Data engineering and analysis (IR4-S7)

Data analysis project report

Lung Cancer Detection

Submitted by
LEBAMA Jacques Jr. Frederic
RODRIGUEZ SENIS Ignacio
GENEV Yordan

Submitted to
LIONTI Fabien

December 14, 2024

Table of contents

Context and motivation

Lung cancer stands as the foremost cause of cancer-related mortality globally, with approximately 2.2 million new cases diagnosed each year. The disease's insidious progression often leads to diagnoses at advanced stages, where therapeutic options are limited and prognoses are poor. Early detection is paramount; studies indicate that about 60% of individuals diagnosed at the earliest stage survive their disease for five years or more, compared to less than 10% for those diagnosed at the most advanced stage. [1]

The primary risk factor for lung cancer is smoking, accounting for approximately 72% of cases in the UK. However, other significant risk factors include exposure to environmental pollutants such as asbestos, radon gas, and air pollution, as well as genetic predisposition. Notably, a growing number of lung cancer cases are being identified among non-smokers, particularly women aged 35 to 54, suggesting that factors beyond tobacco exposure contribute to the disease's incidence. [2]

Common symptoms of lung cancer include persistent coughing, chest pain, and fatigue. However, these symptoms often manifest in later stages, underscoring the necessity for effective screening methods to identify the disease earlier. Traditional screening techniques, such as low-dose computed tomography (LDCT), have been instrumental in detecting lung cancer at more treatable stages. Yet, the implementation of these screening programs faces challenges, including limited accessibility and low uptake among eligible populations.

Recent advancements in artificial intelligence and data analytics have opened new avenues for early detection and risk assessment of lung cancer. For instance, a 2023 study by researchers at MIT and Massachusetts General Hospital introduced "Sybil," an AI tool capable of predicting the risk of developing lung cancer within six years by analyzing low-dose chest computed tomography scans. Sybil demonstrated strong predictive performance across diverse datasets, highlighting AI's potential in early detection. [3]

Similarly, a 2024 study published in *Nature* detailed an AI system trained on millions of pathology images across 19 cancer types. [4] This model not only detected cancer presence with high accuracy but also predicted patient outcomes and treatment responses, showcasing AI's versatility in oncology diagnostics.

Building upon these advancements, this project aims to analyze real-world patient data, including lifestyle habits, medical history, and symptoms to develop a robust predictive model. The goal is to assist clinicians in early diagnosis and to formulate targeted treatment strategies, ultimately improving patient outcomes and reducing lung cancer mortality.

By leveraging data analytics and machine learning, this approach aspires to enhance early detection strategies, ultimately improving patient outcomes and reducing lung cancer mortality.

Exploratory data analysis

Descriptive statistics

The dataset contains 16 columns with a total of 309 entries, all of which have non-null values. The columns represent a mix of categorical and numerical data. Key features include demographic information like GENDER and AGE, lifestyle factors such as SMOKING and ALCOHOL CONSUMING, and health indicators such as WHEEZING, COUGHING, and CHEST PAIN. The target variable is LUNG_CANCER, indicating whether the individual has been diagnosed with lung cancer.

#	Column	Non-Null Count	Dtype
0	GENDER	309 non-null	object
1	AGE	309 non-null	int64
2	SMOKING	309 non-null	int64
3	YELLOW_FINGERS	309 non-null	int64
4	ANXIETY	309 non-null	int64
5	PEER_PRESSURE	309 non-null	int64
6	CHRONIC DISEASE	309 non-null	int64
7	FATIGUE	309 non-null	int64
8	ALLERGY	309 non-null	int64
9	WHEEZING	309 non-null	int64
10	ALCOHOL CONSUMING	309 non-null	int64
11	COUGHING	309 non-null	int64
12	SHORTNESS OF BREATH	309 non-null	int64
13	SWALLOWING DIFFICULTY	309 non-null	int64
14	CHEST PAIN	309 non-null	int64

Table 1. Dataset summary

Column	Unique Values
GENDER	[‘M’, ‘F’]
AGE	[69, 74, 59, 63, 75, 52, 51, 68, 53, 61, 72, 60, 58, 48, 57, 44, 64, 21, 65, 55, 62, 56, 67, 77, 70, 54, 49, 73, 47, 71, 66, 76, 78, 81, 79, 38, 39, 87, 46]
SMOKING	[1, 2]
YELLOW_FINGERS	[2, 1]
ANXIETY	[2, 1]
PEER_PRESSURE	[1, 2]
CHRONIC DISEASE	[1, 2]
FATIGUE	[2, 1]
ALLERGY	[1, 2]
WHEEZING	[2, 1]
ALCOHOL CONSUMING	[2, 1]
COUGHING	[2, 1]
SHORTNESS OF BREATH	[2, 1]
SWALLOWING DIFFICULTY	[2, 1]
CHEST PAIN	[2, 1]
LUNG_CANCER	[‘YES’, ‘NO’]

Table 2. Unique values in each column

Statistic	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE	FATIGUE	ALLERGY	WHEEZING	ALCOHOL CONSUMING	COUGHING	SHORTNESS OF BREATH	SWALLOWING DIFFICULTY	CHEST PAIN
Count	309.000	309.000	309.000	309.000	309.000	309.000	309.000	309.000	309.000	309.000	309.000	309.000	309.000	309.000
Mean	62.673	1.563	1.570	1.498	1.502	1.505	1.673	1.557	1.557	1.557	1.579	1.641	1.469	1.557
Std	8.210	0.497	0.496	0.501	0.501	0.501	0.470	0.498	0.498	0.498	0.494	0.481	0.500	0.498
Min	21.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
25%	57.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
50%	62.000	2.000	2.000	1.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000	1.000	2.000
75%	69.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000
Max	87.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000

Table 3. Descriptive statistics

The dataset contains a total of **309 observations**, with all variables fully populated. This means there are no missing values, simplifying subsequent analysis and model building.

The variable GENDER, representing the gender of individuals, is binary (encoded as 0 and 1). The mean value of **0.52** suggests that the dataset is approximately balanced between males (1) and females (0).

