Lung Cancer Prediction Using Machine Learning: A Data-Driven Approach for Early Diagnosis

Bardh Ademi Bimal Kumal Srijana Shrestha

March 5, 2025

1. Introduction

"More than 80% of people with lung cancer die within five years of diagnosis—and more than half within the first year. Lung cancer causes death in a few different ways. Tumors can grow and spread so much that organs can no longer function. The cancer can also lead to fatal infections or blood clots" (Lynne Eldridge, MD) [8]. An alarming statistic as such makes the following research and data analysis very important, and should be taken seriously by everyone, regardless of age or health condition. Cancer is a word that gives people the chills, and the bad news is there is a lot of data with patients who suffer from Lung Cancer. The good news is that through the use of data science, this data can be analyzed to create good in the world and help prevent the progression of the disease.

Lung cancer, similar to other cancers, often are not diagnosed until symptoms rise to a non-tolerable stage. At this point, patients have no choice but to see a professional and receive treatment. By shedding light on the correlation and causation between our predictors and lung cancer levels, patients and others will be highly informed of the hazards associated with these variables, which should in turn reward them with an understanding of how important it is to keep one's health in check periodically. Ultimately, the goal is to build a model that can support patients and healthcare professionals in prioritizing one's health before receiving a diagnosis.

This project utilizes data science techniques to develop a model capable of using an individual's reported lung cancer level, and predict how likely their cancer is to grow into higher levels of severity based on various

environmental, genetic, and lifestyle-related factors. By identifying key risk and symptom-based indicators, analyzing these indicators using visualizations, and using them to create a machine learning model, this research project will aim to enhance early diagnosis, treatment outcomes including self-discipline to reduce the leading factors, and inform the target audience with prevention strategies to ensure patients live the longest, and healthiest life possible.

2. Methodology

The dataset was first cleaned by removing irrelevant columns, converting categorical variables into appro priate formats, and adjusting skewed features to improve model performance. Exploratory data analysis (EDA) was conducted using correlation matrices and visualizations to identify potential predictors of lung cancer severity. To further explore statistical relationships between variables and the target outcome, tests such as Spearman correlation, Kruskal-Wallis, and Chi-square were applied. A Random Forest classification model was then trained on a subset of selected risk factors and evaluated using a confusion matrix to assess prediction accuracy. Eventually, a prediction function was developed to estimate the probability of lung cancer severity based on individual patient input.

2.1 Dataset Overview

The dataset used in this project is titled "Lung Cancer Prediction Dataset", sourced from Kaggle [6]. It is open-access and publicly available for research and educational use. The dataset consists of 1,000 patient records and 26 variables, encompassing demographic details, risk factors, symptoms, and the target variable. Below is a breakdown of the key components:

Demographic Data:

- Age
- Gender
- Patient ID (removed during preprocessing)

- Risk Factors:
- Air Pollution
- Alcohol Use
- Smoking
- Genetic Risk
- Chronic Lung Disease
- Occupational Hazards
- Balanced Diet
- Obesity
- Passive Smoking

Symptoms:

- Chest Pain
- Coughing of Blood
- Fatigue
- Weight Loss
- Shortness of Breath
- Wheezing
- Swallowing Difficulty
- Clubbing of Finger Nails
- Frequent Cold

- Dry Cough
- Snoring

Target Variable:

- Level: Represents the severity of lung cancer (Low, Medium, High)

2.2 Data Cleaning

The dataset underwent several preprocessing steps to ensure consistency and prepare it for analysis:

- **Index and Patient ID Removal**: The `index` and `Patient.Id` columns were removed to preserve anonymity and eliminate irrelevant identifiers.
- **Standardization of Column Names**: All column names were cleaned and standardized using the `janitor` package to ensure consistent formatting.
- **Missing Value Check**: The dataset was checked for missing values, and no missing data was found.
- **Risk Level Encoding**: The categorical variable `level` (risk level: Low, Medium, High) was converted into an ordered factor and then into a numeric variable (`level_numeric`) to support both modeling and correlation analysis.
- **Balanced Diet Recoding**: The `balanced_diet` variable was reversed so that higher values now indicate less healthy diets, improving interpretability during correlation analysis.

These steps helped ensure that the data was clean, structured, and ready for reliable statistical and machine learning analyses.

3 Exploratory Data Analysis

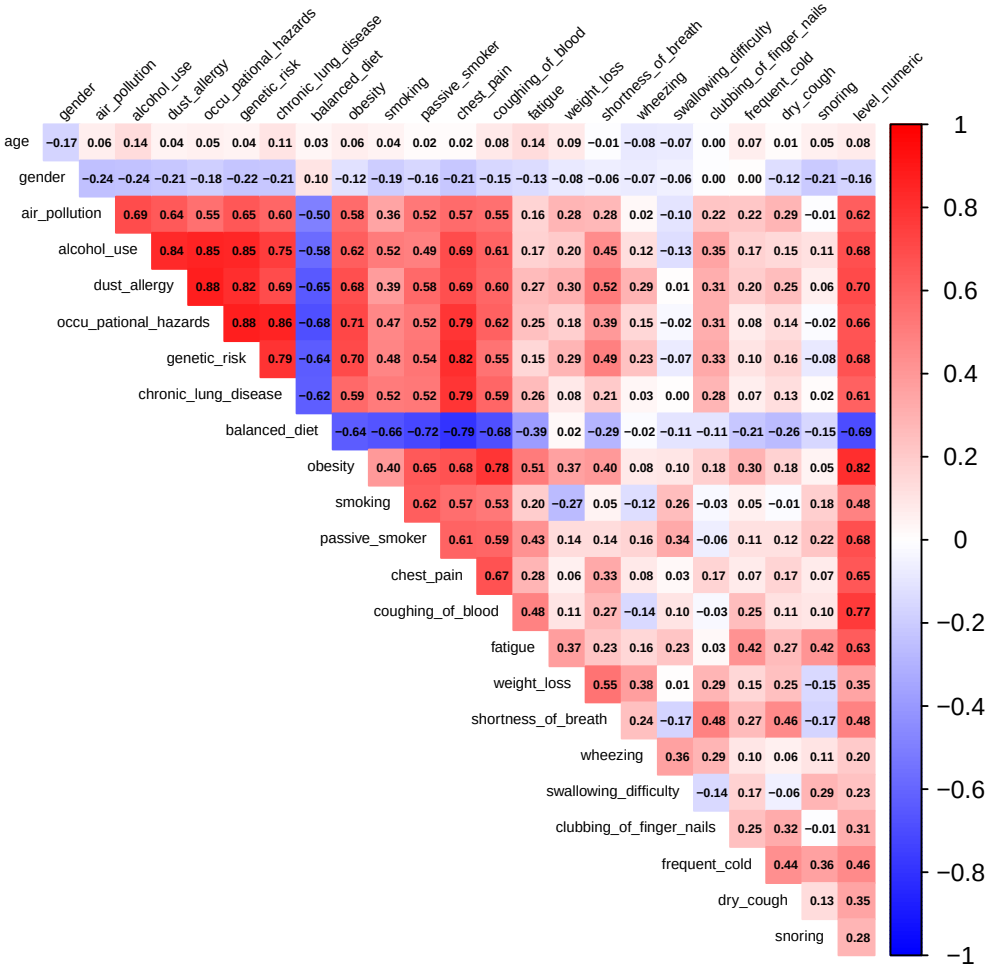
To better understand the dataset and identify potential relationships between variables and lung cancer risk, a series of exploratory data analyses was conducted. This process included visualizing the distributions of

key features, examining correlations between variables, and comparing feature values across different lung cancer risk levels (Low, Medium, High).

