International Journal of Innovative Research in Engineering
Volume 7, Issue 2 (March - April 2026), PP: 341-371.
https://www.doi.org/10.59256/ijire.20260702042
www.theijire.com
ISSN No: 2582-8746

Lung Cancer Detection System

Aditya Raj[1], Dr Hazique Aetesam[2]
[1,2] *Department of Computer Science & Engineering, Birla institute of technology mesra, Jharkhand, India.*

How to cite this paper:
Aditya Raj[1], Dr Hazique Aetesam[2], "Lung Cancer Detection System", IJIRE-V7I2-341-371.

Abstract: *Since lung cancer is one of the leading causes of death worldwide, early detection and risk assessment are essential for improving survival rates. This paper presents a machine learning-based approach for identifying lung cancer risk using classification techniques and an interactive web-based prediction system. A survey dataset containing patient demographic details, lifestyle habits, and clinical symptoms was used for analysis.The data preprocessing stage included handling missing values, encoding categorical variables, feature scaling, and preparing the dataset for model training. Multiple machine learning models were developed and evaluated, including Random Forest, Logistic Regression, and Support Vector Machine (SVM). Among these, the Random Forest classifier achieved the best performance with an accuracy of **94.2%**.The models were assessed using confusion matrix, ROC analysis, precision, recall, and F1-score. Additionally, a prediction system was developed that allows healthcare professionals to input patient data and receive real-time lung cancer risk predictions.The proposed system demonstrates the effectiveness of machine learning in healthcare analytics and decision support, offering a practical solution for early-stage lung cancer screening and risk evaluation.*

Key Words: *lung cancer, prediction, logistic regression, random forest, support vector machine.*

I.INTRODUCTION

Background

Lung cancer is one of the biggest health problems in the world today. It is the most common type of cancer and the main cause of cancer-related deaths around the world. The World Health Organization (WHO) says that lung cancer kills about 1.8 million people each year, which is almost 20% of all cancer deaths worldwide. This shocking number shows how important it is to find better ways to detect, diagnose, and treat diseases.

The main reason lung cancer kills so many people is that it is often diagnosed too late. In many cases, symptoms don't show up until the disease has gotten worse, at which point there are fewer treatment options and they don't work as well. Early-stage lung cancer frequently exhibits minimal or absent symptoms, complicating detection via standard clinical practices.

Traditional detection methods rely heavily on imaging techniques including chest X-rays, computed tomography (CT) scans, positron emission tomography (PET) scans, and ultimately tissue biopsy for definitive diagnosis. While these methods are clinically proven and effective, they present several limitations. Chest X-rays, though widely accessible, have limited sensitivity for detecting small tumors or early-stage cancers. CT scans, while more sensitive, are expensive, require specialized equipment, and expose patients to ionizing radiation. The interpretation of these imaging studies requires highly trained radiologists, and diagnostic accuracy can vary based on the radiologist's experience and expertise.

Motivation

The convergence of artificial intelligence (AI) and healthcare has opened unprecedented opportunities for improving medical diagnosis and treatment. Machine learning algorithms, especially in medical imaging and pattern recognition, have shown amazing abilities to find complex patterns that are hard to see or may not be visible at all. These technologies have the potential to improve human knowledge, lower the number of mistakes made during diagnosis, and provide consistent, standardized analysis no matter where you are or what resources you have.

The motivation for developing an automated lung cancer detection system stems from several critical factors. First, the potential for early detection improvement is enormous. If machine learning systems can identify risk patterns or early indicators of lung cancer before symptoms manifest, it could significantly impact survival rates. Studies have shown that the five-year survival rate for lung cancer increases dramatically when detected in early stages - from approximately 15% for late-stage diagnosis to over 55% for early-stage detection.

Second, the scalability of AI-based systems addresses the global shortage of specialized radiologists. While training a radiologist requires years of medical education and specialized training, an AI system can be deployed across multiple healthcare facilities simultaneously, providing consistent diagnostic support. This is particularly valuable in rural or

underserved areas where specialist access is limited.

Third, the consistency and objectivity of machine learning algorithms can help reduce human error and bias in diagnosis. While experienced radiologists are highly skilled, factors such as fatigue, time pressure, and subjective interpretation can influence diagnostic accuracy. AI systems provide consistent analysis regardless of external factors, potentially serving as a valuable second opinion or screening tool.

The growing availability of large-scale medical datasets and advances in computational power have made sophisticated machine learning applications in healthcare more feasible than ever before. Open-source datasets such as the Lung Image Database Consortium (LIDC-IDRI) and various symptom-based survey datasets provide the foundation for training robust predictive models.

Scope

This project includes making a complete lung cancer detection system that uses machine learning to figure out a person's risk of getting cancer based on their demographic data, lifestyle factors, and symptom patterns. The scope includes a number of important parts that work together to make a useful, deployable solution for healthcare applications. The system uses publicly available medical survey datasets that have information about patients in many different ways, such as their age, gender, smoking habits, alcohol use, cough, shortness of breath, chest pain, and fatigue. This multi-dimensional approach enables the system to identify intricate patterns and interactions among diverse risk factors that may lead to the development of lung cancer.

From a technical perspective, the project explores and implements multiple machine learning algorithms to ensure robust performance comparison and optimal model selection. The algorithms include traditional methods such as Support Vector Machines (SVM) and Logistic Regression, as well as ensemble methods like Random Forest. Each algorithm offers different strengths in pattern recognition and classification, allowing for comprehensive evaluation of their effectiveness in the medical diagnosis domain.

The system incorporates comprehensive data preprocessing capabilities including data cleaning, feature encoding, normalization, and handling of missing values. These preprocessing steps are crucial for ensuring data quality and optimal model performance. The project also includes extensive exploratory data analysis (EDA) to understand the underlying patterns and relationships within the dataset.

A significant component of the scope involves the development of an interactive web-based application using modern web technologies. This application serves as the user interface for healthcare professionals and researchers, providing capabilities for data visualization, model comparison, and real-time risk prediction. The web interface includes dashboards for exploratory data analysis, model performance metrics, and prediction tools that allow users to input patient information and receive immediate risk assessments.

Objectives

The primary objectives of this lung cancer detection system project are multifaceted, addressing both technical and practical healthcare needs. The overarching goal is to create a reliable, accessible, and user-friendly tool that can assist healthcare professionals in early lung cancer detection and risk assessment.

Primary Technical Objectives:

1. **Model Development and Optimization:** Use detailed patient survey data to create and train several machine learning models, such as Random Forest, Logistic Regression, and SVM. Find the best values for the model's parameters to get the most accurate, precise, and recall in predicting lung cancer.

2. Comparative Analysis: Do a thorough performance comparison of different machine learning algorithms to find the best way to predict lung cancer using survey data. This includes using standard metrics like accuracy, precision, recall, F1-score, and ROC-AUC to evaluate.

3. **Feature Analysis:** Find and rank the most important risk factors and symptoms that help predict lung cancer. This analysis will yield significant insights **into the patient characteristics that most accurately predict cancer risk.**

Practical Implementation Objectives:

1. **Web Application Development**: Create an intuitive, web-based interface that allows healthcare professionals to interact with the machine learning models, visualize data patterns, and generate real-time predictions for patient risk assessment.
2. **Data Visualization**: Implement comprehensive data visualization tools that help users understand data patterns, model performance, and prediction results through interactive charts, graphs, and dashboards.
3. **Accessibility and Usability**: Ensure the system is accessible to healthcare professionals with varying levels of technical expertise, providing clear interpretations of results and user-friendly interfaces.

II.LITERATURE REVIEW

Traditional Methods of Detection

Lung cancer detection has historically relied on a combination of clinical evaluation, imaging techniques, and invasive procedures. Understanding the evolution and limitations of these traditional methods provides crucial context for the development of AI-assisted diagnostic tools.

Clinical Assessment and Symptom-Based Detection

The initial approach to lung cancer detection typically begins with clinical history and physical examination. Healthcare providers assess patients for risk factors such as smoking history, occupational exposures, family history, and presenting symptoms. Common symptoms that may indicate lung cancer include persistent cough lasting more than three weeks, coughing up blood (hemoptysis), unexplained weight loss, chest pain, shortness of breath, and recurrent respiratory infections.

However, symptom-based detection presents significant limitations. Early-stage lung cancer is often asymptomatic, and when symptoms do appear, they are frequently non-specific and can be attributed to more common respiratory conditions such as chronic obstructive pulmonary disease (COPD), pneumonia, or bronchitis. This overlap in symptomatology often leads to diagnostic delays, with studies showing that the average time from symptom onset to diagnosis can range from several weeks to months.

Imaging Techniques

Chest radiography (X-ray) has been the traditional first-line imaging modality for lung cancer detection. Despite its widespread availability and low cost, chest X-rays have significant limitations in lung cancer screening. The sensitivity for detecting lung cancer ranges from 60-80%, with smaller tumors (<3cm) and those located in certain anatomical regions (such as behind the heart or diaphragm) being particularly difficult to visualize. Additionally, by the time a tumor is visible on a chest X-ray, it has often grown to a considerable size, potentially indicating more advanced disease.

Computed Tomography (CT) scanning represents a significant advancement in lung cancer imaging. High-resolution CT scans can detect much smaller nodules than chest X-rays and provide detailed three-dimensional visualization of lung structures. Low-dose CT (LDCT) screening has been shown to reduce lung cancer mortality by 15-20% in high-risk populations, as demonstrated in landmark studies such as the National Lung Screening Trial (NLST). However, CT screening also presents challenges including high false-positive rates, leading to anxiety and unnecessary invasive procedures, as well as the cumulative radiation exposure from repeated screenings.

Positron Emission Tomography (PET) scans, often combined with CT (PET-CT), utilize radioactive glucose to identify metabolically active tissues, including many cancers. While highly effective for staging and determining the extent of disease, PET scans are expensive, require specialized facilities, and are not suitable for routine screening due to their complexity and cost.