Examining the age distribution (AGE), we see that it ranges from **21 to 87 years**, with a mean age of approximately **62.67 years**. This indicates that the data focuses primarily on older individuals, which aligns with the context of lung cancer research. The standard deviation of **8.21 years** shows moderate variability in the age distribution.

The dataset also includes several binary features, such as SMOKING (smoking), WHEEZING (wheezing), and YELLOW_FINGERS (yellow fingers). These variables are encoded as 1 and 2, with mean values (e.g., **1.56 for smoking**) suggesting that most individuals in the sample fall into the "yes" category (2).

Regarding the target variable LUNG_CANCER, which indicates whether an individual has lung cancer, the mean value of **0.87** implies that **87% of the observations are cases of lung cancer**. This highlights a significant class imbalance in the dataset, which will need to be addressed during modeling through techniques such as oversampling or class weighting.

Overall, the dataset is clean and ready for further exploration, with key predictors such as SMOKING, WHEEZING, and AGE showing potential importance for lung cancer prediction.

Encoding categorical values

To observe relationships such as those in a correlation heatmap during EDA, it is necessary to encode categorical variables into numerical representations. Using the LabelEncoder method, ensures that categorical values are transformed into numerical format, allowing for meaningful statistical computations and visualizations.

We encoded the categorical variables GENDER and LUNG_CANCER into binaries.

For GENDER:

- Male is represented as 1
- Female is represented as 0

For LUNG_CANCER:

- Having cancer is represented as 1
- Not having cancer is represented as 0

Gender distribution

This pie chart represents the gender distribution within the dataset, showcasing a nearly balanced composition. Approximately 51.45% of the participants belong to one gender category, while the remaining 48.55% belong to the other. This balanced representation ensures that gender-based analyses within the dataset can yield meaningful insights without significant bias.

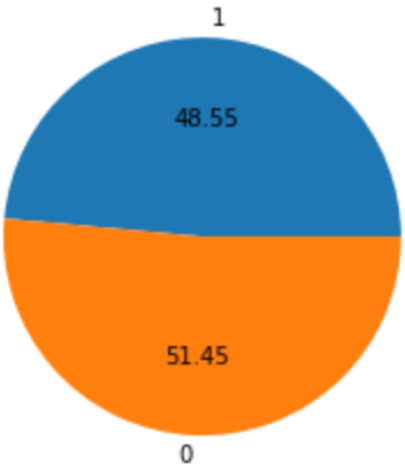

Figure 1. Gender distribution

Feature distributions

The feature distributions reveal that the dataset primarily consists of older individuals, with the majority falling between the ages of 40 and 80, consistent with the demographic most at risk for lung cancer. Several binary features, such as SMOKING, WHEEZING, and YELLOW_FINGERS, are skewed toward "yes" (encoded as 2), indicating that these behaviors or symptoms are prevalent within the dataset.

The target variable LUNG_CANCER appears highly imbalanced, with most cases being positive (1). This highlights the need to address the issue during the modeling phase to ensure fair and accurate predictions.