3.1 Spearman Correlation Heatmap

A correlation heatmap was generated to assess the strength and direction of monotonic relationships among numerical variables using Spearman's correlation method. This was followed by the use of boxplots, bar plots, and statistical summaries to explore how environmental, genetic, and lifestyle factors differ across risk categories. These visualizations provided initial insights into variables that may be most predictive of lung cancer severity and guided model development and hypothesis testing in subsequent stages of the analysis.

Figure 1: Spearman Correlation Heatmap of Key Variables

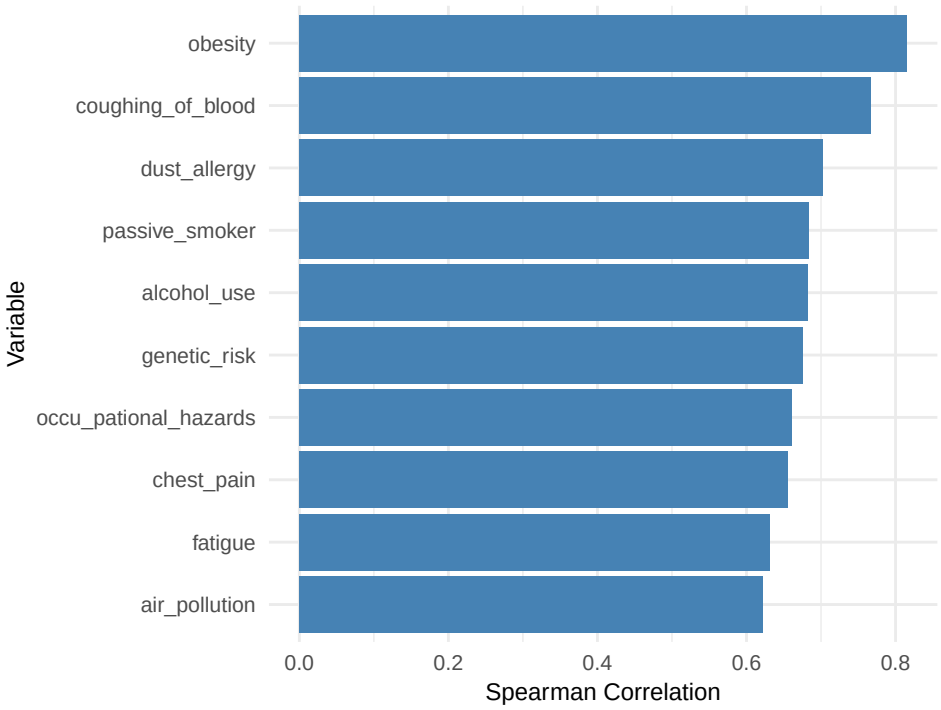


A Spearman correlation heatmap revealed that several variables were strongly associated with lung cancer risk. Obesity exhibited the highest correlation (r = 0.82), although this may reflect dataset bias rather than clinical significance. Other established risk factors, such as passive smoking (r = 0.68), alcohol use (r = 0.68), air pollution (r = 0.62), and symptoms including coughing of blood (r = 0.77) and chest pain (r = 0.65), also demonstrated strong positive correlations. Balanced diet showed a negative correlation (r = –0.69), consistent with its protective role. These findings emphasize key environmental, behavioral, and symptomatic contributors to lung cancer severity.

3.2 Top Ten Strongest Predictors

To further identify the most influential variables associated with lung cancer severity, the top 10 features most strongly correlated with the ordinal severity levels (Low = 1, Medium = 2, High = 3) were extracted. As illustrated in Figure 2, this focused bar plot provides a clearer depiction of how specific features are related to disease progression.

Figure 2: Top 10 Features Correlated with Lung Cancer



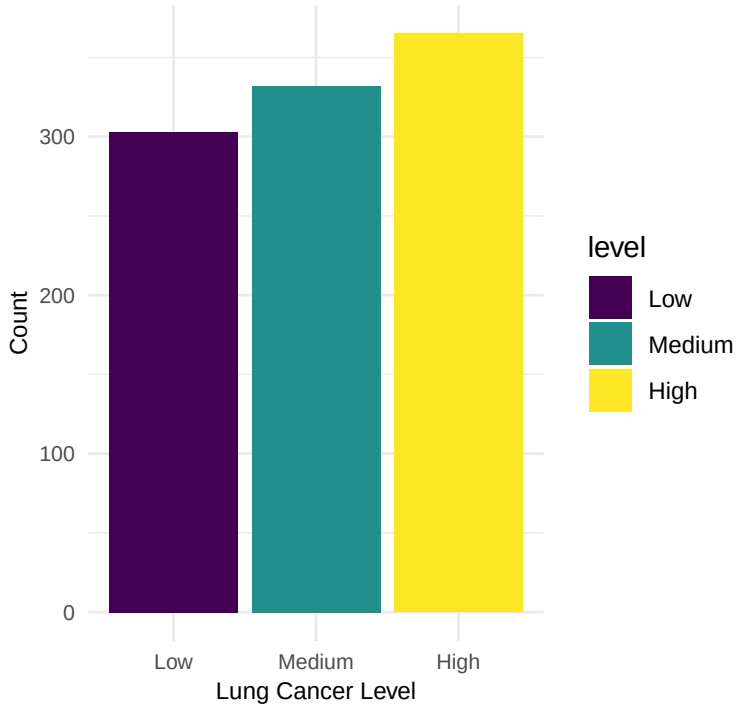
The analysis highlights obesity (correlation = 0.82), coughing of blood (0.77), dust allergy (0.70), genetic risk

(0.68), passive smoking (0.68), and alcohol use (0.68) as the features most strongly associated with lung cancer severity. Environmental exposures such as occupational hazards (0.66) and air pollution (0.62) also exhibit substantial positive correlations. These findings were obtained using Spearman's rank correlation, a non-parametric method suitable for capturing monotonic relationships between ordinal and numeric variables—even when the relationships are not strictly linear.

Interestingly, the strongest correlation was found with obesity, a factor typically associated with other cancers (e.g., breast or colon), but not prominently featured in lung cancer literature. This may point to either dataset-specific patterns or the presence of hidden confounding variables. Nonetheless, the consistency of high correlations across known clinical symptoms (e.g., coughing of blood, chest pain, fatigue) and well-established risk factors (e.g., smoking, air pollution) supports the reliability of the dataset and provides a data-driven rationale for variable selection in subsequent predictive modeling.