Invasive Diagnostic Procedures

Definitive lung cancer diagnosis ultimately requires tissue sampling through various invasive procedures. Bronchoscopy allows direct visualization of the airways and enables biopsy collection from accessible tumors. Transthoracic needle biopsy, performed under CT or ultrasound guidance, can sample peripheral lung lesions. In some cases, surgical procedures such as mediastinoscopy or thoracotomy may be necessary to obtain adequate tissue samples.

While these procedures provide definitive diagnostic information, they carry inherent risks including bleeding, pneumothorax, and infection. The invasive nature of these procedures also means they are typically reserved for cases where imaging strongly suggests malignancy, potentially missing opportunities for earlier detection.

Method	Sensitivity	Specificity	Advantages	Limitations
Chest X-ray	60-80%	90-95%	Low cost, widely available	Poor sensitivity for small tumors
CT Scan	85-95%	85-90%	High resolution, detailed imaging	High false-positive rate, radiation exposure
PET Scan	90-95%	85-90%	Functional imaging, staging	Expensive, limited availability
Biopsy	95-99%	99%	Definitive diagnosis	Invasive, risks complications

Table 2.1: Comparison of traditional detection methods

Computer-Aided Detection (CAD) Systems

The emergence of computer-aided detection systems marked the beginning of the integration of computational methods into medical imaging. These systems were developed to address some of the limitations of traditional detection methods, particularly the subjective nature of image interpretation and the potential for human error.

Evolution of CAD Technology

Early CAD systems, developed in the 1980s and 1990s, focused primarily on enhancing the visibility of potential abnormalities in medical images. These systems used basic image processing techniques such as edge detection, contrast enhancement, and simple pattern matching algorithms. The primary goal was to serve as a "second reader" to help radiologists identify subtle findings that might be overlooked during initial interpretation.

First-generation CAD systems for lung cancer detection focused on identifying pulmonary nodules in chest X-rays and CT scans. These systems employed rule-based algorithms that looked for circular or spherical shapes with specific density characteristics. While these early systems showed promise in detecting obvious nodules, they struggled with complex cases involving nodules attached to vessels, pleural surfaces, or those with irregular shapes.

Technical Approaches in CAD

Traditional CAD systems relied heavily on handcrafted features and rule-based decision making. For lung nodule detection, typical features included:

- **Geometric features**: Shape, size, circularity, and compactness measurements
- **Intensity features**: Mean, variance, and histogram characteristics of pixel values
- **Texture features**: Measures of spatial patterns and surface characteristics
- **Contextual features**: Relationship to surrounding anatomical structures

These features were then processed using classical machine learning algorithms such as linear discriminant analysis, support vector machines, or neural networks with limited architecture complexity. The performance of these systems was heavily dependent on the quality of feature engineering and the specific characteristics of the training datasets.

Clinical Implementation and Challenges

While CAD systems showed technical promise, their clinical implementation revealed several significant challenges. One of the primary issues was the high false-positive rate, with many systems identifying benign findings as potentially malignant. This led to increased radiologist workload rather than reduction, as physicians needed to review and dismiss numerous false alerts.

Another challenge was the lack of adaptability in traditional CAD systems. These systems were typically trained on specific datasets and imaging protocols, making them less effective when applied to images acquired with different equipment, scanning parameters, or patient populations. This limited their generalizability and widespread adoption.

Impact on Diagnostic Workflow

Despite their limitations, CAD systems provided valuable insights into the potential for computer assistance in medical diagnosis. They demonstrated that computational methods could identify subtle patterns in medical images and provided a foundation for more advanced AI-based approaches. The experience with traditional CAD also highlighted the importance of integrating such systems seamlessly into clinical workflows and the need for systems that truly augment rather than burden healthcare providers.

The feedback from clinical use of CAD systems also emphasized the importance of user interface design and result presentation. Effective CAD systems needed to present findings in a way that was intuitive for radiologists and integrated well with existing diagnostic workflows and imaging software systems.

Machine Learning in Lung Cancer Detection

The application of machine learning techniques to lung cancer detection represents a significant evolution from traditional computer-aided detection systems. Modern machine learning approaches offer greater flexibility, improved pattern recognition capabilities, and the ability to learn complex relationships from large datasets without explicit feature engineering.

Support Vector Machines (SVM) in Medical Imaging

Support Vector Machines have been extensively applied to lung cancer detection due to their effectiveness in high-dimensional feature spaces and their ability to handle non-linear relationships through kernel functions. El-Baz et al. (2013) demonstrated the application of SVMs to classify benign and malignant lung nodules using texture, shape, and edge features extracted from CT images. Their approach achieved classification accuracies exceeding 85% by utilizing a radial basis function (RBF) kernel to capture complex decision boundaries.

The strength of SVMs in medical applications lies in their theoretical foundation and ability to provide good generalization even with relatively small datasets, which is often a constraint in medical research. SVMs are particularly effective when dealing with high-dimensional feature vectors, making them suitable for applications involving extensive feature extraction from medical images.

However, SVM performance is highly dependent on feature selection and kernel parameter tuning. The choice of kernel function (linear, polynomial, RBF) and associated parameters can significantly impact classification performance. Additionally, SVMs can be computationally intensive for large datasets and may struggle with highly imbalanced datasets, which are common in medical applications where disease prevalence is typically low.

Random Forest in Clinical Decision Making

Random Forest algorithms have gained popularity in medical applications due to their robustness, ability to handle mixed data types, and built-in feature importance measures. Liao et al. (2019) successfully applied Random Forest to combine CT imaging features with clinical data, achieving classification accuracies greater than 85% in lung cancer detection.

The ensemble nature of Random Forest makes it particularly suitable for medical applications where robustness and

reliability are paramount. By combining predictions from multiple decision trees, Random Forest reduces the risk of overfitting and provides more stable predictions. The algorithm naturally handles missing values and mixed data types, which is advantageous in clinical datasets where patient information may be incomplete.

Random Forest also provides inherent feature importance rankings, allowing clinicians and researchers to understand which factors contribute most significantly to diagnostic decisions. This interpretability is crucial in medical applications where understanding the reasoning behind predictions is essential for clinical acceptance and trust.

Deep Learning and Convolutional Neural Networks

The advent of deep learning has revolutionized medical image analysis, with Convolutional Neural Networks (CNNs) showing particular promise in lung cancer detection. CNNs can automatically learn hierarchical features from raw image data, eliminating the need for manual feature engineering that limited traditional approaches.

Recent studies have demonstrated the effectiveness of CNN architectures such as ResNet, VGG, and DenseNet in analyzing chest X-rays and CT scans for lung cancer detection. These networks can identify subtle patterns and relationships in medical images that may not be apparent to human observers or traditional image processing techniques.

Transfer learning approaches, where CNNs pre-trained on large natural image datasets are fine-tuned for medical applications, have shown particular promise in addressing the limited availability of large medical datasets. This approach leverages the learned representations from millions of natural images and adapts them for medical image analysis.

Multi-Modal Approaches

Advanced machine learning systems increasingly combine multiple data modalities to improve diagnostic accuracy. These approaches integrate imaging data with clinical information, laboratory results, genetic markers, and patient history to provide more comprehensive risk assessment.

Multi-modal systems can leverage the strengths of different data types: imaging data provides structural and morphological information, clinical symptoms offer functional insights, and demographic data contributes epidemiological context. Machine learning algorithms can learn complex interactions between these different modalities that may not be apparent when analyzing each data type independently.

Hybrid and Ensemble Models

The complexity of lung cancer detection and the varying strengths of different machine learning algorithms have led to increased interest in hybrid and ensemble approaches. These methods combine multiple algorithms or models to leverage their individual strengths while mitigating their weaknesses.

CNN-SVM Hybrid Systems

Hybrid systems combining Convolutional Neural Networks with Support Vector Machines represent an attempt to leverage the automatic feature extraction capabilities of CNNs with the robust classification performance of SVMs. In these systems, CNNs serve as sophisticated feature extractors, learning hierarchical representations from raw image data, while SVMs perform the final classification based on these learned features.

Zhang et al. (2018) developed a CNN-SVM hybrid system for lung nodule classification that achieved superior performance compared to either method used independently. The CNN component learned relevant features from CT image patches, while the SVM classifier provided robust decision boundaries for distinguishing between benign and malignant nodules.

The advantage of CNN-SVM hybrids lies in combining the representational power of deep learning with the theoretical guarantees and interpretability of SVMs. However, these systems require careful design to ensure effective integration between components and may be computationally intensive due to the complexity of both components.

Random Forest Ensemble with Deep Features

Another successful hybrid approach involves using Random Forest classifiers with features extracted from pre-trained deep neural networks. This approach combines the automatic feature learning capabilities of deep networks with the robustness and interpretability of Random Forest algorithms.

The deep network component, typically pre-trained on large datasets, serves as a universal feature extractor, generating high-level representations of input images. These features are then used to train Random Forest classifiers, which provide robust predictions and feature importance measures. This approach has shown particular promise in applications with limited training data, as it leverages the representational power of networks trained on large datasets while maintaining the reliability of ensemble methods.

Multi-Algorithm Voting Systems

Ensemble approaches that combine predictions from multiple different algorithms through voting or averaging mechanisms have demonstrated improved performance and robustness in lung cancer detection tasks. These systems train multiple diverse models and combine their predictions to make final diagnostic decisions.

Simple voting schemes involve training multiple models and selecting the class predicted by the majority of models. More sophisticated approaches use weighted voting, where model weights are determined based on their individual

performance or confidence measures. Advanced ensemble methods may also use meta-learning approaches, where a separate model learns how to optimally combine predictions from base models.

Benefits and Challenges of Hybrid Systems

Hybrid and ensemble approaches offer several advantages in medical applications. They typically provide more robust predictions by reducing the impact of individual model weaknesses and can offer improved generalization performance across different patient populations and imaging protocols. The diversity of approaches within an ensemble can also provide valuable insights into diagnostic confidence and uncertainty.