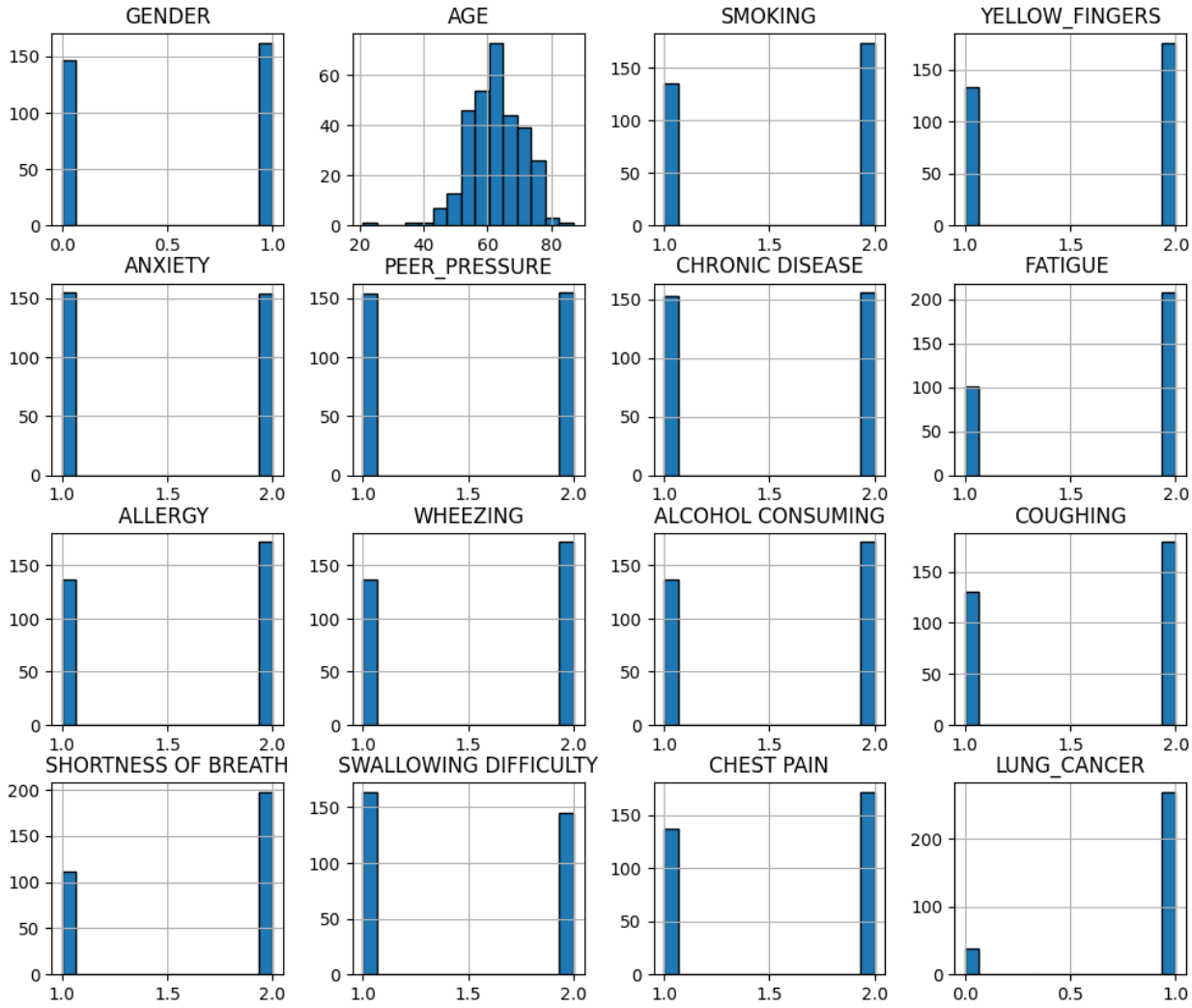

Table 4. Feature distributions

Correlations

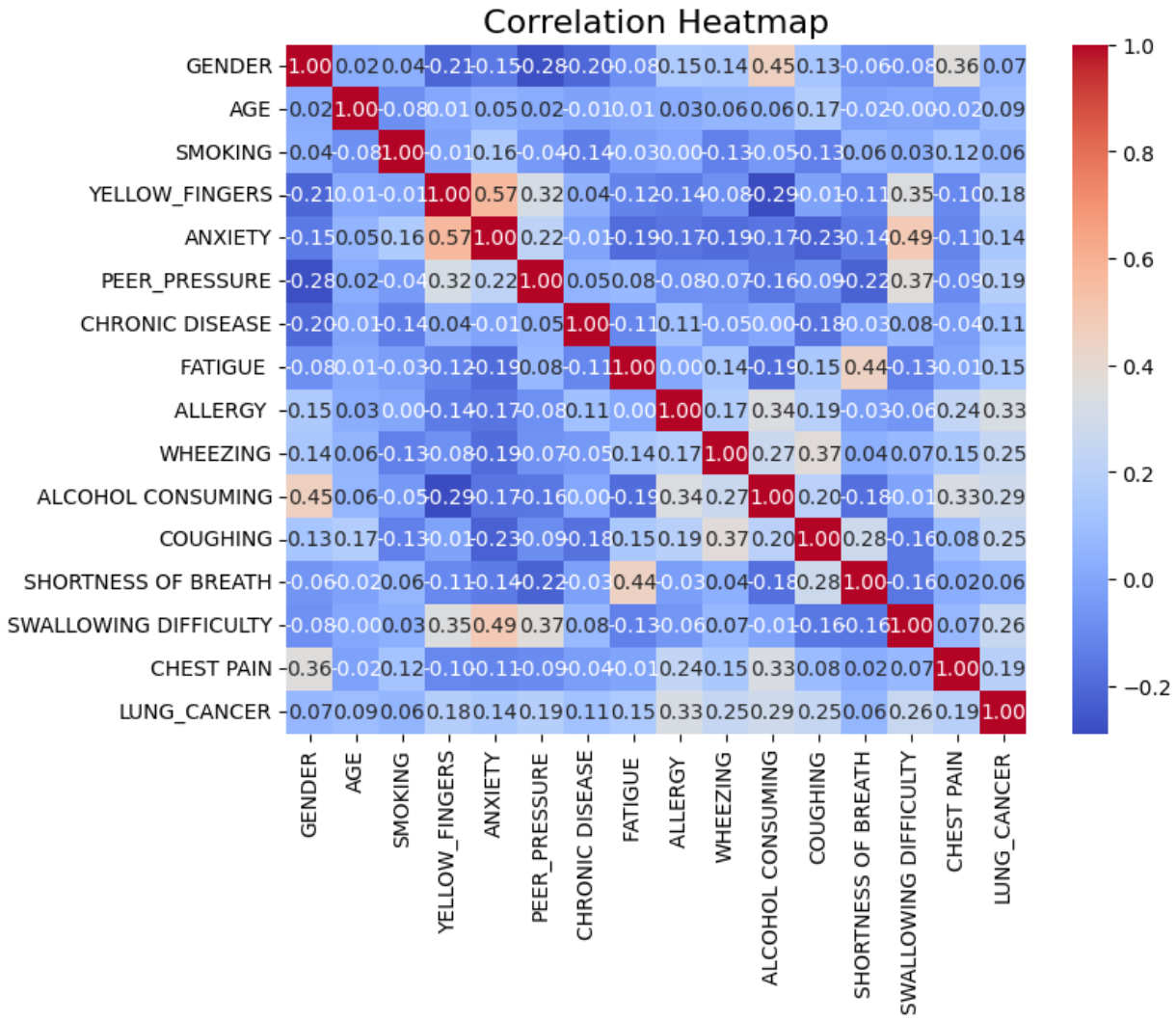

Figure 2. Correlation heatmap

Key features such as ALLERGY, WHEEZING, SHORTNESS OF BREATH, and SWALLOWING DIFFICULTY show positive correlations with LUNG_CANCER, suggesting they could play an important role in predicting its presence. Other certain features like SMOKING and YELLOW_FINGERS exhibit smaller but noticeable correlations, reinforcing their relevance as potential predictors. Most features show weak or negligible correlations with each other, which indicates minimal multicollinearity and supports the inclusion of multiple features in the model without redundancy.

Chi-Square test

The Chi-Square test is a statistical method used to determine whether there is a significant association between two categorical variables. [5] In this analysis, the relationship between smoking (SMOKING) and lung cancer diagnosis (LUNG_CANCER) was assessed. Smoking is a binary variable categorized as non-smokers (1) and smokers (2), while lung cancer is categorized as no lung cancer (0) and lung cancer (1).

Visualization of the data indicated that most smokers were diagnosed with lung cancer, whereas non-smokers had fewer lung cancer cases. However, statistical results from the test provided a different perspective. The Chi-Square statistic was calculated as 0.163, with a p-value of 0.686. Since the p-value is greater than 0.05, the result suggests that the relationship between smoking and lung cancer in this dataset is not statistically significant. This implies that smoking does not appear to have a significant association with lung cancer within this sample.