3.3 Lung Cancer Distribution

Figure 3: Distribution of Lung Cancer Levels

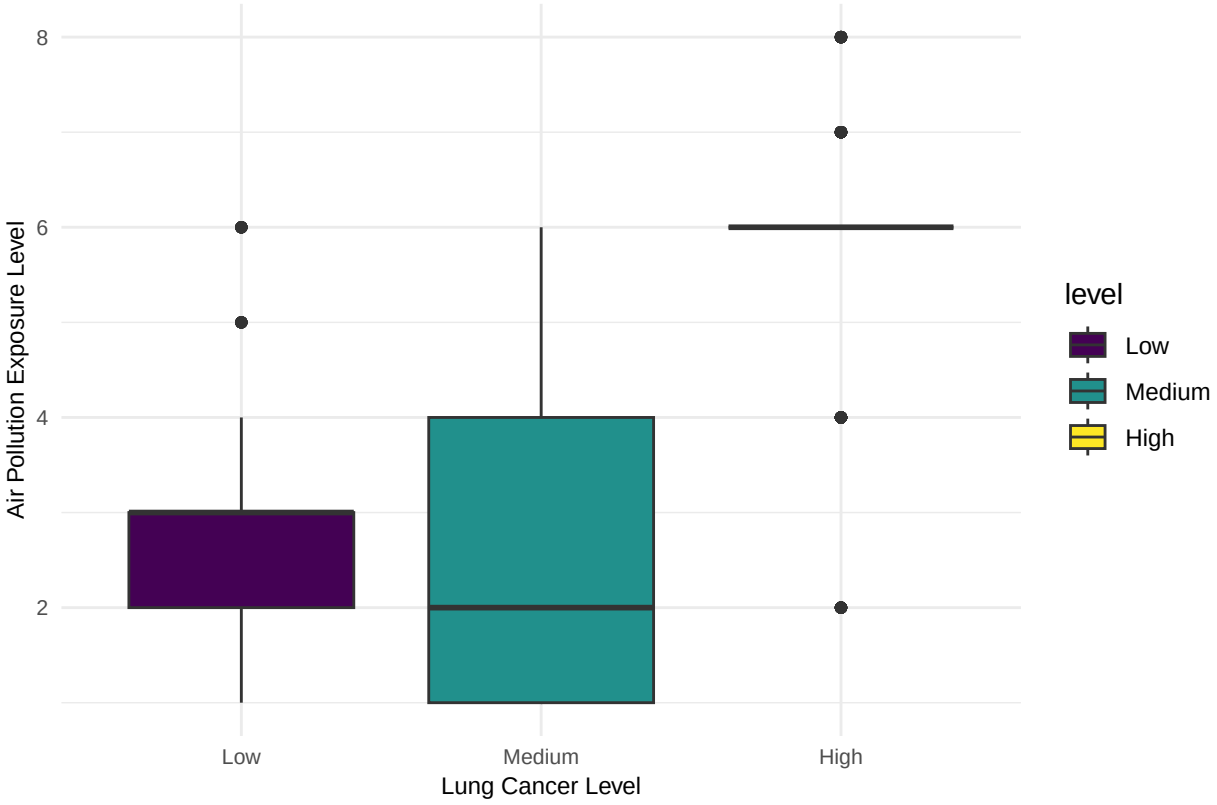


To better understand the composition of the dataset,the distribution of lung cancer risk levels was visualized

across the three categories: Low, Medium, and High. As shown above in figure 3, the classes are relatively balanced, with each category containing a comparable number of records. This balance is crucial for machine learning applications, as it minimizes the risk of model bias toward the majority class and promotes more stable and generalizable predictions. A well-balanced dataset also ensures that evaluation metrics such as accuracy, precision, and recall are meaningful and not skewed by class imbalance. Moreover, it enables more consistent comparisons across risk levels when analyzing feature associations and model behavior throughout the study.

3.4 Air Pollution Across Lung Cancer Levels

Figure 4: Air Pollution Exposure vs Lung Cancer Level



According to figure 4 above, there is a clear relationship present between the air pollution and lung cancer level of the patients. The boxplot showed that patients with a high level of lung cancer experienced higher levels of air pollution, while patients with a low and medium level of lung cancer experienced lower levels of air pollution. Patients in the "High" cancer level category had a median value of 6, while patients with

"Medium" and "Low" cancer levels had medians of 2 and 3.

Using statistical analysis, a Kruskal-Wallis rank sum test was generated to determine whether there is a significant difference between the Low, Medium, and High cancer levels. This was done due to the fact that the three levels may be perceived as close to one another. Therefore, for accurate and clear results, a Kruskal-Wallis rank sum test presented the accuracy and results desired. The test did reveal a difference among the three levels, with a chi-squared value of 463.14 and a p-value less than 2.2e-16. With a degree of freedom being 2, and choosing a significance level of 0.05, the critical value is 5.9914645, which is lower than the observed chi-squared value. Therefore, the null hypothesis can be rejected and a conclusion can be confidently made that the distribution of High, Medium, and Low cancer levels differ in air pollution exposure.

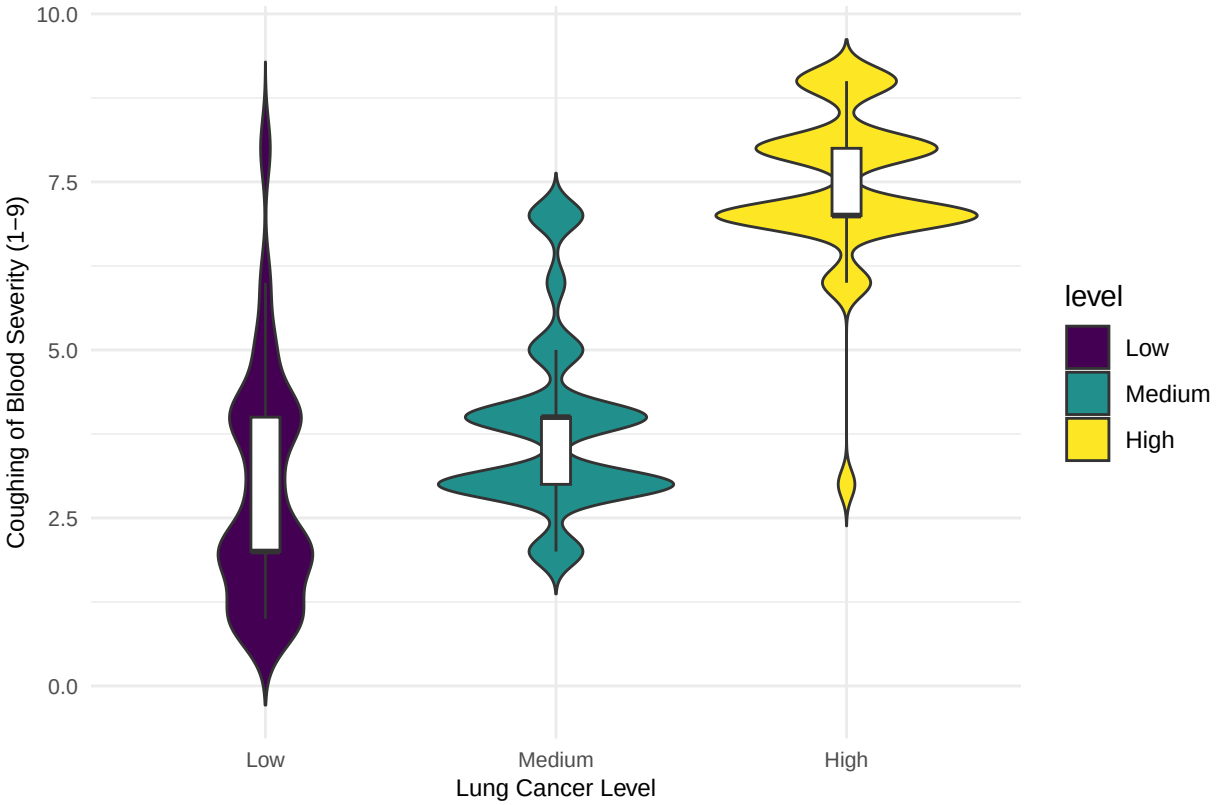
Additionally, The US Environmental Protection Agency states that "airway inflammation increases airway responsiveness to irritants (e.g., cold air, particle pollution, allergens, lipopolysaccharides, and gaseous pollutants) and may reduce lung function by causing bronchoconstriction. At a cellular level, inflammation may damage or kill cells and compromise the integrity of the alveolar-capillary barrier. Repeated exposure to particle pollution aggravates the initial injury and promotes chronic inflammation with cellular proliferation and extracellular matrix reorganization" (Berend, 2016) [7]. Scientifically, it is apparent that air pollution is damaging to the respiratory system, which further corroborates the distribution revealed in figure 4.

