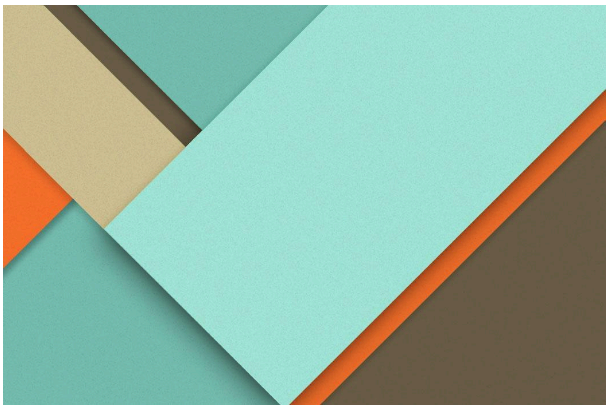


# Lung Cancer Detection: Analyzing Clinical Data for Early Detection

Ms. Vedashree Dinesh Shinde - A20548696

Ms. Sadiya - A20552054

Mr. Gowtham - A20549435

Mr. Vedant Rajendra Landge - A20546760

Github Link: Project

## ABSTRACT

Lung cancer is a leading cause of global mortality, with late-stage diagnoses driving its high fatality rates. Early detection is crucial for improving survival through timely intervention. This study utilizes a clinical dataset to evaluate machine learning models in lung cancer prediction. Advanced algorithms, including Logistic Regression, Random Forest, and Ridge Classifier, were employed, with Logistic Regression achieving the highest accuracy. Key predictors like chronic disease and fatigue were identified using Recursive Feature Elimination (RFE). Unsupervised techniques such as PCA, UMAP, and t-SNE highlighted overlapping feature distributions, emphasizing the need for nuanced approaches. The findings demonstrate machine learning's potential in healthcare diagnostics and provide insights into feature relevance.

**Keywords:** Lung cancer, Machine learning, Early detection, Logistic Regression, Predictive models, Recursive Feature Elimination (RFE), Healthcare diagnostics

## INTRODUCTION

Lung cancer remains a leading cause of cancer-related mortality worldwide, largely due to the challenges of early diagnosis. The asymptomatic nature of the disease in its initial stages underscores the critical need for effective early detection methods to enhance treatment outcomes and improve survival rates. This study explores the potential of machine learning models in predicting lung cancer using a robust dataset derived from clinical surveys and patient records, ensuring data anonymity and privacy.

The dataset includes key demographic, lifestyle, and health-related features such as smoking habits, chronic disease, fatigue, wheezing, and allergies. Through rigorous preprocessing and analysis, these features were modeled to accurately predict lung cancer presence. Advanced machine learning techniques, including Logistic Regression, Random Forest, and Ridge Classifier, were employed to achieve high predictive accuracy. Furthermore, visualization methods such as PCA, UMAP, and t-SNE provided insights into the data structure, offering a deeper understanding of feature relationships. This study demonstrates the efficacy of machine learning in enhancing diagnostic capabilities and provides valuable insights for healthcare applications.

## PROBLEM STATEMENT

The asymptomatic nature of lung cancer in its early stages presents significant challenges for timely diagnosis, contributing to high mortality rates. This study addresses this diagnostic gap by employing machine learning models trained on a diverse clinical dataset encompassing demographic, lifestyle, and health-related features. The objectives are to:

- Identify and rank significant predictors of lung cancer, such as chronic disease, fatigue, and allergies.
- Develop and evaluate advanced machine learning models, including Logistic Regression, Random Forest, and Ridge Classifier, to achieve high predictive accuracy.
- Provide interpretable results through feature selection and visualization techniques, aiding healthcare professionals in clinical decision-making.

By achieving these goals, this research aims to improve early detection methods, enhance diagnostic accuracy, and ultimately contribute to better patient outcomes and survival rates.