Despite smoking being a well-established risk factor for lung cancer, the lack of statistical significance in this analysis could be attributed to the limited size of the dataset or the presence of other confounding variables that may influence the relationship.

Feature-target relationships

The visualization below illustrates the feature-target relationships in the dataset, emphasizing how various attributes relate to the target variable, LUNG_CANCER. Each subplot corresponds to a specific feature, showing its distribution across the two categories of the target variable: individuals diagnosed with lung cancer (LUNG_CANCER = 1) and those not diagnosed (LUNG_CANCER = 0).

Detailed Observations:

1. GENDER: The distribution of lung cancer appears relatively consistent across genders, though there may be slight variations in frequency. This suggests that gender alone may not be a strong predictor of lung cancer.
2. AGE: The AGE feature shows a significant trend, with individuals diagnosed with lung cancer generally skewing toward the older age group. The histogram highlights that lung cancer cases are more prevalent in individuals above 60 years, reinforcing age as a potential risk factor.
3. SMOKING: A strong association is visible between SMOKING and lung cancer. Individuals who are categorized as heavy or frequent smokers (SMOKING = 2) have a noticeably higher count of lung cancer cases compared to non-smokers or less frequent smokers.
4. YELLOW_FINGERS: This feature, likely a proxy for smoking intensity or exposure, shows a clear trend where individuals with YELLOW_FINGERS = 2 are more likely to

have lung cancer. This aligns with the known correlation between smoking and lung cancer.

5. ANXIETY and PEER_PRESSURE: Both features show balanced distributions across the target categories. This suggests that while they may have some correlation with other factors, their direct relationship with lung cancer might be weaker.
6. CHRONIC DISEASE and FATIGUE: Both features display higher counts in the lung cancer group. This indicates that individuals with chronic conditions or persistent fatigue may be more prone to lung cancer, potentially due to weakened immunity or other associated health issues.
7. ALLERGY and WHEEZING: These respiratory-related features demonstrate higher associations with lung cancer cases. Specifically, wheezing (WHEEZING = 2) is more prevalent in individuals with lung cancer, which is consistent with the impact of lung-related disorders on respiratory symptoms.
8. ALCOHOL CONSUMING: Alcohol consumption shows some variability across the target groups. While not as significant as smoking, it could act as a co-factor in lung cancer risk when combined with other features.
9. COUGHING, SHORTNESS OF BREATH, SWALLOWING DIFFICULTY, and CHEST PAIN: These features show a strong relationship with lung cancer, with higher counts in the positive diagnosis group (LUNG_CANCER = 1). These are likely direct symptoms or indicators of lung cancer, making them critical predictors for the target variable.

The feature-target relationships highlight that certain features, such as AGE, SMOKING, YELLOW_FINGERS, WHEEZING, and health symptoms (e.g., COUGHING, SHORTNESS OF BREATH), have a strong association with lung cancer. These features are likely to play a significant role in predictive modeling. Conversely, features like GENDER, ANXIETY, and PEER_PRESSURE show less direct relationships with the target variable, suggesting they may have limited predictive value. This analysis provides valuable insights for feature selection, enabling the identification of the most impactful variables for further modeling and analysis.