These results suggest that patients in the High lung cancer risk group were experiencing much higher levels of air pollution compared to those in the Low and Medium groups, highlighting the relationship between extreme pollution exposure and high lung cancer severity.

3.5 Coughing of Blood Across Lung Cancer Levels

Patients who reported coughing of blood had the second highest correlation value (0.77). For further investigation, a violin plot was generated (see figure 5) to highlight the relationship between the reported coughing of blood across the Low, Medium, and High cancer severities. This plot revealed that patients who were categorized as suffering from high levels of lung cancer also reported a high severity of blood in their cough.

Figure 5: Distribution of Coughing of Blood Severity Across Lung Cancer Level



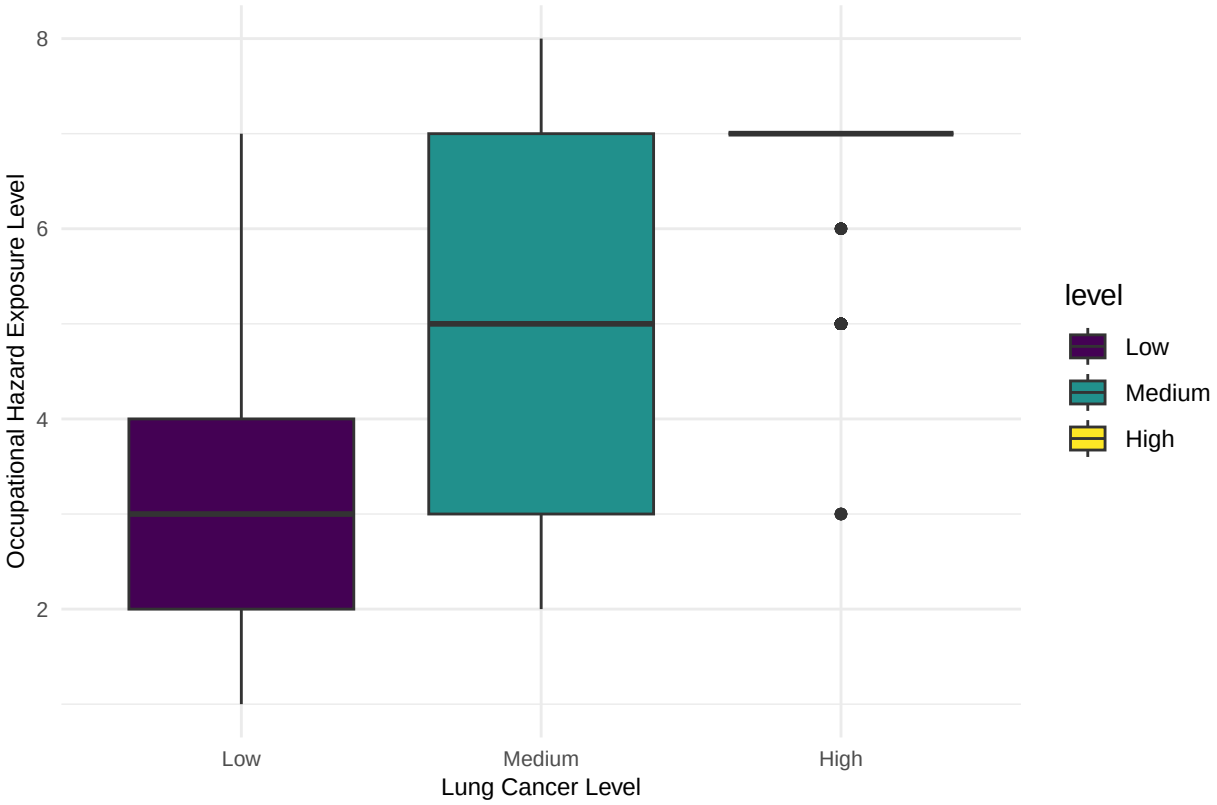
The shape of the violins is a strong representation of the density of patients whom reported each level of coughing severity within each cancer level. Patients that were diagnosed with high lung cancer fell toward the higher severity scores in terms of reported coughing of blood. Although the Low and Medium patients seem somewhat similar in terms of their reported coughing of blood severity, High cancer-level patients are much further up with a higher concentration of patients (wide violin) from 6.25 to 9 as their reported levels of coughing.

On the other hand, patients in the Low cancer severity group display a narrower distribution, skewed toward the lower severity scores (between 1 and 5), suggesting that coughing of blood is much less severe in these patients. The Medium cancer severity group has a similar violin shape to the High group, but is closer to the Low group in terms of median and overall box and whiskers placement. Such distribution and median is expected, since patients in this category do not have the highest lung cancer level, but are still experiencing moderate coughing of blood.

This boxplot visually strengthens the relationship between coughing of blood and lung cancer levels. By visualizing a violin for each level, an added feature of distribution is added to the figure, a key technique in data analysis when searching for trends. While correlation alone does not imply causation, this distribution supports the hypothesis that in most patients, coughing of blood could serve as a strong clinical indicator for elevated cancer severity, warranting further medical investigation.