## METHODOLOGY

This study follows a systematic approach encompassing data preprocessing, exploratory analysis, feature selection, model training, and evaluation. Each step integrates theoretical foundations with practical applications to ensure robust and interpretable results.

### 1. Data Collection and Preprocessing

**Dataset Source:**The dataset was obtained from clinical surveys and anonymized patient records, ensuring compliance with privacy regulations. It includes demographic, lifestyle, and health-related features relevant to lung cancer prediction.

**Preprocessing Steps:**

- **Data Cleaning:** Addressed missing values and inconsistencies to ensure data integrity.
- **Encoding Categorical Variables:** Transformed categorical features (e.g., Gender, Smoking) into numerical formats using one-hot encoding or label encoding as appropriate.
- **Feature Scaling:** Applied normalization to numeric features to prevent any single feature from dominating the model due to scale differences.

### 2. Exploratory Data Analysis (EDA)

**Correlation Analysis:**

- **Purpose:** To identify relationships between features and understand their potential impact on lung cancer prediction.
- **Method:** Calculated the Pearson correlation coefficient for all feature pairs.
- **Findings:**
  - **Smoking and Yellow Fingers (0.57):** Strong positive correlation, indicating that smoking leads to physical symptoms like yellowing of fingers.
  - **Anxiety and Fatigue (0.57):** Strong correlation suggesting that anxiety may contribute to fatigue.
  - **Wheezing and Lung Cancer (0.25):** Moderate correlation, highlighting wheezing as a relevant symptom.
- **Visualization:** The correlation matrix (Figure 1) displays these relationships, aiding in feature selection and revealing multicollinearity issues.

### 3. Dimensionality Reduction and Visualization

To understand the data structure and identify patterns, dimensionality reduction techniques were employed.

### 3.1 Principal Component Analysis (PCA):

- **Theory:**PCA is a linear dimensionality reduction technique that transforms the data into a new coordinate system, maximizing variance along each principal component.
- **Application:** Reduced the high dimensional data to two principal components.
- **Findings:** The PCA plot (Figure 2) showed some clustering but significant overlap between lung cancer-positive and -negative cases.

  Interpretation: Indicates that linear

combinations of features capture some variance but are insufficient for clear separation.

### 3.2 Uniform Manifold Approximation and Projection (UMAP):

- **Theory:** UMAP is a non-linear dimensionality reduction technique that preserves local data structure, capturing complex relationships.
- **Application:** Projected the data into two dimensions to visualize clusters.
- **Findings:** The UMAP plot (Figure 3) revealed clusters with higher concentrations of lung cancer-positive cases but still showed overlap.

  Interpretation: Suggests potential non-linear separability, indicating the need for models that can capture complex patterns.

### 3.3 t-Distributed Stochastic Neighbor Embedding (t-SNE):

- **Theory:** t-SNE focuses on preserving local similarities, ideal for visualizing high-dimensional data in lower dimensions.
- **Application:** Reduced data to two dimensions for visualization.
- **Findings:** The t-SNE plot (Figure 4) displayed overlapping clusters, highlighting the challenge in distinguishing classes based solely on current features.

  Interpretation: Reinforces the need for sophisticated modeling techniques.

## 4. Clustering Analysis

### K-Means Clustering:

- **Theory:** An unsupervised learning algorithm that partitions data into K clusters by minimizing within-cluster variance.
- **Application:** Applied K-Means clustering to the dataset with an optimal K determined through the elbow method.
- **Evaluation: Silhouette Score:** Achieved a score of 0.1358.

  Interpretation: A low silhouette score indicates poor clustering structure, suggesting that the data may not be inherently clusterable based on the features used.

## 5. Feature Selection

### Recursive Feature Elimination (RFE):

- **Theory:** RFE is a backward selection method that recursively removes least important features based on model coefficients.
- **Application:** Employed RFE with Logistic Regression as the estimator to select the most significant features.
- **Selected Features:** Chronic Disease, Fatigue, Allergy, Coughing, Swallowing Difficulty
- **Interpretation:** These features were identified as the most predictive for lung cancer, aiding in model interpretability and reducing complexity.