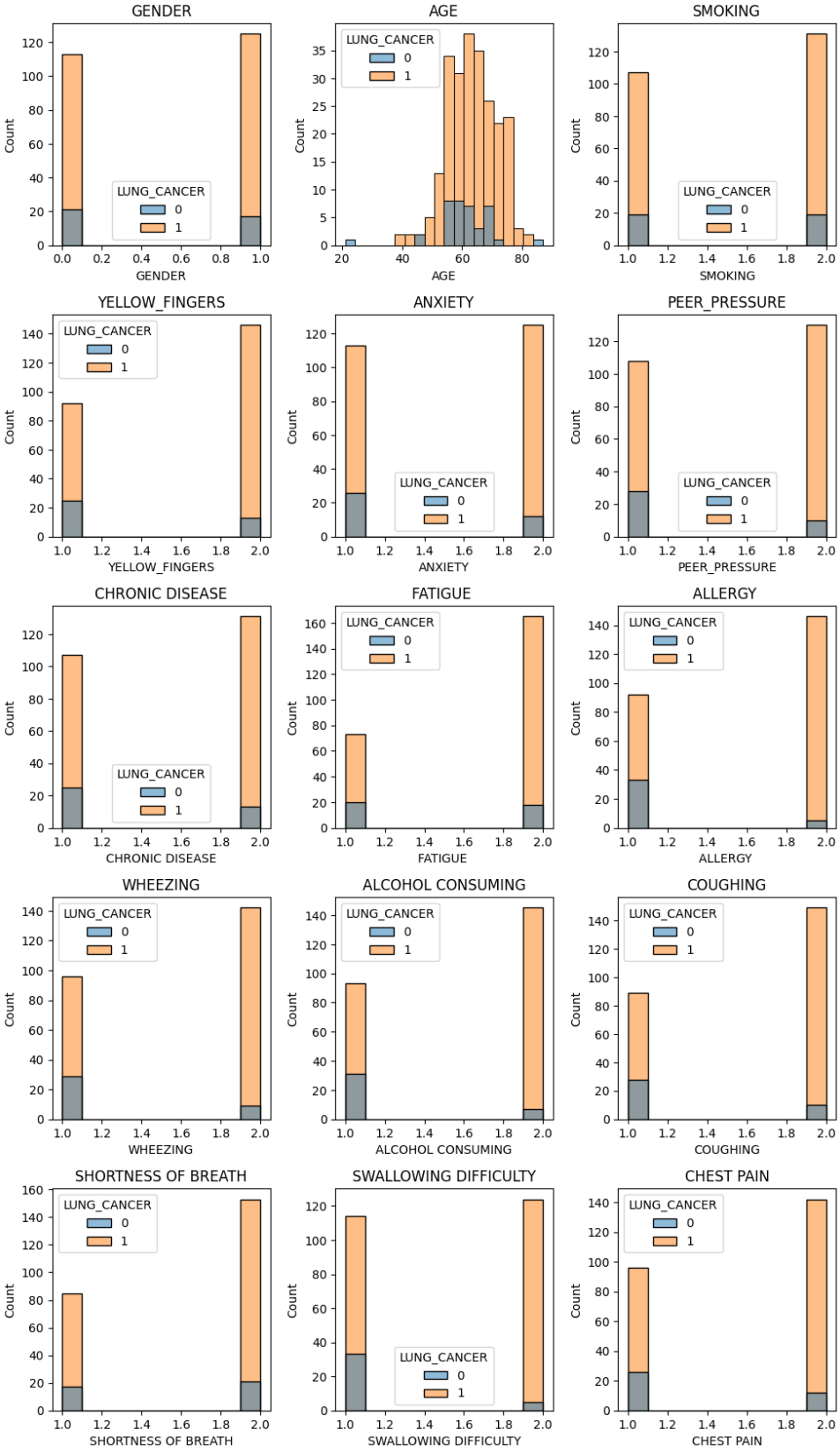

Figure 3. Feature-target relationships

Dataset preprocessing

Scalling

Standardizing numerical features is a crucial preprocessing step in machine learning, especially for algorithms sensitive to feature scales, such as Logistic regression. Standardization transforms features to have a mean of zero and a standard deviation of one, ensuring that each feature contributes equally during model training. This process enhances the efficiency and performance of algorithms that rely on gradient-based optimization methods. [6]

GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE	FATIGUE	ALLERGY	WHEEZING	ALCOHOL CONSUMING	COUGHING	SHORTNESS OF BREATH	SWALLOWING DIFFICULTY	CHEST PAIN
1	51	2	1	1	1	1	2	1	2	2	2	2	1	2
1	53	1	1	1	1	2	2	2	1	2	1	2	1	2
1	67	1	2	1	1	1	2	1	2	2	2	2	1	1
1	77	1	2	1	2	1	2	2	2	2	2	1	1	1
1	74	2	1	1	1	2	2	2	2	2	1	1	2	2

Table 5. Unscaled numeric features example

GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE	FATIGUE	ALLERGY	WHEEZING	ALCOHOL CONSUMING	COUGHING	SHORTNESS OF BREATH	SWALLOWING DIFFICULTY	CHEST PAIN
0.93339	-1.39515	0.87423	-1.08896	-0.95641	-0.98793	-0.96420	0.64143	-1.07137	0.91831	0.91831	0.84574	0.71787	-0.86705	0.91831
0.93339	-1.14705	-1.14386	-1.08896	-0.95641	-0.98793	1.03713	0.64143	0.93339	-1.08896	0.91831	-1.18240	0.71787	-0.86705	0.91831
0.93339	0.58960	-1.14386	0.91831	-0.95641	-0.98793	-0.96420	0.64143	-1.07137	0.91831	0.91831	0.84574	0.71787	-0.86705	-1.08896
0.93339	1.83006	-1.14386	0.91831	-0.95641	1.01222	-0.96420	0.64143	0.93339	0.91831	0.91831	0.84574	-1.39301	-0.86705	-1.08896
0.93339	1.45792	0.87423	-1.08896	-0.95641	-0.98793	1.03713	0.64143	0.93339	0.91831	0.91831	-1.18240	-1.39301	1.15334	0.91831

Table 6. Scaled numeric features example

Feature engineering

Feature engineering is the process of using domain knowledge to extract features from raw data that make machine learning algorithms work more effectively. [7] It involves selecting, modifying, and creating variables (features) that enhance the performance of machine learning models. By transforming raw data into a format that better represents the underlying problem, feature engineering helps bridge the gap between raw inputs and meaningful outputs in various applications. [8]

New feature – SMOKING_YELLOW_RISK

The newly engineered feature, SMOKING_YELLOW_RISK, combines the smoking habits of individuals (SMOKING) with the physical manifestation of heavy smoking (YELLOW_FINGERS). This feature is designed to better represent the severity of smoking-related behavior, as YELLOW_FINGERS often reflects long-term or intense smoking exposure. By multiplying these two variables, the feature captures both the behavioral and physical dimensions of smoking, potentially strengthening the dataset's predictive capabilities for lung cancer.

In the dataset, SMOKING_YELLOW_RISK takes on three distinct values: 1, 2, and 4. The distribution of these values shows that 139 instances correspond to higher combined smoking risks (value of 2), followed by 85 cases for the highest severity (value of 4), and 52 cases with the lowest risk (value of 1).

The plot reveals that individuals diagnosed with lung cancer tend to have higher SMOKING_YELLOW_RISK values compared to those without lung cancer. Specifically, the median SMOKING_YELLOW_RISK for lung cancer cases is higher, and there is a wider range of values, including an outlier at the highest end of the scale. In contrast, non-cancer cases show a more concentrated distribution of lower SMOKING_YELLOW_RISK values.