3.6 Occupational Hazards Across Lung Cancer Levels

Figure 6: Occupational Hazards vs Lung Cancer Level

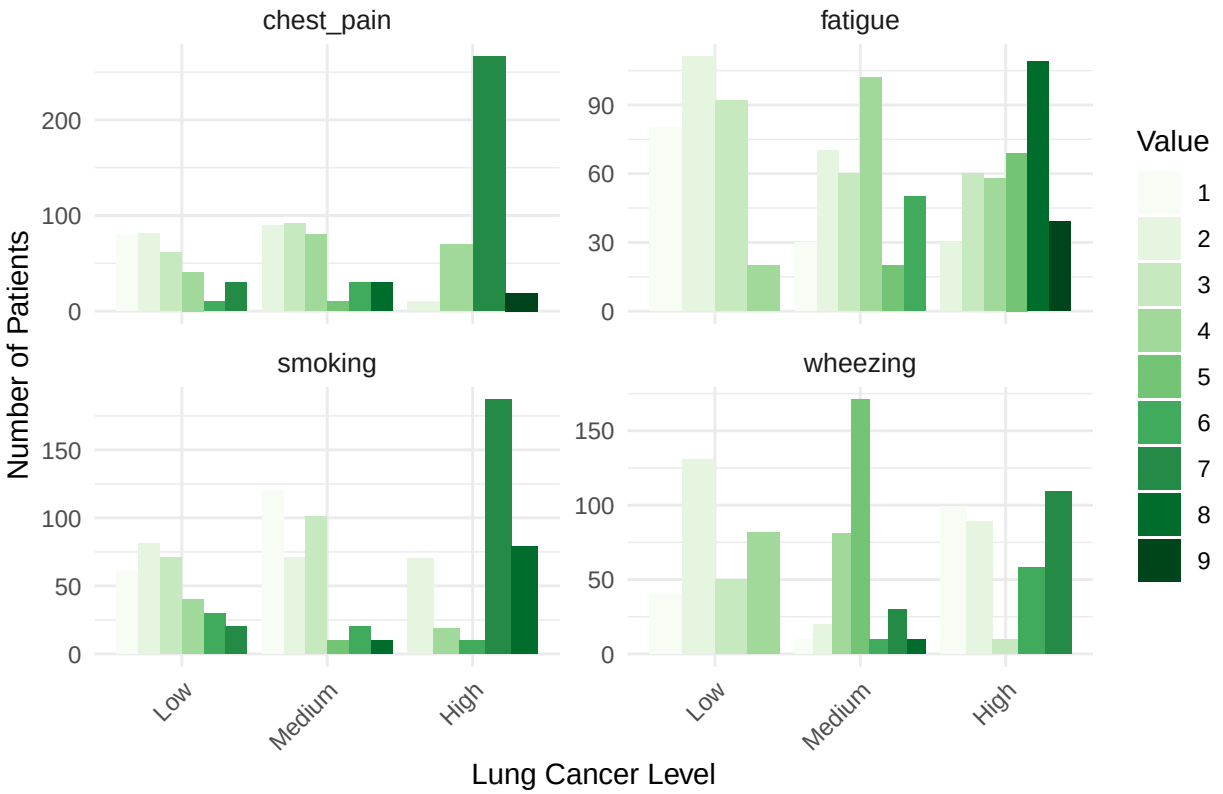


To evaluate the relationship between occupational hazard exposure and lung cancer level, a boxplot was created (Figure 6) comparing occupational hazards across the cancer level groups. The boxplot revealed that individuals in the High-cancer group experienced much higher occupational hazard exposure compared to those in the Low and Medium groups. A Kruskal-Wallis rank sum test confirmed that these differences were highly significant (chi-squared = 436.58, p-value < 2.2e-16). These findings strongly suggest that occupational exposure to hazardous environments is an important factor contributing to lung cancer level.

3.7 Multiple Predictors Across Lung Cancer Risk Levels

The faceted bar plots below in figure 6 visualize the distribution of lung cancer risk levels across different levels of key predictors—such as chest pain, fatigue, smoking, and wheezing. The height of each bar represents the number of patients associated with their assigned predictor, allowing for a clear comparison of how each factor relates to cancer risk across subgroups.

Figure 7: Distribution of Predictors Across Lung Cancer Levels



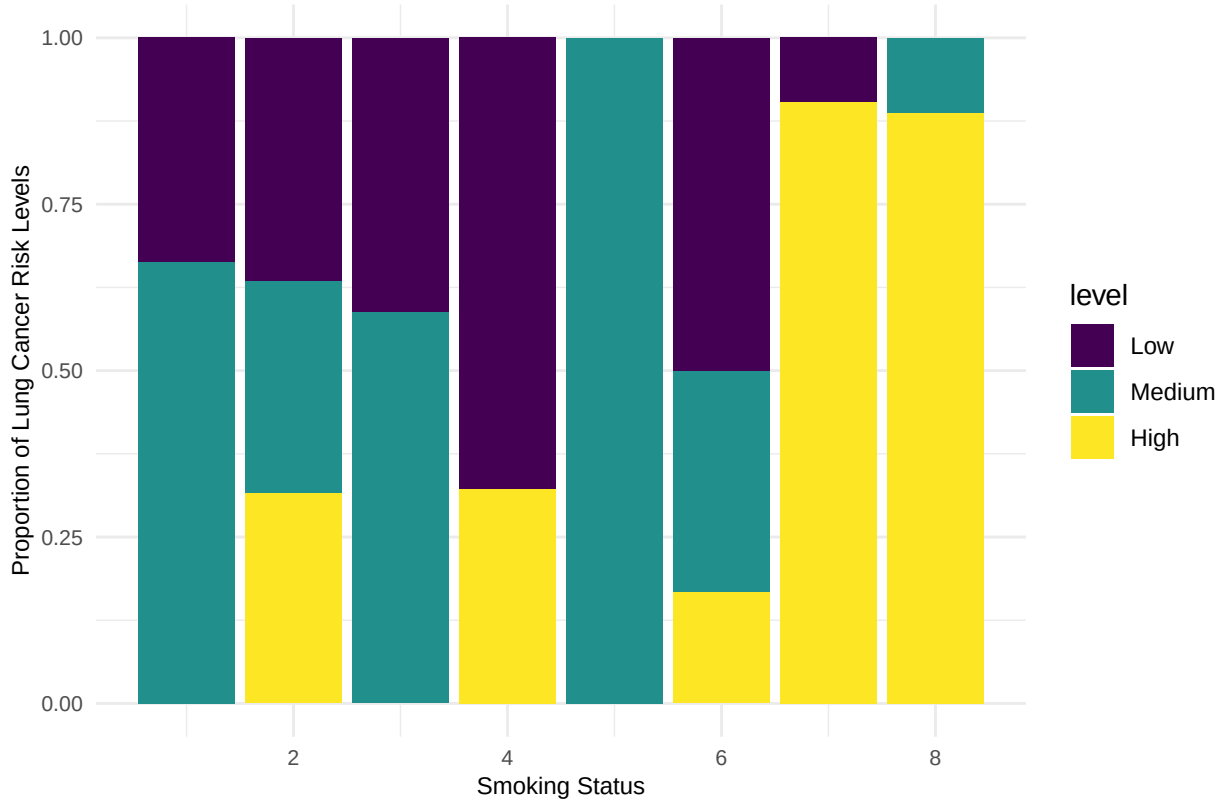
The colors of each bar represent different levels of the predictor variable. For example, smoking frequency from "None" to "Heavy" (1 to 9). Furthermore, patients with a higher reports of chest pain (darker green) have a higher lung cancer level. This result is present in patients with fatigue, smoking, as well as wheezing, highlighting that patients who report these predictors are more likely of increasing their level of lung cancer, and further progressing their disease.