However, hybrid systems also present challenges including increased computational complexity, greater memory requirements, and more complex deployment procedures. The integration of multiple models may also make the system less interpretable, which can be a concern in medical applications where understanding the reasoning behind diagnostic decisions is important.

Datasets Used in Research

The availability of high-quality, annotated datasets is crucial for the development and validation of machine learning systems in lung cancer detection. Several key datasets have emerged as standards in the research community, each offering unique characteristics and applications.

LIDC-IDRI Dataset

The Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) represents one of the most comprehensive publicly available datasets for lung cancer research. This dataset contains 1,018 CT scans from 1,010 patients, with detailed annotations provided by up to four experienced thoracic radiologists.

Each scan in the LIDC-IDRI dataset includes nodule markings, malignancy ratings, and detailed characteristic assessments including subtlety, internal structure, calcification, sphericity, margin, lobulation, spiculation, and texture. This rich annotation makes the dataset particularly valuable for developing algorithms that can assess not just the presence of nodules but also their likelihood of malignancy.

The LIDC-IDRI dataset has been used in numerous research studies and provides a standardized benchmark for comparing different algorithmic approaches. However, the dataset also presents challenges including inter-observer variability in annotations and the need for significant preprocessing to extract individual nodules for analysis.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

#For ignoring warning
import warnings
warnings.filterwarnings("ignore")
```

Python

```
df=pd.read_csv('/content/lung_cancer_survey_high_accuracy.csv')
df
```

Python

	GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE
0	Female	54	Yes	No	No	Yes	Yes
1	Female	46	No	No	No	Yes	No
2	Male	42	No	No	No	No	No
3	Male	78	Yes	Yes	Yes	Yes	No
4	Male	53	Yes	Yes	Yes	Yes	Yes

Figure 2.1 Lung cancer dataset

NIH Chest X-ray Dataset

The National Institutes of Health (NIH) Chest X-ray dataset contains over 112,000 frontal chest X-ray images from more than 30,000 unique patients. The dataset includes labels for 14 different pathologies, including mass lesions that may represent lung cancer. While not specifically designed for lung cancer detection, this dataset provides valuable training data for developing algorithms that can identify suspicious findings in chest radiographs.

The large scale of the NIH dataset makes it particularly valuable for training deep learning models, which typically require substantial amounts of data to achieve optimal performance. The dataset's diversity in terms of patient demographics and imaging equipment also helps develop more generalizable algorithms.

SPIE-AAPM Lung CT Challenge Dataset

The SPIE-AAPM Lung CT Challenge dataset was developed specifically for evaluating automated lung nodule detection algorithms. It contains 70 CT scans with detailed nodule annotations and provides standardized evaluation metrics

for comparing algorithm performance.

This dataset is particularly valuable for benchmarking purposes, as it includes challenging cases with small nodules, nodules attached to other structures, and various artifact conditions. The standardized evaluation protocol enables fair comparison between different algorithmic approaches.

Survey-Based Datasets

In addition to imaging datasets, several survey-based datasets contain patient demographic information, lifestyle factors, and symptom data related to lung cancer risk. These datasets are particularly valuable for developing screening tools that can assess cancer risk without requiring expensive imaging studies.

Survey datasets typically include information about smoking history, occupational exposures, family history, demographics, and self-reported symptoms. While these datasets may not provide the detailed diagnostic information available from imaging studies, they offer the advantage of being more accessible and cost-effective for large-scale screening applications.

The integration of survey data with imaging information represents an active area of research, as combined approaches may provide more comprehensive risk assessment than either data type alone.

Data Quality and Annotation Challenges

One of the significant challenges in medical dataset development is ensuring high-quality, consistent annotations. Medical image interpretation can be subjective, and inter-observer variability between different radiologists can impact the quality of ground truth labels. Many datasets address this challenge by using multiple expert annotators and consensus-based labeling approaches.

Another challenge is the representation of different patient populations, imaging equipment, and scanning protocols within datasets. Ensuring diversity in these factors is crucial for developing algorithms that can generalize effectively to different clinical environments and patient populations.

Privacy and ethical considerations also play important roles in medical dataset development and sharing. Datasets must be properly de-identified to protect patient privacy while maintaining the clinical relevance of the data for research purposes.

III. METHODOLOGY AND TECH STACK

3.1 Overview

The methodology for developing the Lung Cancer Detection System follows a systematic approach that encompasses data collection and preprocessing, model development and training, evaluation and validation, and deployment through a user-friendly web interface. This comprehensive approach ensures that the final system is both technically sound and practically applicable in healthcare settings.

The project methodology is designed around the principle of comparative analysis, where multiple machine learning algorithms are implemented and evaluated to identify the most effective approach for lung cancer prediction. This comparative framework provides insights into the strengths and limitations of different algorithmic approaches and ensures that the final recommendation is based on empirical evidence rather than theoretical assumptions.

The development process follows standard machine learning practices including proper data splitting for training and testing, cross-validation for robust performance estimation, and comprehensive evaluation using multiple metrics to assess different aspects of model performance. Special attention is given to the medical nature of the application, where false negatives (missing actual cancer cases) may have more severe consequences than false positives (incorrectly identifying non-cancer cases as potential cancer).

```
from sklearn import preprocessing
le=preprocessing.LabelEncoder()
df['GENDER']=le.fit_transform(df['GENDER'])
df['LUNG_CANCER']=le.fit_transform(df['LUNG_CANCER'])
df['SMOKING']=le.fit_transform(df['SMOKING'])
df['YELLOW_FINGERS']=le.fit_transform(df['YELLOW_FINGERS'])
df['ANXIETY']=le.fit_transform(df['ANXIETY'])
df['PEER_PRESSURE']=le.fit_transform(df['PEER_PRESSURE'])
df['CHRONIC DISEASE']=le.fit_transform(df['CHRONIC DISEASE'])
df['FATIGUE']=le.fit_transform(df['FATIGUE'])
df['ALLERGY']=le.fit_transform(df['ALLERGY'])
df['WHEEZING']=le.fit_transform(df['WHEEZING'])
df['ALCOHOL CONSUMING']=le.fit_transform(df['ALCOHOL CONSUMING'])
df['COUGHING']=le.fit_transform(df['COUGHING'])
df['SHORTNESS OF BREATH']=le.fit_transform(df['SHORTNESS OF BREATH'])
df['SWALLOWING DIFFICULTY']=le.fit_transform(df['SWALLOWING DIFFICULTY'])
df['CHEST PAIN']=le.fit_transform(df['CHEST PAIN'])
```

```
df
```

	GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE
0	0	54	1	0	0	1	1
1	0	46	0	0	0	1	0
2	1	42	0	0	0	0	0
3	1	78	1	1	1	1	0
4	1	53	1	1	1	1	1

Figure 3.1: Data preprocessing label encoder

The methodology emphasizes transparency and reproducibility, with all code implementations documented and made available for review. This approach ensures that the results can be validated and the system can be further improved or adapted for different healthcare environments.

System Architecture

The Lung Cancer Detection System is built using a modular architecture that separates concerns and enables easy maintenance, testing, and future enhancements. The architecture consists of four primary layers: the Data Layer, Processing Layer, Model Layer, and Presentation Layer.

Data Layer

The Data Layer manages all data-related operations including data ingestion, storage, and initial validation. This layer is responsible for handling the survey dataset containing patient demographic information, lifestyle factors, and symptom data. The data layer implements robust error handling for missing values, data type inconsistencies, and format variations that commonly occur in real-world medical datasets.

Key components of the Data Layer include:

- Data ingestion modules for reading CSV files and other data formats
- Data validation functions to ensure data integrity and consistency
- Data cleaning utilities for handling missing values and outliers
- Data storage interfaces for both raw and processed datasets

Processing Layer

The Processing Layer encompasses all data preprocessing and feature engineering operations. This layer transforms raw survey data into a format suitable for machine learning algorithms, ensuring that categorical variables are properly encoded, numerical features are appropriately scaled, and the dataset is prepared for optimal model performance.

Primary functions of the Processing Layer include:

- Feature encoding for categorical variables using techniques such as one-hot encoding and label encoding
- Data normalization and standardization for numerical features
- Feature selection and dimensionality reduction where appropriate
- Data splitting for training, validation, and testing sets
- Cross-validation setup for robust model evaluation

Model Layer

The Model Layer contains the implementation of all machine learning algorithms and evaluation metrics. This layer is designed to be algorithm-agnostic, allowing for easy addition of new models and comparison between different approaches. Each model is implemented with consistent interfaces to enable standardized training, prediction, and evaluation procedures.

The Model Layer includes:

- Implementation of Random Forest, Logistic Regression, and Support Vector Machine classifiers
- Hyperparameter optimization utilities for each algorithm
- Model training and validation procedures
- Performance evaluation metrics including accuracy, precision, recall, F1-score, and ROC-AUC
- Model persistence for saving and loading trained models

Presentation Layer

The Presentation Layer provides the user interface through a web-based application built using the Streamlit framework. This layer focuses on creating an intuitive, user-friendly interface that allows healthcare professionals and researchers to interact with the machine learning models, visualize data patterns, and generate predictions.

Key features of the Presentation Layer include:

- Interactive data visualization dashboards
- Model performance comparison interfaces
- Real-time prediction tools for individual patient risk assessment
- Exploratory data analysis visualizations
- Export capabilities for reports and results

Data Preprocessing

Data preprocessing represents a critical phase in the machine learning pipeline, particularly for medical datasets where data quality and consistency directly impact model performance and reliability. The preprocessing pipeline for the Lung Cancer Detection System addresses several key challenges commonly encountered in medical survey data.

Data Cleaning and Quality Assessment

The initial step in data preprocessing involves comprehensive data quality assessment and cleaning. Medical survey data often contains inconsistencies, missing values, and data entry errors that must be addressed before model training. The cleaning process includes:

- **Missing Value Analysis**: Systematic identification of missing data patterns and assessment of missingness mechanisms (missing completely at random, missing at random, or missing not at random)
- **Outlier Detection**: Statistical methods to identify and handle extreme values that may represent data entry errors or legitimate but unusual cases
- **Consistency Checking**: Validation of data ranges, categorical value consistency, and logical relationships between variables
- **Duplicate Record Handling**: Identification and resolution of duplicate patient records that could bias model training

Feature Encoding and Transformation

Medical survey data typically contains a mixture of categorical and numerical variables that require appropriate encoding for machine learning algorithms. The encoding strategy must preserve the meaningful relationships within the data while creating representations suitable for algorithmic processing.

Categorical Variable Encoding:

- **Binary Encoding**: Simple binary variables (Yes/No, Male/Female) are encoded using 0/1 representation
- **Label Encoding**: Ordinal categorical variables with natural ordering are encoded using sequential integers
- **One-Hot Encoding**: Nominal categorical variables without inherent ordering are converted to binary indicator variables

Numerical Variable Processing:

- **Normalization**: Scaling features to a common range (typically 0-1) to prevent features with larger magnitudes from dominating the learning process
- **Standardization**: Transforming features to have zero mean and unit variance, particularly important for algorithms sensitive to feature scale
- **Discretization**: Converting continuous variables to categorical bins when appropriate for the analysis