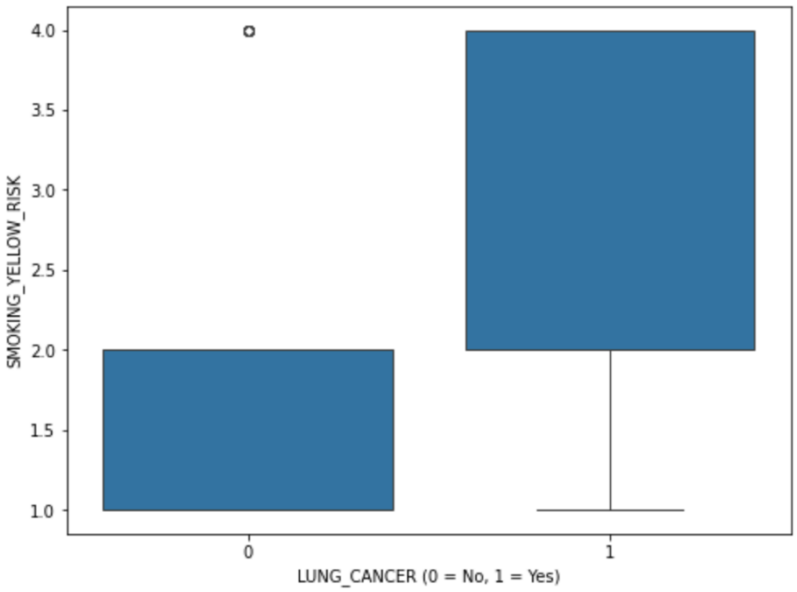

Figure 4. Relationship Between SMOKING_YELLOW_RISK and LUNG_CANCER

This suggests that higher values of SMOKING_YELLOW_RISK, which capture the combined effect of smoking intensity and its physical consequences (yellow-stained fingers), are associated with a greater likelihood of lung cancer. The box plot underscores the potential utility of this feature in predictive modeling.

New feature – BREATHING_ISSUES

The newly engineered feature – BREATHING_ISSUES, combines the variables SHORTNESS_OF_BREATH and WHEEZING, both of which are symptoms often associated with lung cancer and other respiratory diseases. [9] Individually, these variables provide information about respiratory distress, but when combined, they offer a comprehensive representation of the overall severity of breathing difficulties.

The feature is created by summing the values of SHORTNESS_OF_BREATH and WHEEZING, resulting in three distinct levels of severity: 2, 3, and 4. The distribution shows that the majority of individuals (129) experience a moderate level of breathing issues (value of 3), followed by 98 cases with more severe issues (value of 4), and 49 cases with lower levels of difficulty (value of 2).

BREATHING_ISSUES is expected to correlate strongly with lung cancer, as higher values reflect cumulative respiratory problems, which are often linked to lung-related health conditions. By capturing the combined impact of multiple respiratory symptoms, this variable enhances the dataset's ability to model lung cancer risk effectively.

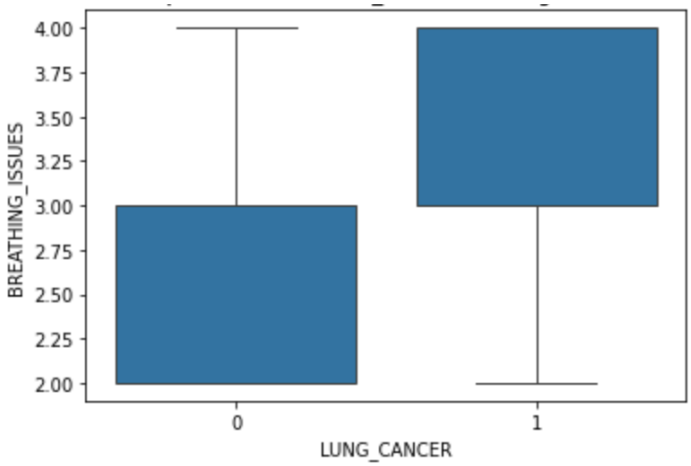

Figure 5. Relationship Between BREATHING_ISSUES and LUNG_CANCER

The boxplot shows that individuals diagnosed with lung cancer (LUNG_CANCER = 1) tend to have higher median BREATHING_ISSUES values compared to those without lung cancer (LUNG_CANCER = 0). The interquartile range for lung cancer cases is also wider, indicating more variability in the severity of breathing issues among diagnosed individuals. Conversely, non-cancer cases exhibit a narrower distribution with lower overall values of BREATHING_ISSUES.

Machine learning model selection

Random Forest

Random Forest is a machine learning algorithm that uses ensemble learning to build multiple decision trees and aggregates their outputs for prediction. It is particularly robust, capable of handling non-linear relationships, and resistant to overfitting, making it an ideal choice for datasets with complex interactions. It provides insights into feature importance, which enhances interpretability.

For this analysis, Random Forest was selected due to its ability to effectively manage mixed data types and interactions between features such as SMOKING_YELLOW_RISK and BREATHING_ISSUES. The model demonstrated strong predictive performance on the test data, achieving an accuracy of 91%. Notably, its precision for detecting lung cancer cases (Class 1) was 94%, indicating a low rate of false positives, and its recall was 96%, reflecting a high sensitivity in correctly identifying lung cancer cases. The F1-score for lung cancer detection was 95%, showing a strong balance between precision and recall.

Class	Precision	Recall	F1-Score	Support
Class 0	0.71	0.62	0.67	8
Class 1	0.94	0.96	0.95	48
Accuracy			0.91	56
Macro Avg	0.83	0.79	0.81	56
Weighted Avg	0.91	0.91	0.91	56

Table 7. Classification report for the Random Forest algorithm

The confusion matrix highlights some limitations in detecting non-lung cancer cases (Class 0), with a recall of 62%. This suggests occasional misclassification of healthy individuals as having lung cancer. Despite this, the high recall for lung cancer (Class 1) underscores its suitability for medical diagnostics where minimizing false negatives is critical.

Predicted \ Actual	Class 0	Class 1
Class 0	5	3
Class 1	2	46

Table 8. Confusion matrix for the Random Forest algorithm

The findings suggest that Random Forest is a strong candidate for lung cancer prediction. Future improvements could involve hyperparameter tuning to address false positives and enhance recall for non-lung cancer cases.

Decision Tree

The Decision Tree algorithm was chosen for its simplicity and interpretability, making it particularly suited for medical applications where clear and transparent decision-making is essential. Its structure offers valuable insights into how various features influence predictions, allowing practitioners to trace decisions through explicit rules. Additionally, its ability to handle both categorical and numerical data without extensive preprocessing made it an ideal choice for this dataset. [10]

The model performed well, achieving an accuracy of 89% on the test set. For lung cancer detection, it demonstrated strong metrics, with a precision, recall, and F1-score all at 94%. These results highlight its effectiveness in identifying lung cancer patients while maintaining a balanced performance between precision and recall. However, the confusion matrix revealed some limitations in distinguishing non-cancer cases, as 3 healthy individuals were misclassified as having lung cancer, and 3 lung cancer cases were missed.