These faceted plots provide intuitive, visual evidence of the predictive power of these variables. They also highlight differences between the four predictors. Chest pain, fatigue, and smoking show a strong relationship

to lung cancer, while wheezing shows a weaker correlation. There is a strong possibility that this is due to the fact that wheezing is a symptom of all lung cancer patients, no matter what level. Nonetheless, such visualizations are critical for identifying high-impact risk factors and for developing targeted prevention strategies.

3.8 Smoking Severity Across Lung Cancer Levels

Figure 8: Smoking Status vs Lung Cancer Risk Level



A proportional bar plot (Figure above) was generated to visualize the relationship between smoking status and lung cancer risk levels. The plot clearly demonstrates that as smoking exposure increases, the proportion of individuals classified under High risk increases dramatically. Individuals with lower smoking statuses (e.g., 0–2) tend to have greater proportions of Low and Medium risk outcomes, whereas higher smoking statuses (6–8) are predominantly associated with High lung cancer risk. A Pearson's Chi-square test was conducted to statistically assess the relationship between smoking status and lung cancer risk level.

4. Statistical Testing

Interestingly, both air pollution exposure and occupational hazard exposure demonstrated very strong associations with lung cancer risk, as evidenced by similar Kruskal-Wallis test results. This may reflect the interconnected nature of environmental risk factors, where individuals exposed to one harmful condition (e.g., polluted air) are often simultaneously exposed to occupational hazards, compounding their overall risk.

The Pearson's Chi-square test yielded a highly significant result (chi-squared = 684.5, df = 14, p-value < 2.2e-16), indicating a strong association between smoking behavior and lung cancer risk levels. Although a warning regarding small expected cell counts was noted, the large sample size and extremely small p-value support the reliability of this finding.

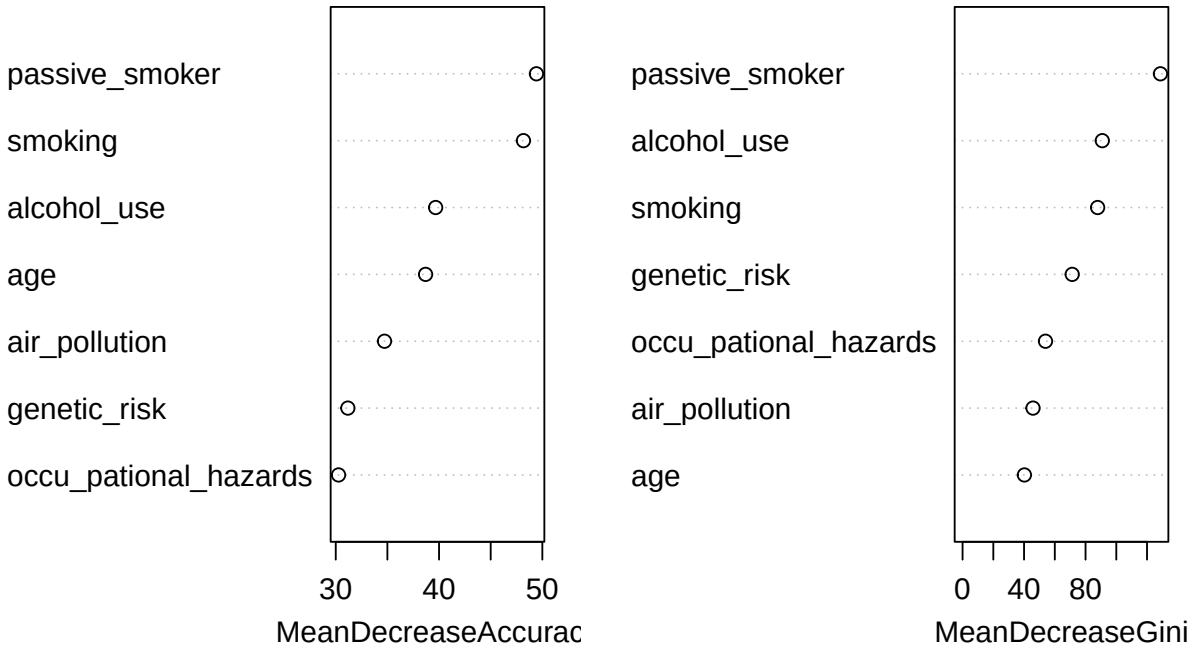
5. Machine Learning Models

5.1 Risk-Based Model

A Random Forest classification model was constructed using major risk factors, including air pollution exposure, smoking status, passive smoking exposure, genetic predisposition, occupational hazards, alcohol use, and age. The model achieved an overall accuracy of 98.99% on the test dataset, with a 95% confidence interval ranging from 96.42% to 99.88%.

Feature importance analysis (See Figure 9) revealed that passive smoking, active smoking, and alcohol use were the most influential predictors of lung cancer risk, followed by age and air pollution. These results underscore the dominant role of lifestyle and environmental exposures in determining lung cancer risk and demonstrate the potential of machine learning models for early risk prediction.

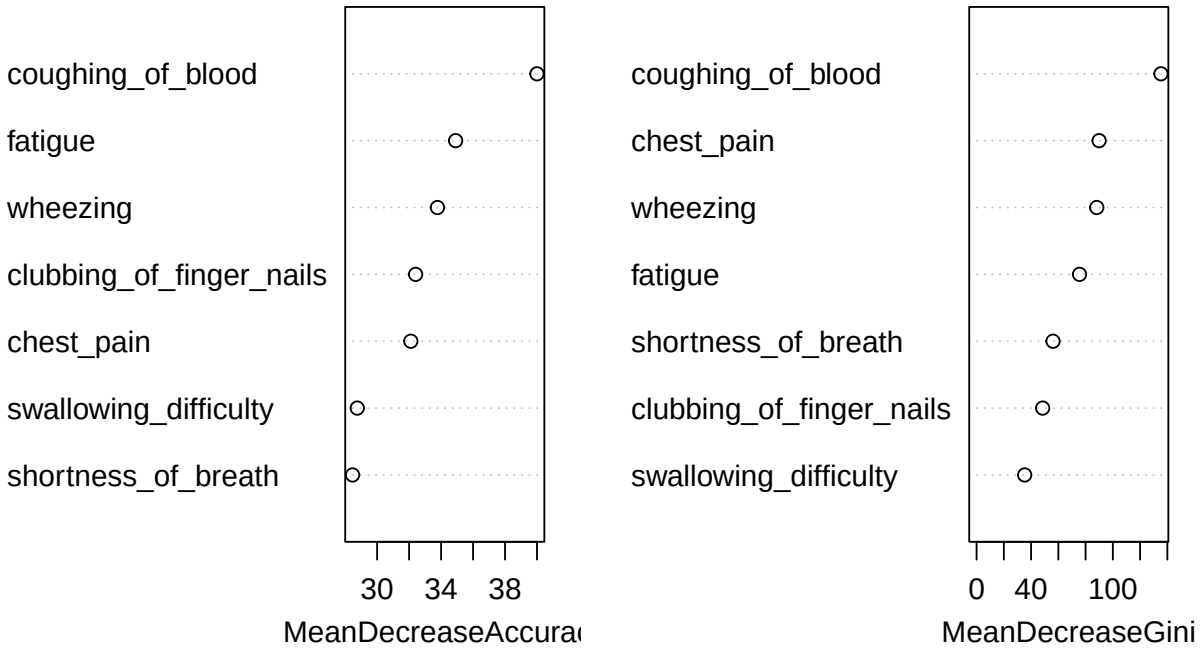
Figure 9: Feature Importance (Risk−Based Model)



5.2 Symptom-Based Model

A Random Forest model was trained using only symptom-related features to evaluate how well these clinical symptoms could predict lung cancer severity.