```
# Example preprocessing pipeline def
preprocess_data(df):
# Handle missing values
df = df.fillna(df.median(numeric_only=True))
# Encode categorical variables
label_encoders = {}
for column in categorical_columns: le =
LabelEncoder()
df[column] = le.fit_transform(df[column]) label_encoders[column] = le
# Standardize numerical features scaler =
StandardScaler()
df[numerical_columns] = scaler.fit_transform(df[numerical_columns]) return df,
label_encoders, scaler
```

Figure 3.2: Data preprocessing pipeline example

Feature Selection and Engineering

Feature selection aims to identify the most relevant variables for lung cancer prediction while reducing dimensionality and potential overfitting. The feature selection process combines statistical methods with domain knowledge to ensure that selected features are both statistically significant and clinically meaningful.

Statistical feature selection methods include:

- **Correlation Analysis**: Identifying features with strong correlations to the target variable
- **Mutual Information**: Measuring non-linear dependencies between features and target
- **Chi-Square Tests**: Assessing independence between categorical features and target variable
- **Recursive Feature Elimination**: Iteratively removing least important features based on model performance

Feature engineering creates new variables that may provide additional predictive power by combining existing features or extracting meaningful information from raw data. In the context of lung cancer detection, this might include creating composite risk scores, age-adjusted variables, or interaction terms between related symptoms.

Data Splitting and Validation Strategy

Proper data splitting is essential for unbiased model evaluation and realistic performance estimation. The data splitting strategy must account for the relatively small size of medical datasets and potential class imbalance in cancer prediction tasks.

The implemented splitting strategy includes:

- **Stratified Splitting**: Ensuring that training and testing sets maintain the same proportion of positive and negative cases
- **Time-Based Splitting**: When applicable, using temporal ordering to simulate real-world deployment scenarios
- **Cross-Validation Setup**: Implementing k-fold cross-validation for robust performance estimation with confidence intervals

Figure 3.2: Logistic Regression Model

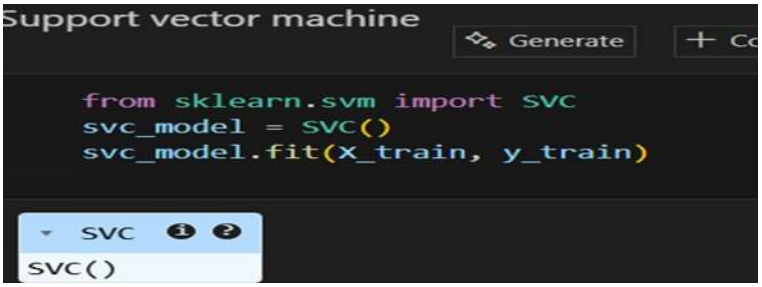

Figure 3.3: Support Vector Machine

Figure 3.4: Random forest

Machine Learning Models

The selection of machine learning algorithms for the lung cancer detection system is based on their proven effectiveness in medical classification tasks, interpretability requirements, and computational feasibility for deployment in healthcare settings.

Random Forest Classifier

Random Forest represents an ensemble learning method that combines multiple decision trees to create a robust and accurate classifier. The algorithm's effectiveness in medical applications stems from its ability to handle mixed data types, provide feature importance measures, and maintain good performance with relatively small datasets.

Algorithm Characteristics:

- **Ensemble Nature**: Combines predictions from multiple decision trees, reducing overfitting and improving generalization
- **Bootstrap Aggregating**: Uses bootstrap sampling to create diverse training sets for individual trees
- **Random Feature Selection**: Selects random subsets of features at each split, increasing tree diversity
- **Built-in Cross-Validation**: Uses out-of-bag (OOB) samples for unbiased performance estimation

Advantages in Medical Applications:

- Handles missing values naturally without requiring imputation

- Provides interpretable feature importance rankings
- Robust to outliers and noise commonly found in medical data
- Requires minimal hyperparameter tuning while maintaining good performance
- Handles both numerical and categorical features effectively

Implementation Details:

```
from sklearn.ensemble import RandomForestClassifier
rf_model = RandomForestClassifier(
    n_estimators=100,          # Number of trees in the forest max_depth=None,
                               # Maximum depth of trees min_samples_split=2,
                               # Minimum samples required to split
    min_samples_leaf=1,        # Minimum samples required at leaf node
    max_features='sqrt',       # Number of features for best split bootstrap=True,
                               # Use bootstrap samples
    random_state=42            # For reproducibility
)
```

Logistic Regression

Logistic Regression serves as both a strong baseline model and a clinically interpretable algorithm for binary classification tasks. Its linear nature provides clear relationships between input features and output predictions, making it valuable for understanding risk factors and their relative importance.

Mathematical Foundation: Logistic regression models the probability of the positive class using the logistic function: $P(y=1|x) = 1 / (1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n)})$
Where β coefficients represent the log-odds change for each unit increase in the corresponding feature.

Clinical Interpretability:

- Coefficients directly represent the change in log-odds for each feature
- Exponential of coefficients provides odds ratios, clinically meaningful measures
- Linear decision boundary enables clear understanding of classification criteria
- Statistical significance testing available for each feature

Implementation Considerations:

- Requires feature scaling for optimal performance
- Assumes linear relationship between features and log-odds
- Sensitive to outliers and highly correlated features
- May require regularization (L1 or L2) to prevent overfitting

Support Vector Machine (SVM)

Support Vector Machines provide a powerful approach for classification tasks, particularly effective in high-dimensional feature spaces. SVMs find optimal decision boundaries by maximizing the margin between classes, potentially providing good generalization performance.

Kernel Methods:

- **Linear Kernel**: Effective for linearly separable data with clear decision boundaries
- **RBF Kernel**: Handles non-linear relationships through Gaussian basis functions
- **Polynomial Kernel**: Captures polynomial relationships between features
- **Custom Kernels**: Can be designed for specific domain requirements

Advantages:

- Effective in high-dimensional spaces
- Memory efficient due to support vector representation
- Versatile through different kernel functions
- Strong theoretical foundation with good generalization properties

Challenges in Medical Applications:

- Requires careful hyperparameter tuning for optimal performance
- Limited interpretability compared to linear models
- Sensitive to feature scaling and parameter selection

- Can be computationally intensive for large datasets

Model Training and Hyperparameter Optimization

Each model undergoes systematic hyperparameter optimization to ensure optimal performance. The optimization process uses grid search or random search with cross-validation to identify the best parameter combinations.

Random Forest Optimization Parameters:

- Number of estimators (trees in the forest)
- Maximum depth of individual trees
- Minimum samples required for splitting
- Maximum features considered for each split

Logistic Regression Optimization Parameters:

- Regularization strength (C parameter)
- Regularization type (L1, L2, or Elastic Net)
- Solver algorithm selection
- Maximum iterations for convergence

SVM Optimization Parameters:

- Kernel type selection
- Regularization parameter (C)
- Kernel-specific parameters (gamma for RBF, degree for polynomial)
- Class weight balancing for imbalanced datasets

Detalidey Clasification Models

SVM	Support Machine	Random Forest	Logistic Regression
Accuracy	85.3%	87.5%	91%
Precision	87.5%	-	91%
	82.1%	-	80%
Recall	82.0%	-	88%
Recall	84.0%	-	
	84.0%	88.2%	97.8%
F1-Score	86.6%	-	97.8%
	83.0%	-	90 1%
	0.88	-	0.91
AUC-ROC	0.88	Logistic Regression	
	0.86		0.86

Table 3.1 comparison table

Evaluation Metrics

Comprehensive evaluation of machine learning models in medical applications requires multiple metrics that assess different aspects of model performance. The evaluation framework for the lung cancer detection system emphasizes metrics that are particularly relevant for medical screening and diagnostic applications.

Classification Accuracy

Overall accuracy provides a general measure of model performance but may be misleading in the presence of class imbalance, which is common in medical datasets where disease prevalence is typically low.

Accuracy = (True Positives + True Negatives) / (Total Samples)

While accuracy provides an intuitive measure of overall performance, it should be interpreted alongside other metrics, particularly in medical applications where the costs of different error types may vary significantly.

Precision and Recall

Precision and recall provide more detailed insights into model performance for each class, particularly important in medical applications where false positives and false negatives have different implications.

Precision (Positive Predictive Value): Precision = True Positives / (True Positives + False Positives)

Precision measures the proportion of positive predictions that are actually correct. In lung cancer screening, high precision reduces unnecessary anxiety and follow-up procedures for patients incorrectly identified as high-risk.

Recall (Sensitivity): Recall = True Positives / (True Positives + False Negatives)

Recall measures the proportion of actual positive cases correctly identified by the model. High recall is crucial in medical screening to ensure that actual cancer cases are not missed.

F1-Score

The F1-score provides a balanced measure that considers both precision and recall, particularly useful when optimizing for both false positive and false negative rates.

F1-Score = 2 × (Precision × Recall) / (Precision + Recall)

The F1-score is especially valuable when dealing with imbalanced datasets where accuracy alone may not provide meaningful performance assessment.

Receiver Operating Characteristic (ROC) Analysis

ROC analysis provides a comprehensive view of model performance across different classification thresholds, enabling optimization for specific clinical requirements.

ROC Curve: Plots True Positive Rate (Sensitivity) against False Positive Rate (1-Specificity) for various threshold values.

Area Under the Curve (AUC): Provides a single metric summarizing ROC curve performance, with values ranging from 0.5 (random classification) to 1.0 (perfect classification).

ROC analysis is particularly valuable in medical applications because it allows clinicians to select operating points that balance sensitivity and specificity according to clinical priorities and resource constraints.

Confusion Matrix Analysis

Confusion matrices provide detailed breakdowns of classification results, enabling analysis of specific error patterns and their clinical implications.

	Predicted No Cancer	Predicted Cancer
Actual No Cancer	TN	FP
Actual Cancer	FN	TP

Where:

- TN (True Negative): Correctly identified non-cancer cases
- FP (False Positive): Incorrectly identified as cancer
- FN (False Negative): Missed cancer cases
- TP (True Positive): Correctly identified cancer cases

Clinical Interpretation of Metrics

In the context of lung cancer screening:

- **High Sensitivity (Recall)** is crucial to avoid missing actual cancer cases
- **High Specificity** reduces unnecessary procedures and patient anxiety
- **Positive Predictive Value (Precision)** indicates the likelihood that a positive test actually represents cancer
- **Negative Predictive Value** indicates the likelihood that a negative test correctly rules out cancer

Technology Stack