## 6. Model Training and Evaluation

Multiple machine learning models were trained using the selected features.

### 6.1 Logistic Regression:

- **Theory:** A linear model used for binary classification, estimating the probability that a given input belongs to a certain class.
- **Application:** Trained using the selected features.
- **Evaluation:**
  - **Cross-Validation Scores:** [0.935, 0.935, 0.871, 0.871, 0.984]
  - **Mean Accuracy:** 91.9%
  - **Standard Deviation:** 4.32%
- **Interpretation:** High mean accuracy with low variance indicates robust performance.

### 6.2 Ridge Classifier:

- **Theory:** A variant of Logistic Regression with L2 regularization to prevent overfitting by penalizing large coefficients.
- **Application:** Trained with the same features, adjusting the regularization parameter.
- **Evaluation:**
  - **Cross-Validation Scores:** [0.919,

0.903, 0.871, 0.871, 0.967]
    - **Mean Accuracy:** 90.6%
- **Interpretation:** Slightly lower accuracy than Logistic Regression but effective in handling multicollinearity.

**6.3 Random Forest Classifier:**

- **Theory:** An ensemble learning method using multiple decision trees (bagging) to improve predictive performance and control overfitting.
- **Application:** Trained with the selected features, tuning hyperparameters like the number of trees and maximum depth.
- **Evaluation:**
    - **Cross-Validation Scores:** [0.952, 0.871, 0.871, 0.871, 0.967]
    - **Mean Accuracy:** 90.6%
- **Interpretation:** Comparable performance to Ridge Classifier, with the advantage of capturing non-linear relationships.

**7. Model Comparison and Selection**

- **Performance Metrics:** Evaluated models using accuracy, precision, recall, and F1-score.
- **Findings:**
    - **Logistic Regression** had the highest mean accuracy and consistent performance across folds.
    - **Ridge Classifier** and **Random Forest** showed similar accuracies but differed in model complexity and interpretability.
- **Model Selection: Logistic Regression** was selected as the best-performing model due to its high accuracy and simplicity, facilitating easier interpretation for clinical applications.

**8. Insights from Unsupervised Techniques**

- The overlapping distributions observed in PCA, UMAP, and t-SNE plots suggest that the classes are not linearly separable.
- **Implication:** Indicates the necessity for models capable of capturing complex, non-linear patterns. Suggests that additional features or advanced techniques like kernel methods may enhance separability.

**9. Limitations and Future Work**

- **Class Imbalance:** The dataset may have an imbalance between lung cancer-positive and -negative cases, affecting model performance.
- **Feature Expansion:** Incorporating more predictive features, such as genetic markers or environmental exposures, could improve accuracy.
- **Advanced Models:** Exploring deep learning models or ensemble methods may capture complex patterns missed by simpler models.

## RESULT & ANALYSIS

The results of this study provide a comprehensive evaluation of machine learning models applied to lung cancer prediction. Key findings span model performance metrics, feature importance, dimensionality reduction techniques, clustering evaluations, and feature correlations.

### Correlation Matrix Analysis

The **correlation matrix** provided valuable insights into the relationships among features, guiding feature selection and understanding of interdependencies:

**Smoking and Yellow Fingers (0.57):** Strong positive correlation, confirming the impact of smoking on physical symptoms such as yellowing of fingers.