Figure 10: Feature Importance (Symptom–Based Model)



As shown in Figure 10, coughing of blood, chest pain, and wheezing were the most influential features in predicting lung cancer severity. While the risk-based model performed exceptionally well, the symptom-based model achieved perfect classification, which feels almost too good to be true. The 100% accuracy and flawless performance across all metrics raise the possibility of overfitting or some underlying data bias. In real-world clinical scenarios, such perfection is rare—so something might be a little fishy here. Further validation with external datasets is needed to confirm whether this model is truly generalizable.

Still, both models highlight the importance of environmental and symptomatic factors in predicting lung cancer severity.

6. Results

6.1 Correlation Analysis Summary

Obesity showed the strongest positive correlation (0.82) with lung cancer severity, followed by coughing of blood (0.77), dust allergy (0.70), and passive smoking (0.68). Other notable factors with moderate positive correlation included alcohol use, genetic risk, and occupational hazards. Weaker correlations were observed for age (0.08), wheezing (0.20), and swallowing difficulty (0.23). Gender and balanced diet showed negative correlations, with balanced diet having a moderately strong negative correlation (-0.69), suggesting a possible protective effect.

6.2 Kruskal-Wallis Test Results

The Kruskal-Wallis test was used to assess whether there are statistically significant differences in exposure levels across lung cancer risk categories.

- Air Pollution Exposure vs Lung Cancer Risk Level The Kruskal-Wallis test yielded a chi-squared value of 463.14 with 2 degrees of freedom and a p-value $< 2.2e\text{-}16$, indicating a statistically significant difference in air pollution exposure among different lung cancer risk levels.
- Occupational Hazard Exposure vs Lung Cancer Risk Level The test produced a chi-squared value of 436.58 with 2 degrees of freedom and a p-value $< 2.2e\text{-}16$, also suggesting a significant difference in occupational hazard exposure across lung cancer risk groups.

These results imply that both air pollution and occupational hazards are significantly associated with variations in lung cancer risk levels.

6.3 Chi-Square Test Results

The Pearson Chi-square test was applied to evaluate the association between categorical variables—specifically, smoking behavior and lung cancer severity levels. The test yielded a chi-squared value of 684.5

with 14 degrees of freedom and a p-value less than 2.2e-16, indicating a statistically significant relationship. This result provides strong evidence that varying levels of smoking exposure are associated with increased lung cancer risk.

While a warning was issued regarding small expected counts in some cells, the large sample size and extremely low p-value suggest the result is robust. These findings reinforce the role of smoking as a critical lifestyle-related risk factor in lung cancer development.

6.4 Random Forest Model Results

Two Random Forest classification models were developed: one using risk-related factors and the other using symptom-based features.

Risk-Based Model :The model was trained on features such as smoking, passive smoking, alcohol use, air pollution, genetic risk, occupational hazards, and age.

- Accuracy: 98.99%
- Kappa: 0.9849 (almost perfect agreement)
- Balanced Accuracy: Above 98% for all classes
- Sensitivity: 96.7% (Low), 100% (Medium, High)

Most important features: Passive smoking, smoking, and alcohol use, followed by genetic risk and occupational hazards

Symptom-Based Model :This model used clinical symptoms such as chest pain, coughing of blood, shortness of breath, fatigue, and others.

- Accuracy: 100%
- Kappa: 1.00 (perfect agreement)
- Sensitivity, Specificity, and Predictive Values: 1.00 for all classes

- Balanced Accuracy: 1.00 for all risk levels

Top contributing symptoms: Coughing of blood, shortness of breath, fatigue, and chest pain

The symptom-based model outperformed the risk-based model slightly, achieving perfect classification. However, both models demonstrated strong predictive power, highlighting the importance of both environmental exposures and clinical signs in lung cancer severity prediction.

6.5 Confusion Matrix and Accuracy

The confusion matrices confirm the reliability and precision of the models:

Risk-Based Model Confusion Matrix:

	Predicted Low	Predicted Medium	Predicted High
Actual Low	58	2	0
Actual Medium	0	66	0
Actual High	0	0	73

- **Accuracy**: 98.99%
- **Balanced Accuracy**: >98% for all classes
- Minor misclassification: 2 Medium-risk cases predicted as Low-risk

Symptom-Based Model Confusion Matrix:

	Predicted Low	Predicted Medium	Predicted High
Actual Low	60	0	0
Actual Medium	0	66	0
Actual High	0	0	73

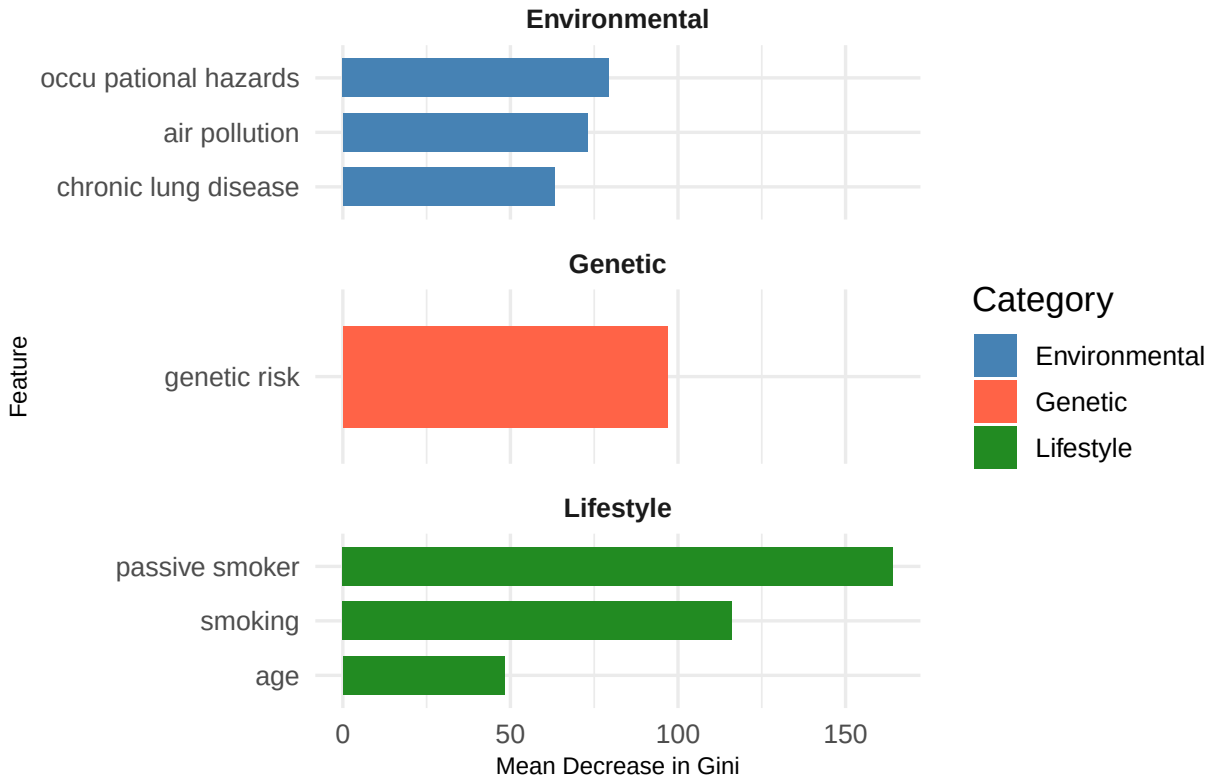
- **Accuracy**: 100%
- **Balanced Accuracy**: 100% for all classes
- No misclassifications