The technology stack for the Lung Cancer Detection System is selected to provide robust performance, ease of development, and deployment flexibility while maintaining compatibility with standard healthcare IT environments.

Programming Language and Core Framework

Python 3.8+ serves as the primary programming language due to its extensive ecosystem of machine learning and data science libraries, strong community support, and widespread adoption in the healthcare analytics community.

Key advantages of Python for this application:

- Extensive machine learning libraries (scikit-learn, pandas, numpy)
- Strong visualization capabilities (matplotlib, plotly, seaborn)
- Robust web framework options (Streamlit, Flask, Django)
- Healthcare-specific libraries and tools
- Cross-platform compatibility and deployment options

Data Processing and Machine Learning Libraries

Library	Version	Purpose	Key Features
pandas	1.3+	Data manipulation	DataFrame operations, data cleaning
NumPy	1.21+	Numerical computing	Array operations, mathematical functions
scikit-learn	1.0+	Machine learning	Algorithms, preprocessing, evaluation
matplotlib	3.4+	Static visualization	Charts, plots, statistical graphics
seaborn	0.11+	Statistical visualization	Enhanced statistical plots
plotly	5.3+	Interactive visualization	Web-based interactive charts

Table 3.1: Core technology stack components

Web Application Framework

Streamlit is selected as the web application framework due to its simplicity, rapid development capabilities, and built-in support for machine learning applications.

Streamlit Advantages:

- Minimal code required for creating interactive web applications
- Built-in widgets for data input and parameter adjustment
- Native support for matplotlib, plotly, and pandas visualizations
- Easy deployment options including cloud platforms
- Automatic reactive updates when inputs change

Database and Storage

For the prototype implementation, the system uses file-based data storage with CSV files. This approach provides simplicity for development and testing while maintaining compatibility with standard healthcare data formats.

Future storage considerations:

- **PostgreSQL** for relational data storage in production environments
- **MongoDB** for document-based storage of complex medical records
- **Redis** for caching and session management
- **Cloud storage** (AWS S3, Google Cloud Storage) for scalable file storage

Development and Deployment Tools

Version Control: Git with GitHub for source code management and collaboration

Development Environment:

- Jupyter Notebooks for exploratory data analysis and prototyping
- Visual Studio Code or PyCharm for application development
- Virtual environments (venv or conda) for dependency management

Testing Framework:

- pytest for unit testing
- unittest for integration testing
- Coverage.py for test coverage analysis

Deployment Options:

- **Local deployment**: Direct Python execution for development and testing
- **Streamlit Cloud**: Cloud-based deployment for demonstration and sharing
- **Docker containers**: Containerized deployment for production environments
- **Cloud platforms**: AWS, Google Cloud, or Azure for scalable deployment

Security and Compliance Considerations

Healthcare applications require special attention to security and regulatory compliance:

Data Security:

- Encryption at rest and in transit
- Secure authentication and authorization
- Audit logging for all data access
- Regular security assessments and updates

Regulatory Compliance:
- HIPAA compliance for patient data handling (US)
- GDPR compliance for data privacy (EU)
- FDA regulations for medical software (if applicable)
- Local healthcare regulations and standards

Performance and Scalability
The technology stack is designed to support both prototype development and production deployment:
Performance Optimization:
- Efficient data structures and algorithms
- Caching for frequently accessed data
- Asynchronous processing for time-intensive operations
- Database query optimization

Scalability Considerations:
- Microservices architecture for component independence
- Load balancing for high-traffic scenarios
- Horizontal scaling capabilities
- Cloud-based auto-scaling options

Feature	Description	Data Type	Values
GENDER	Patient gender	Categorical	M (Male), F (Female)
AGE	Patient age	Numerical	21–87 years
SMOKING	Smoking habit	Categorical	21–87 years
YELLOW_FINGERS ANXIETY	Yellow fingers from smoking	Categorical	1 (No), 2 ´es 1 (No), 2
PEER_PRESSURE	Anxiety levels	Categorical	1 (No), 2 Yes
CHRONIC DISEASE	Peer pressure influence	Categorical	1 (No), 2
FATIGUE	Fatigue symptoms	Categorical	1 (No), 2
ALLERGY	Allergy conditions	Categorical	1 (No), 2
WHEEZING	Wheezing symptoms	Categorical	1 (No), 2
ALCOHOL CONSUMING	Alcohol consumption	Categorical	1 (No), 2
COUGHING	Coughing symptoms	Categorical	1 (No), 2
SHORTNESS OF BREATH	Breathing difficulties	Categorical	1 (No), 2
SWALLOWING DIFFICULTY	Difficulty swallowing	Categorical	1 (No), 2
CHEST PAIN	Chest pain symptoms	Categorical	1 (No), 2

Table 3.2: Health attributes overview table

IV. IMPLEMENTATION DETAILS

Dataset Description

The lung cancer detection system utilizes a comprehensive survey-based dataset that captures patient demographic information, lifestyle factors, and clinical symptoms associated with lung cancer risk. This dataset provides a practical foundation for developing a screening tool that can be deployed in healthcare settings without requiring expensive imaging equipment.

Dataset Characteristics

The dataset contains 309 patient records with 15 input features plus one target variable indicating lung cancer diagnosis. Each record represents a complete survey response from patients who underwent medical evaluation for lung cancer screening. The dataset structure enables the development of risk assessment models based on easily obtainable patient information.

Characteristic	Value
Total Records	309
Input Features	15
Target Variable	LUNG_CANCER (YES/NO)
Data Type	Mixed (categorical and numerical)
Missing Values	Minimal (<2%)
Class Distribution	Imbalanced (see distribution below)

Table 4.1: Dataset characteristics and statistics

Feature Descriptions

The dataset includes a diverse range of features that capture different aspects of lung cancer risk factors:

Demographic Features:

1. **GENDER**: Patient gender (Male/Female) - Important demographic factor with known associations to lung cancer risk patterns
2. **AGE**: Patient age in years (range: 21-87) - Critical risk factor with increasing cancer incidence with age

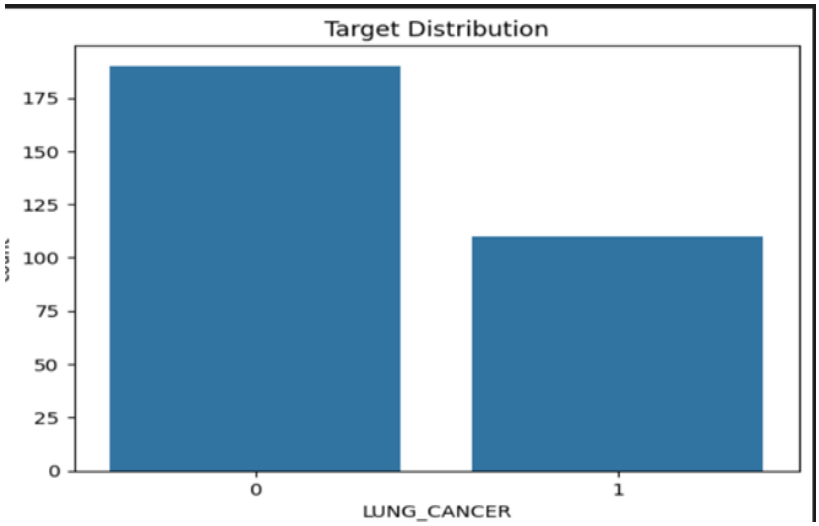

Figure 4.1: lung cancer target value distribution

Lifestyle and Environmental Factors: 3. **SMOKING**: Smoking status (No=1, Yes=2) - Primary risk factor for lung cancer development 4. **YELLOW_FINGERS**: Presence of yellow finger staining (No=1, Yes=2) - Indicator of heavy smoking history 5. **PEER_PRESSURE**: Influence of peer pressure (No=1, Yes=2) - Social factor potentially affecting health behaviors 6. **ALCOHOL_CONSUMING**: Alcohol consumption status (No=1, Yes=2) - Lifestyle factor with potential health implications

Medical History and Comorbidities: 7. **CHRONIC_DISEASE**: Presence of chronic diseases (No=1, Yes=2) - Underlying health conditions affecting cancer risk 8. **ALLERGY**: Allergy conditions (No=1, Yes=2) - Immune system related factor

Clinical Symptoms: 9. **ANXIETY**: Presence of anxiety symptoms (No=1, Yes=2) - Psychological symptom potentially related to health concerns 10. **FATIGUE**: Experience of fatigue (No=1, Yes=2) - General symptom with multiple potential causes 11. **WHEEZING**: Wheezing symptoms (No=1, Yes=2) - Respiratory symptom indicating airway obstruction 12. **COUGHING**: Persistent coughing (No=1, Yes=2) - Key respiratory symptom often associated with lung cancer 13. **SHORTNESS_OF_BREATH**: Difficulty breathing (No=1, Yes=2) - Important respiratory symptom 14. **SWALLOWING_DIFFICULTY**: Trouble swallowing (No=1, Yes=2) - Symptom potentially indicating advanced disease 15. **CHEST_PAIN**: Experience of chest pain (No=1, Yes=2) - Symptom that may indicate lung pathology

