Pavithra, R. (2022, May). Lung cancer detection using machine learning. In 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC) (pp. 539-543). IEEE.*
26. *Khan, V. A., Yadav, A. K., Arshad, M., & Akhtar, N. (2025). Lung Cancer Prediction Using an Enhanced Neutrosophic Set Combined with a Machine Learning Approach. Neutrosophic Sets and Systems, 88(1), 64.*
27. *Hussain, L., Alsolai, H., Hassine, S. B. H., Nour, M. K., Duhayyim, M. A., Hilal, A. M., ... & Rizwanullah, M. (2022). Lung cancer prediction using robust machine learning and image enhancement methods on extracted gray-level co-occurrence matrix features. Applied Sciences, 12(13), 6517.*
28. *Yamini, B., Sudha, K., Nalini, M., Kavitha, G., Subramanian, R. S., & Sugumar, R. (2023, June). Predictive modelling for lung cancer detection using machine learning techniques. In 2023 8th International Conference on Communication and Electronics Systems (ICCES) (pp. 1220-1226). IEEE.*
29. *Shakya, S. R., Ceh-Varela, E., & Sanjaya, I. (2025, June). Lung Cancer Classification Using Deep Learning Models for Edge Computing. A Comparative Analysis. In 2025 IEEE International Conference on AI and Data Analytics (ICAD) (pp. 1-6). IEEE.*
30. *Provath, M. A. M., Deb, K., Dhar, P. K., & Shimamura, T. (2023). Classification of lung and colon cancer histopathological images using global context attention based convolutional neural network. IEEE Access, 11, 110164-110183.*

ACCOUNTS CLEARANCE



PLAGIARISM REPORT

213-35-768

ORIGINALITY REPORT

20%	14%	16%	11%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.frontiersin.org Internet Source	1%
2	Submitted to Tampere University Student Paper	1%
3	findresearcher.sdu.dk Internet Source	1%
4	S.P. Jani, M. Adam Khan. "Applications of AI in Smart Technologies and Manufacturing", CRC Press, 2025 Publication	1%
5	"Proceeding of the 2nd International Conference on Machine Intelligence and Emerging Technologies", Springer Science and Business Media LLC, 2025 Publication	1%
6	eprint.innovativepublication.org Internet Source	1%
7	vbn.aau.dk Internet Source	1%
8	migrationletters.com Internet Source	1%
9	Mohan Timilsina, Samuele Buosi, Maria Torrente, Mariano Provencio et al. "Large language model vs. traditional machine learning: Evaluating predictive models for	1%