The results demonstrate that both models can effectively classify lung cancer severity, with the symptom-based model achieving perfect classification on the test set. However, due to the extremely high performance of Symptoms-Based Model, further investigation is recommended to ensure there is no potential data leakage.

6.6 Feature Importance Analysis Based on Predictors

To gain deeper insight into how specific variables contribute to lung cancer severity prediction, feature importance scores was analyzed from the Random Forest model. The importance of each feature was measured using the Mean Decrease in Gini index, which reflects the contribution of each variable to reducing classification impurity across the trees in the ensemble. To enhance interpretability, features were categorized into three domains—Genetic, Environmental, and Lifestyle factors—and visualized in a faceted bar plot. This grouping allows for a more intuitive understanding of which risk domains have the greatest predictive influence, and which individual variables stand out within each group. The resulting plot not only confirms the significance of known risk factors such as smoking and passive smoking, but also highlights the added predictive value of environmental exposures and genetic predispositions.

Figure 11: Feature Importance: Genetic, Environmental, and Lifestyle Factors



7. Discussion

This study explains how machine learning models can predict the seriousness of lung cancer based on multiple risk factors and symptom data. The research revealed a number of factors that exhibited a substantial direct relationship to lung cancer severity, such as obesity and blood coughing, alongside passive smoking and alcohol use, and dust allergy. Although this analysis identified obesity as a powerful correlation(r=0.82), this may reflect dataset bias rather than clinical significance. Some studies showed that obesity is associated with other cancers (e.g., breast or colon), but not prominently featured in lung cancer.

Two machine learning models were implemented using the Random Forest algorithm to create two different classifiers that predicted lung cancer severity based on significant risk factors and major symptoms. The risk-based model uses demographic and environmental information to predict future cancer development. This model achieved excellent results by delivering both a 98.99% accurate prediction and maintaining bal-

anced classes. Epidemiological studies have previously shown that passive smoking, along with smoking and alcohol use, represent key predictors of cancer development. The statistical test, which included Spearman correlation along with Kruskal-Wallis tests, revealed that smoking and occupational hazards, and air pollution significantly impact lung cancer risk factors. The statistical analysis through the Pearson Chi-square test established a significant connection between smoking habits and cancer development, which strengthens the previous statistical findings.

The Symptoms-based model determines which individuals possibly suffer from existing cancer. The model analyzed patient-provided symptoms, which included chest pain alongside coughing blood and shortness of breath and wheezing and fatigue and clubbing of finger nails, and difficulty swallowing. The test dataset produced a perfect result with 100% accuracy and a Kappa value of 1.0, and the model maintained perfect sensitivity and specificity for all three risk categories. The examination of feature importance determined that coughing blood, together with chest pain and wheezing, stood out as the main risk predictors for lung cancer. These findings demonstrate that present clinical symptoms can strongly indicate the risk level of lung cancer, which may aid in faster diagnosis and treatment decision-making. However, the exceptional model performance requires additional investigation to identify any potential data leakage risks.

A risk-based model supports healthcare providers in detecting lung cancer before symptoms begin, while the symptom-based model enables doctors to diagnose early-stage lung cancer. This approach allows medical professionals to observe symptom changes throughout time and adjust treatment plans when necessary.

The limitation of this study is the use of a small dataset and a potentially biased dataset sourced from Kaggle. The model's generalizability might be low because of a lack of clinical verification and the diversity of the patient dataset. Alcohol Use, Smoking, and Balanced Diet are common self-reported features in survey data. Self-reported data often encounters social desirability bias, which causes patients to provide inaccurate information.

8. Future Work

While the current study produced highly accurate predictive models using both risk and symptom-related features, several areas remain for future development and exploration:

Use of Clinical Data: Future studies should incorporate real-world clinical datasets from hospitals or cancer registries to improve generalizability and clinical relevance. Public datasets, while useful for prototyping, may not represent diverse patient populations or verified outcomes.

Temporal & Longitudinal Data: Integrating time-series data on patient health changes could support modeling disease progression and predict transitions from low to high severity risk levels.

Integration of Radiological and Genomic Data: Combining imaging, biopsy, or genetic mutation data with clinical and lifestyle variables may enhance model precision and offer a more comprehensive view of risk factors.

Deep Learning Approaches: Future iterations could explore neural networks and ensemble models to evaluate whether these methods outperform traditional machine learning in classifying lung cancer severity.

Deployable Tools: Building interactive web applications or mobile tools (e.g., using R Shiny or iOS/Android frameworks) would make the model more accessible to clinicians and the public, enabling real-time screening support.

9. References

1. World Health Organization. (2021). *Cancer*. Retrieved from https://www.who.int/news-room/fact-sheets/detail/cancer

2. GLOBOCAN 2020. (2021). *Lung Cancer Fact Sheet*. International Agency for Research on Cancer. Retrieved from https://gco.iarc.fr/today/data/factsheets/cancers/15-Lung-fact-sheet.pdf

3. Agonsanou, S., Houngbeme, F., & Degboe, A. (2020). *Machine learning techniques for lung cancer risk prediction: A comparative analysis*. *Journal of Biomedical Informatics*, 103, 103388. https://www.

explorationpub.com/Journals/em/Article/1001201

4. Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

5. Liaw, A., & Wiener, M. (2002). *Classification and Regression by randomForest*. *R News*, 2(3), 18–22. Retrieved from https://cran.r-project.org/doc/Rnews/Rnews_2002-3.pdf

6. Kaggle. (2020). *Lung Cancer Prediction Dataset*. Retrieved from https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link

7. US Environmental Protection Agency. (2024). "Particle Pollution and Respiratory Effects." www.epa.gov, 20 June 2024. Retrieved from [www.epa.gov/pmcourse/particle-pollution-and-respiratory-effects]

8. Eldridge, Lynne. (2023). "How Does Lung Cancer Kill People?" Verywell Health, 26 June 2023, [www.verywellhealth.com/how-do-people-die-from-lung-cancer-2249013].