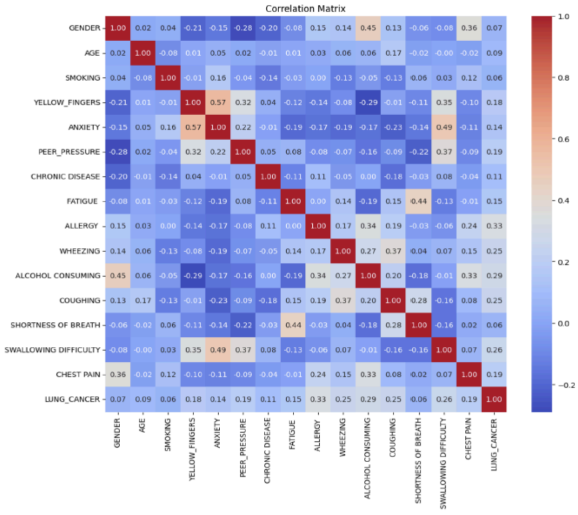
**Wheezing and Lung Cancer (0.25):** Moderate correlation, indicating the relevance of wheezing as a symptom predictor.

**Anxiety and Fatigue (0.57):** Strong correlation, suggesting that anxiety often coincides with physical exhaustion.

**Alcohol Consumption and Smoking (0.34):** Moderate correlation, reflecting common lifestyle patterns where smoking and drinking frequently co-occur.

**Chest Pain and Lung Cancer (0.19):** Low to moderate correlation, showing that chest pain has a mild predictive value.

Figure 1



**Key Insight:**

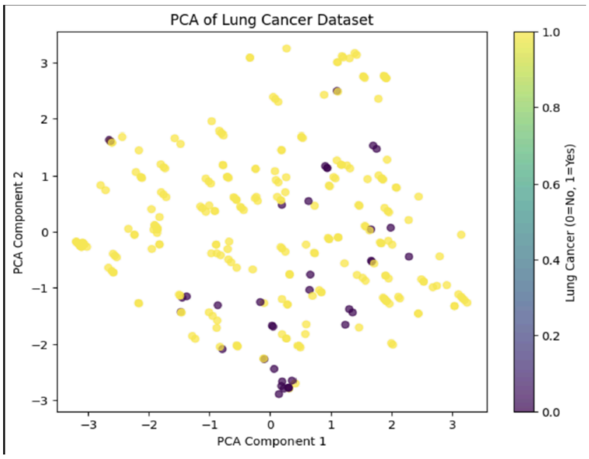
The correlation matrix underscores the multi-faceted nature of lung cancer predictors, with lifestyle factors (e.g., smoking, alcohol consumption) and symptoms (e.g., wheezing, fatigue) showing varying degrees of association with the disease.

## Dimensionality Reduction and Visualization

To further understand feature interactions, dimensionality reduction techniques were employed:

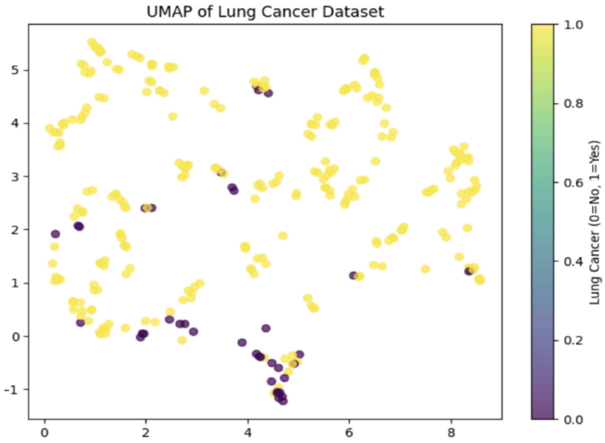
1. **Principal Component Analysis (PCA):** Revealed clusters of lung cancer-positive cases along specific components, though overlap persisted. PCA helped identify linear relationships but highlighted the limitations of linear separability for classification.