The Decision Tree's structure identified SWALLOWING_DIFFICULTY, YELLOW_FINGERS, and ALCOHOL_CONSUMING as key predictors. Patients with high values for these features were often classified as having lung cancer, underscoring their importance in risk assessment. This interpretability, combined with its solid predictive performance, makes the Decision Tree a valuable tool in medical diagnostics. Nevertheless, its susceptibility to false positives emphasizes the need for complementary methods to minimize unnecessary interventions.

Class	Precision	Recall	F1-Score	Support
Class 0	0.62	0.62	0.62	8
Class 1	0.94	0.94	0.94	48
Accuracy			0.89	56
Macro Avg	0.78	0.78	0.78	56
Weighted Avg	0.89	0.89	0.89	56

Table 9. Classification report for the Decision Tree algorithm

Predicted \ Actual	Class 0	Class 1
Class 0	5	3
Class 1	3	45

Table 10. Confusion matrix for the Decision Tree algorithm

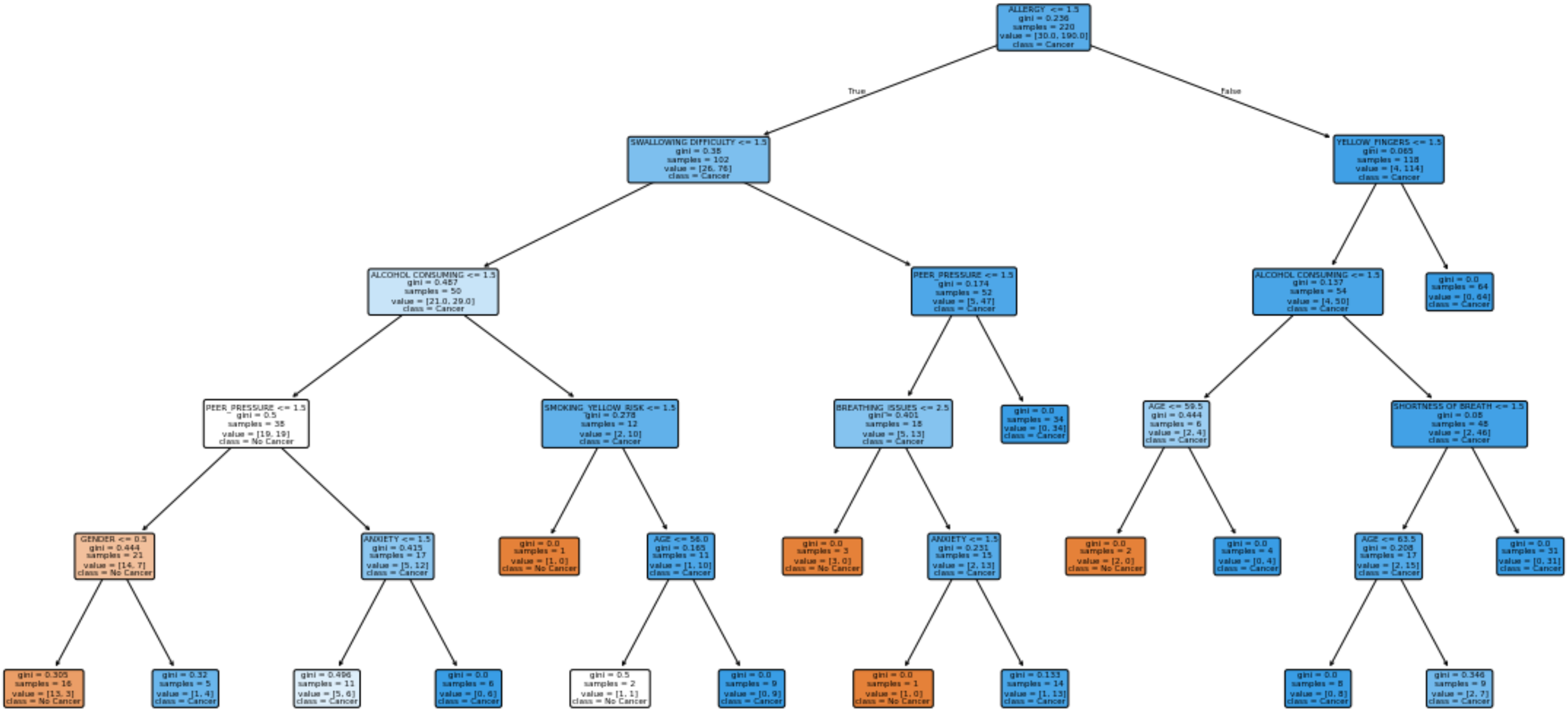

Figure 6. Decision Tree visualization

Support Vector Machine

The Support Vector Machine model was applied to evaluate its effectiveness in predicting lung cancer cases, leveraging its capability to handle complex relationships between features. The model is particularly suitable for datasets with non-linear interactions, as demonstrated by its use of the radial basis function kernel. The built-in regularization parameters of SVM enhance its robustness by controlling model complexity and reducing the risk of overfitting.

Class	Precision	Recall	F1-Score	Support
Class 0	0.50	0.25	0.33	8
Class 1	0.88	0.96	0.92	48
Accuracy			0.86	56
Macro Avg	0.69	0.60	0.63	56
Weighted Avg	0.83	0.86	0.84	56

Table 11. Classification report for the SVM model

Predicted \ Actual	Class 0	Class 1
Class 0	2	6
Class 1	2	46

Table 12. Confusion matrix for the SVM model

The SVM model achieved strong performance, with an accuracy of 85.7% on the test set. For lung cancer detection (Class 1), the model demonstrated high precision (88.5%) and recall (95.8%), resulting in an impressive F1-score of 92%. These metrics indicate its effectiveness in identifying lung cancer cases. The confusion matrix reveals that the model correctly identified 46 lung cancer patients (True Positives) while misclassifying 2 as non-cancer (False Negatives). However, the model struggled with distinguishing non-cancer cases (Class 0), correctly classifying only 2 healthy individuals (True Negatives) while misclassifying 6 as lung cancer (False Positives). This results in a recall of only 25% for non-cancer cases.

The findings emphasize the SVM model's strength in minimizing false negatives, making it particularly valuable in medical diagnostics where missing a cancer diagnosis can have severe consequences. However, the relatively low performance for non-cancer cases highlights a need for improvement, as the false positives could lead to unnecessary follow-up tests and anxiety for patients. Despite this limitation, the high sensitivity for lung cancer cases suggests that the SVM model could be a reliable tool for early diagnosis, especially when combined with additional techniques to address its weaknesses in identifying healthy individuals. Further optimization, such as adjusting hyperparameters or exploring complementary models, could enhance its overall performance.

Conclusion

This report comprehensively explored the application of machine learning models to predict lung cancer based on patient attributes. The study began with a contextual overview highlighting the importance of early lung cancer detection and the motivation to leverage data-driven approaches for this task. EDA revealed critical patterns and relationships within the dataset, providing a solid foundation for model selection and evaluation.

Descriptive statistics and feature distributions emphasized the dataset's balance and variability. Encoding categorical variables ensured compatibility with analytical methods, while visualizations like gender distribution, feature distributions, and correlations provided a deeper understanding of data patterns. Statistical tests, including the Chi-Square test, assessed the association between key features and the target variable, laying the groundwork for feature engineering.

The creation of SMOKING_YELLOW_RISK and BREATHING_ISSUES as new features demonstrated the value of combining related variables to capture nuanced relationships. These features enriched the dataset, contributing significantly to the performance of machine learning models.

The model selection process involved Random Forest, Decision Tree, and Support Vector Machine. Random Forest emerged as the most effective model, achieving the highest accuracy (91%) and demonstrating balanced performance across precision, recall, and F1-score. Decision Tree, with an accuracy of 89%, offered unmatched interpretability, making it valuable for clinical applications requiring transparency. The SVM model, while achieving slightly lower accuracy (85.7%), excelled in recall for lung cancer cases, minimizing false negatives.

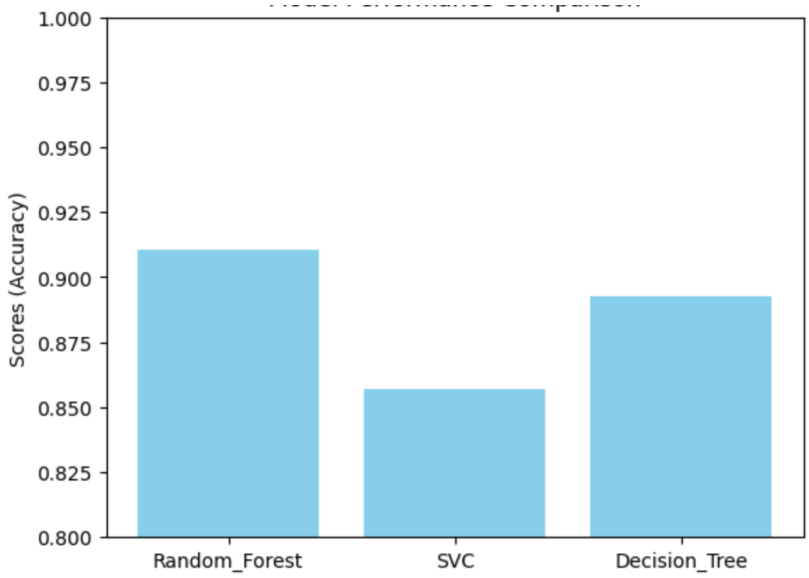

Figure 7. Model performance comparison

These findings highlight the trade-offs between model accuracy, interpretability, and sensitivity. Random Forest is ideal for maximizing predictive performance, Decision Tree is well-suited for settings where transparency is crucial, and SVM offers strong sensitivity for minimizing missed diagnoses.

Future efforts could focus on addressing class imbalances, refining feature engineering, and combining models into ensembles to leverage their complementary strengths. These approaches promise to enhance diagnostic accuracy, improve patient outcomes, and support healthcare professionals in early detection and intervention.

Bibliography

[1] "Staging and Treatment of Early-Stage Lung Cancer," Verywell Health, [Online]. Available: https://www.verywellhealth.com/what-is-early-stage-lung-cancer-2249025.

[2] "Causes and risk factors of lung cancer," Macmillan Cancer Support, [Online]. Available: https://www.macmillan.org.uk/cancer-information-and-support/lung-cancer/causes-and-risk-factors-of-lung-cancer.

[3] "MIT researchers develop an AI model that can detect future lung cancer risk," Massachusetts Institute of Technology, [Online]. Available: https://news.mit.edu/2023/ai-model-can-detect-future-lung-cancer-0120.

[4] "A pathology foundation model for cancer diagnosis and prognosis prediction," Nature, [Online]. Available: https://www.nature.com/articles/s41586-024-07894-z.

[5] "8. The Chi squared tests," BMJ, [Online]. Available: https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/8-chi-squared-tests.

[6] "Importance of Feature Scaling," scikit-learn, [Online]. Available: https://scikit-learn.org/stable/auto_examples/preprocessing/plot_scaling_importance.html.

[7] "Feature engineering," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Feature_engineering.

[8] "Machine Learning Engineering Unit 2 – Data Prep & Feature Engineering for ML," Fiveable, [Online]. Available: https://library.fiveable.me/machine-learning-engineering/unit-2.

[9] "Coping with breathlessness when you have lung cancer," Cancer Research, [Online]. Available: https://www.cancerresearchuk.org/about-cancer/lung-cancer/living-with/coping-with-breathlessness.

[10] "A Survey of Decision Trees: Concepts, Algorithms, and Applications," IEEE Xplore, [Online]. Available: https://ieeexplore.ieee.org/document/10562290.