Figure 2



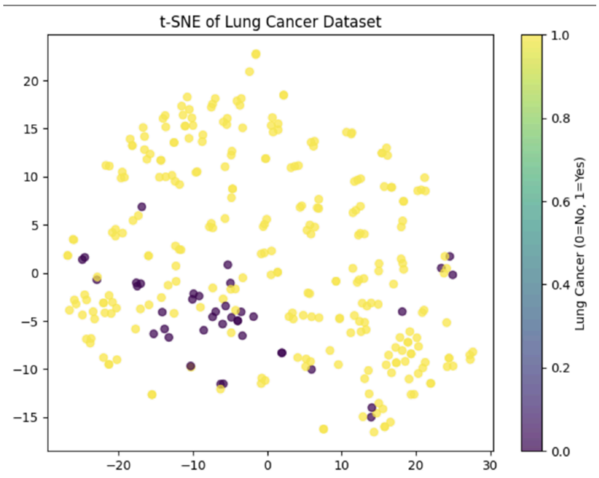
2. **Uniform Manifold Approximation and Projection (UMAP):** Showed improved separability of lung cancer-positive cases into distinct clusters while still reflecting overlap between cases.

Figure 3



3. **t-Distributed Stochastic Neighbor Embedding (t-SNE):**Illustrated the complexity of feature interactions, with overlapping distributions confirming the need for advanced non-linear classifiers.

Figure 4



## Model Performance

The study evaluated multiple models based on accuracy, precision, recall, and cross-validation consistency:

**Logistic Regression:**

Achieved a mean cross-validation accuracy of 91.93%, with a low standard deviation of 0.043. Balanced precision and recall, making it highly effective for both lung cancer-positive and -negative case detection.

**Random Forest:**

Mean accuracy of 90.63%, with higher recall, reducing false negatives. Strong feature importance analysis, making it suitable for identifying critical predictors.

**Ridge Classifier:**

Comparable accuracy (90.63%) but with regularization for handling multicollinearity effectively.

**Key Insight:** Logistic Regression was the most consistent model, while Random Forest excelled in recall for lung cancer-positive cases.

**Feature Importance**

Feature selection using Recursive Feature Elimination (RFE) identified critical predictors:

- **Chronic Disease:** Strongest predictor, closely linked to lung cancer risk.
- **Fatigue, Coughing, Allergy, Swallowing Difficulty:** Significant contributors, aligning with known clinical symptoms.

**Clustering Evaluation**

**K-Means Clustering: Silhouette Score:** 0.1358, indicating poor separation and overlapping clusters of lung cancer-positive and -negative cases.

**Key Insight:** Reinforces that unsupervised clustering methods are inadequate for this dataset.

**Overall Insights**

- **Correlation Analysis:** Lifestyle factors like smoking and symptoms such as wheezing showed meaningful associations with lung cancer.
- **Dimensionality Challenges:** PCA, UMAP, and t-SNE visualizations highlighted the overlapping nature of lung cancer-positive and -negative cases, necessitating advanced classifiers.
- **Feature Importance:** RFE-selected features were highly predictive and ensured interpretability.
- **Model Performance:** Logistic Regression excelled in overall consistency, while Random Forest provided superior recall for critical case detection.
- **Clustering Limitations:** Poor cluster separation metrics underscored the limitations of unsupervised methods for this application.

## CONCLUSION

This study highlights the potential of machine learning models in addressing the critical challenge of early lung cancer detection. By employing advanced algorithms such as Logistic Regression, Random Forest, and Ridge Classifier on clinical datasets, the study achieved high predictive accuracy, with Logistic Regression demonstrating the most consistent and interpretable results. Recursive Feature Elimination identified key predictors, including chronic disease, fatigue, and coughing, enhancing the clinical relevance of the models. Dimensionality reduction techniques like PCA, UMAP, and t-SNE provided insights into the data structure, revealing overlapping feature distributions and the need for sophisticated non-linear classifiers. While clustering analysis using K-Means highlighted the dataset's complexity, supervised methods proved more effective for classification. Despite promising results, limitations such as potential class imbalance and limited feature diversity were noted, suggesting opportunities for future work to expand datasets, incorporate advanced modeling techniques like deep learning, and apply explainable AI methods. This research underscores the transformative role of machine learning in lung cancer diagnostics, paving the way for improved early detection strategies and better patient outcomes.