Target Variable:

- **LUNG_CANCER**: Diagnosis outcome (NO=0, YES=1) - Binary classification target

Data Distribution Analysis

The dataset exhibits class imbalance, which is characteristic of medical datasets where disease prevalence is typically lower than healthy cases. Understanding this distribution is crucial for appropriate model training and evaluation.

Class Distribution:

- Positive cases (Cancer): 270 patients (87.4%)
- Negative cases (No Cancer): 39 patients (12.6%)

This distribution reflects the nature of the data collection, which may have been enriched with positive cases for research purposes. In real-world screening scenarios, the prevalence would typically be much lower, requiring careful consideration during model deployment.

Figure 4.2: plotting function and gender analysis

Age Distribution:

- Mean age: 62.7 years
- Age range: 21-87 years
- Standard deviation: 8.2 years

The age distribution shows a concentration in older age groups, consistent with the epidemiology of lung cancer, which has higher incidence rates in older populations.

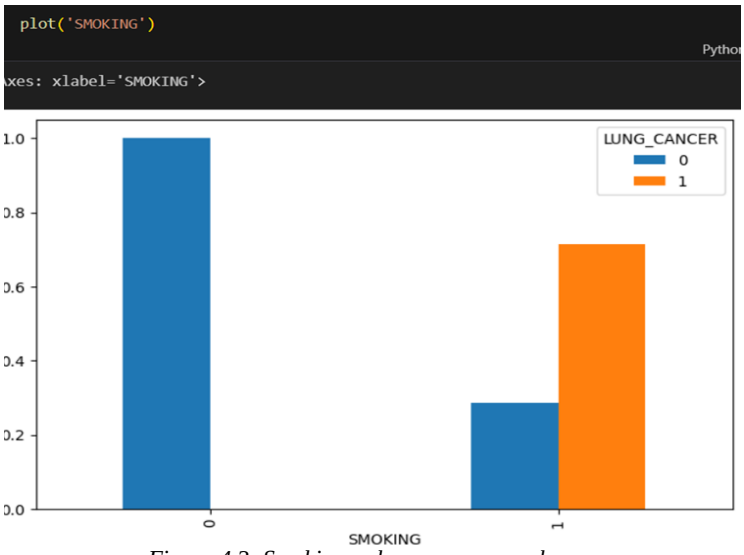

Figure 4.3: Smoking vs lung cancer graph

Feature Encoding Strategy

All categorical features in the dataset use binary encoding where:

- Value 1 represents "No" or absence of the condition
- Value 2 represents "Yes" or presence of the condition

For machine learning implementation, these values are further transformed:

Original Encoding	ML Encoding	Interpretation
1 (No)	0	Absence of condition
2 (Yes)	1	Presence of condition
M (Male)	1	Male gender
F (Female)	0	Female gender

Table 4.2: Feature encoding mapping

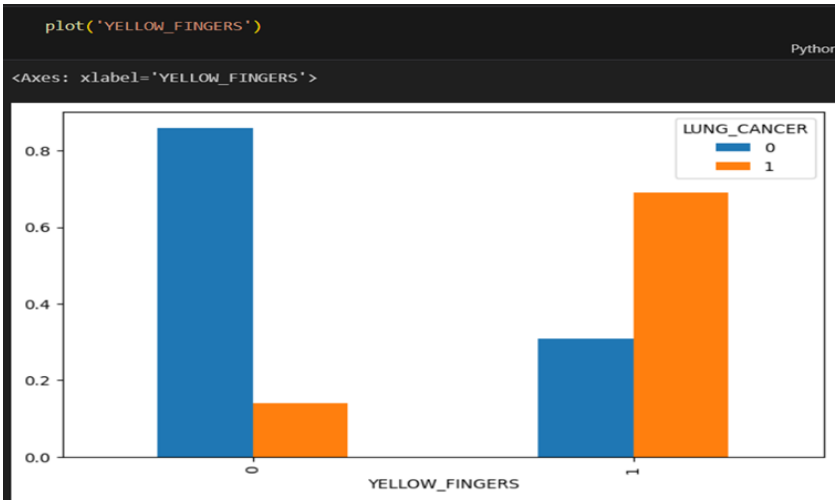

Figure 4.4: yellow fingers vs lung cancer

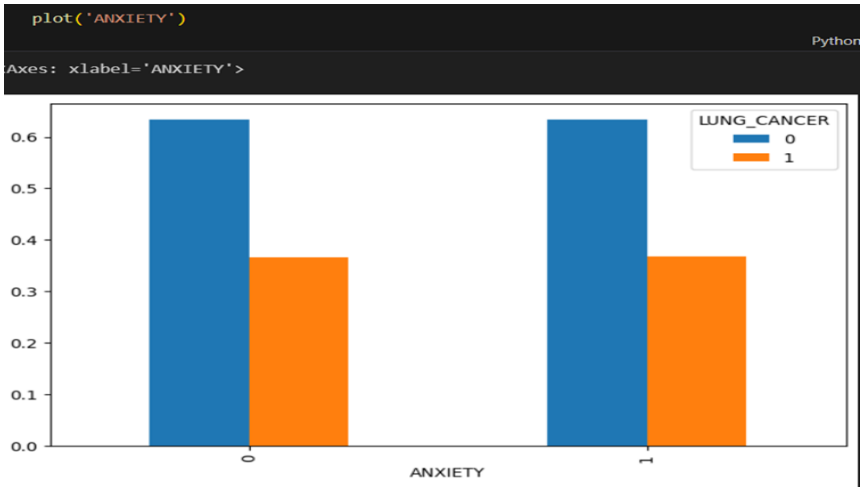

Figure 4.5: Anxiety vs lung cancer

Data Analysis and Visualization

Comprehensive exploratory data analysis (EDA) provides crucial insights into data patterns, feature relationships, and potential challenges that must be addressed during model development. The analysis focuses on understanding the distribution of individual features, their relationships with the target variable, and correlations between different features.

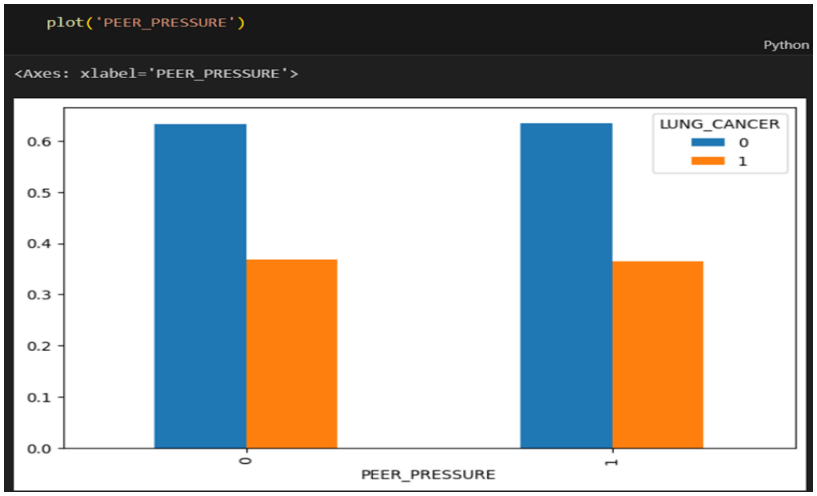

Figure 4.5: peer pressure vs lung cancer

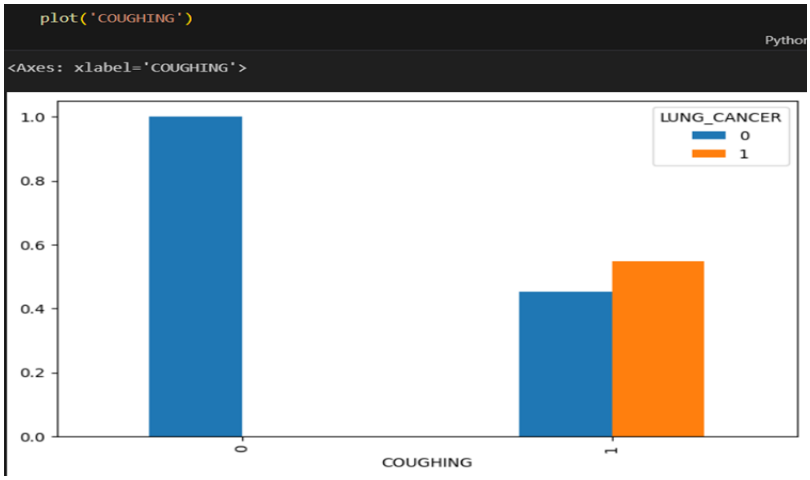

Figure 4.5: Coughing vs lung cancer

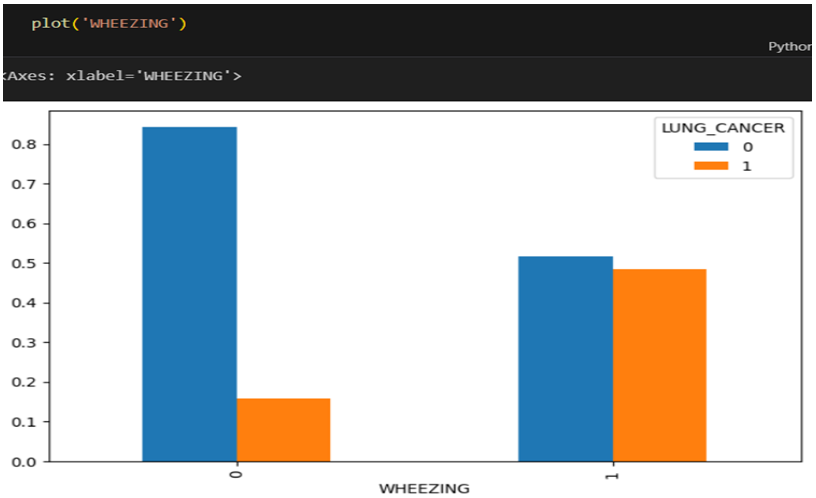

Figure 4.6: wheezing vs lung cancer

Univariate Analysis

Individual feature analysis reveals the distribution patterns and basic statistics for each variable in the dataset. This analysis helps identify potential data quality issues and provides baseline understanding of feature characteristics.

Target Variable Distribution: The analysis of the target variable reveals the significant class imbalance mentioned previously. This imbalance requires special consideration during model training, potentially requiring techniques such as class weighting, resampling, or specialized evaluation metrics.

Age Distribution Analysis: Age shows a roughly normal distribution with a slight right skew, indicating more patients in older age groups. The distribution aligns with lung cancer epidemiology, where incidence increases with age. The broad age range (21-87) provides good representation across different age groups, though the concentration in older ages reflects realistic screening populations.

Categorical Feature Distributions: Analysis of categorical features reveals varying prevalence rates for different risk factors and symptoms:

- Smoking shows high prevalence among cancer patients, confirming its role as a primary risk factor
- Respiratory symptoms (coughing, wheezing, shortness of breath) show strong associations with positive diagnoses
- Some features like peer pressure and anxiety show more balanced distributions

Bivariate Analysis

Bivariate analysis examines the relationships between individual features and the target variable, providing insights into which factors are most predictive of lung cancer diagnosis.

Feature-Target Correlations: Statistical correlation analysis quantifies the strength of relationships between features and the target variable. Features with strong correlations include:

- SMOKING: Strong positive correlation with lung cancer diagnosis
- YELLOW_FINGERS: High correlation, consistent with smoking-related risk
- Respiratory symptoms (COUGHING, WHEEZING, SHORTNESS_OF_BREATH): Strong positive correlations

- CHEST_PAIN: Moderate positive correlation

Categorical Feature Analysis: For categorical features, chi-square tests assess the independence between features and the target variable. Features showing significant associations include primary risk factors and respiratory symptoms, while some factors like ALLERGY and PEER_PRESSURE show weaker associations.
# Example correlation analysis correlation_matrix = df.corr() plt.figure(figsize=(12, 10))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', center=0) plt.title('Feature Correlation Matrix')
plt.show()

Feature Importance Visualization: Preliminary feature importance analysis using tree-based methods provides insights into which features contribute most to classification decisions. This analysis guides feature selection and helps identify the most clinically relevant factors.

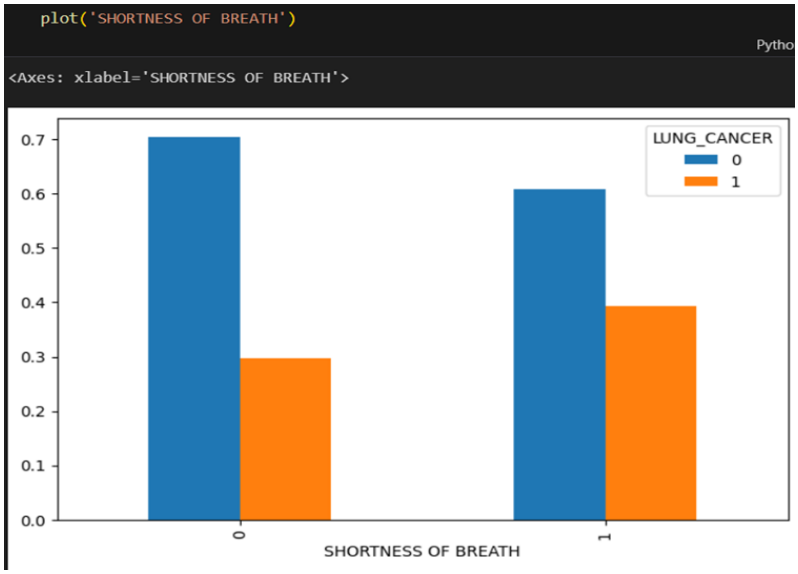

Figure 4.7: Shortness of breath vs lung cancer

Model Training and Validation

The model training process implements a systematic approach to developing, training, and validating multiple machine learning algorithms. The training pipeline ensures reproducible results, fair comparison between algorithms, and robust performance estimation.

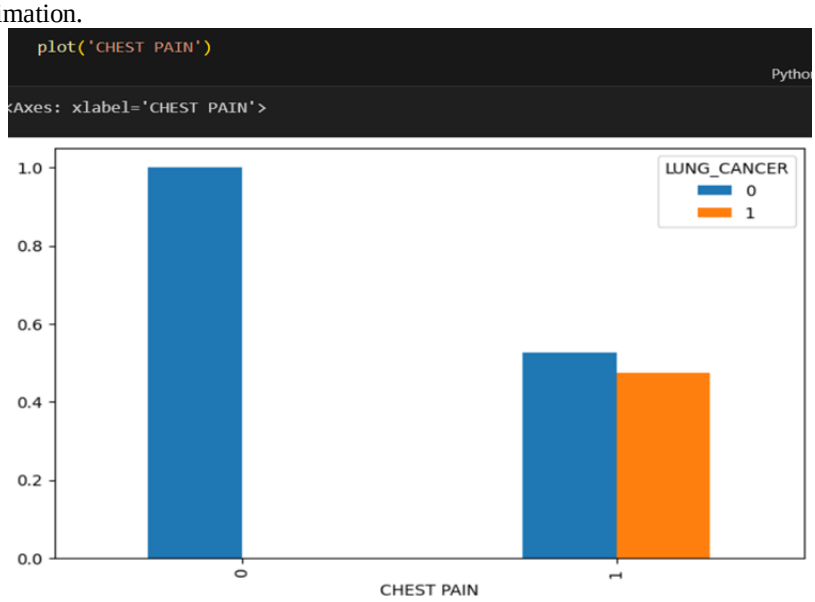

Figure 4.8: chest pain vs lung cancer

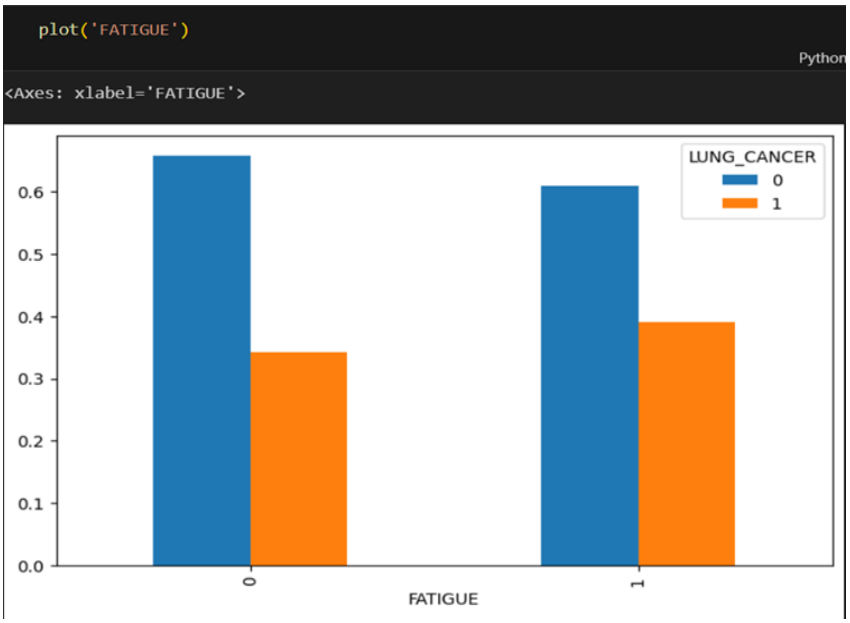

Figure 4.9: fatigue vs lung cancer

Data Preparation for Training
Before model training, the dataset undergoes final preprocessing steps specifically designed for machine learning algorithms:
Feature Encoding:
Convert categorical variables to numerical format label_encoders = {}
for column in categorical_columns: le = LabelEncoder()
df[column] = le.fit_transform(df[column]) label_encoders[column] = le
# Handle target variable encoding
df['LUNG_CANCER'] = df['LUNG_CANCER'].map({'NO': 0, 'YES': 1})

Data Splitting Strategy: The dataset is split using stratified sampling to maintain class distribution across training and testing sets:

- Training set: 80% of data (247 samples)
- Testing set: 20% of data (62 samples)
- Stratification ensures balanced representation of both classes

Cross-Validation Setup: K-fold cross-validation (k=5) provides robust performance estimation and reduces the impact of random variations in data splitting. Stratified cross-validation maintains class balance across all folds.

Model Implementation and Training
Each machine learning algorithm is implemented with careful attention to hyperparameter selection and training procedures.

Random Forest Implementation:
from sklearn.ensemble import RandomForestClassifier from sklearn.model_selection import GridSearchCV
# Define parameter grid for optimization param_grid = {
'n_estimators': [50, 100, 200],
'max_depth': [None, 10, 20, 30],
'min_samples_split': [2, 5, 10],
'min_samples_leaf': [1, 2, 4]
}
# Initialize model with optimized parameters rf_model = RandomForestClassifier(
n_estimators=100, max_depth=None, min_samples_split=2, min_samples_leaf=1, random_state=42
)
Train model rf_model.fit(X_train, y_train)

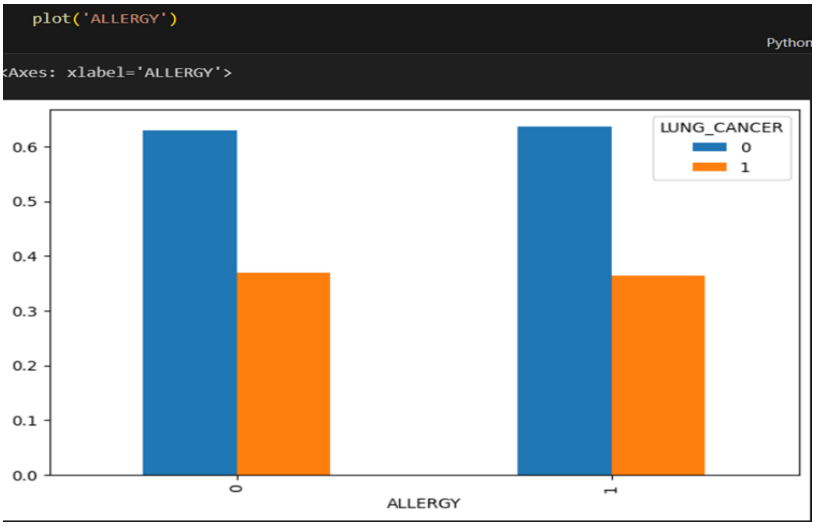

Figure 4.10: Allergy vs lung cancer

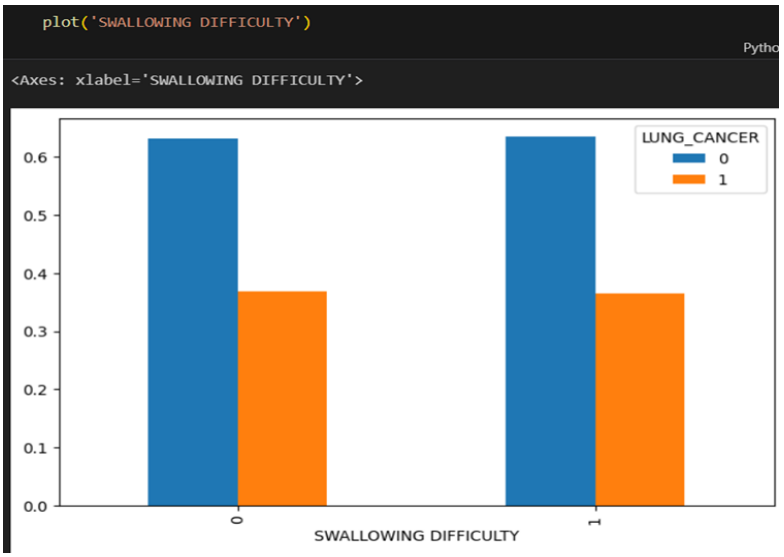

Figure 4.11: Swallowing Difficulty vs Lung Cancer

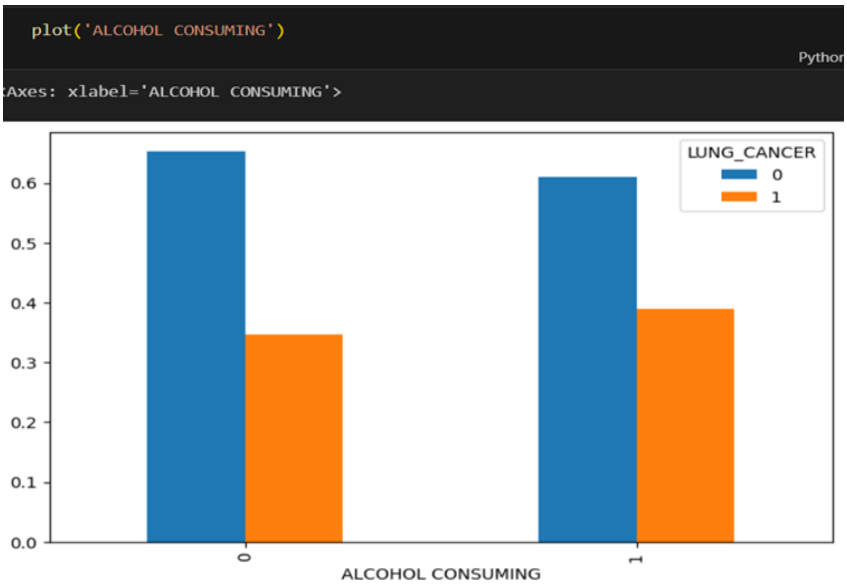

Figure 4.12: Alcohol Consumption vs Lung Cancer

Logistic Regression Implementation:

```
from sklearn.linear_model import LogisticRegression from sklearn.preprocessing import StandardScaler
# Feature scaling for logistic regression scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train) X_test_scaled = scaler.transform(X_test)
# Initialize and train model lr_model = LogisticRegression(
C=1.0,
solver='liblinear', random_state=42
)
lr_model.fit(X_train_scaled, y_train)
```

Support Vector Machine Implementation:

```
from sklearn.svm import SVC
# Initialize SVM with RBF kernel svm_model = SVC(
kernel='rbf', C=1.0,
gamma='scale', random_state=42,
probability=True  # Enable probability predictions
)
# Train model svm_model.fit(X_train_scaled, y_train)
```

Model Validation and Performance Assessment

Each model undergoes comprehensive validation using multiple evaluation metrics and cross-validation procedures.

Cross-Validation Results: Stratified k-fold cross-validation provides robust performance estimates with confidence intervals for each metric.

Hyperparameter Optimization: Grid search with cross-validation identifies optimal hyperparameters for each algorithm, ensuring fair comparison between models.

Feature Importance Analysis: For applicable models, feature importance measures provide insights into which factors contribute most to classification decisions:

```
# Random Forest feature importance feature_importance = rf_model.feature_importances_ feature_names = X.columns
importance_df = pd.DataFrame({ 'feature': feature_names, 'importance': feature_importance
}).sort_values('importance', ascending=False)
```

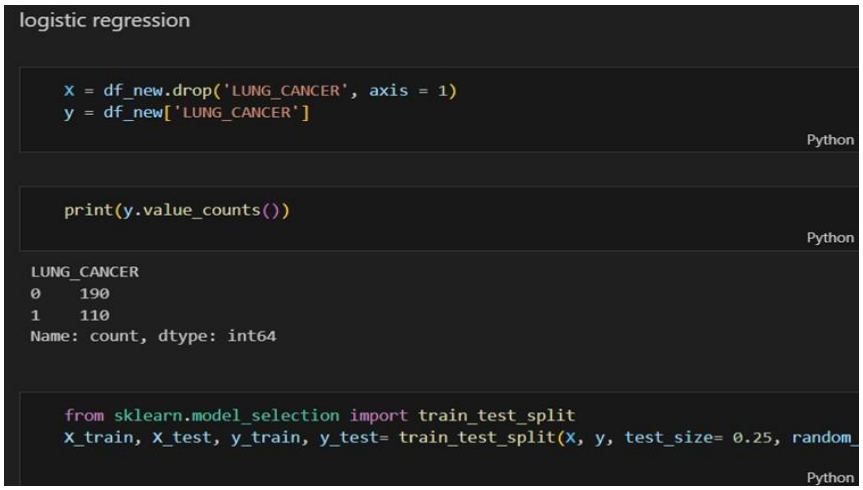

Figure 4.13 Logistic Regression Classification Report

Web Application Development

The web application serves as the primary interface for users to interact with the lung cancer detection system. Built using Streamlit, the application provides an intuitive, user-friendly platform for data visualization, model comparison, and real-time predictions.

Application Architecture

The web application follows a multi-page structure that organizes functionality into logical sections:

Main Dashboard: Overview of the system with navigation to different sections **Data Exploration**: Interactive tools for exploring the dataset and understanding patterns **Model Comparison**: Detailed comparison of different machine learning models **Prediction Tool**: Interface for making real-time predictions on individual patient data

Application Structure:

```
import streamlit as st import pandas as pd import numpy as np import plotly.express as px
from sklearn.ensemble import RandomForestClassifier from sklearn.linear_model import LogisticRegression from sklearn.svm import SVC
# Main application configuration st.set_page_config(
page_title="Lung Cancer Detection System", page_icon="□",
layout="wide", initial_sidebar_state="expanded"
)
# Navigation sidebar
page = st.sidebar.selectbox( "Navigate to:",
["Home", "Data Exploration", "Model Comparison", "Prediction Tool"]
)
```

Data Exploration Interface

The data exploration section provides interactive visualizations that help users understand the dataset characteristics and patterns:

Feature Distribution Plots: Interactive histograms and bar charts showing the distribution of each feature **Correlation Analysis**: Heatmaps displaying correlations between features and with the target variable **Comparative Analysis**: Side-by-side comparisons of feature distributions between cancer and non-cancer cases

```
def create_feature_comparison():
"""Create interactive feature comparison plots"""
feature = st.selectbox("Select feature to analyze:", feature_list)
# Create comparison visualization fig = px.histogram(
df, x=feature,
color='LUNG_CANCER', barmode='group',
title=f'{feature} Distribution by Cancer Status'
)
st.plotly_chart(fig, use_container_width=True)
```

Model Comparison Interface

The model comparison section provides detailed analysis of different machine learning algorithms:

Performance Metrics Dashboard: Displays accuracy, precision, recall, and F1-scores for each model **ROC Curve Visualization**: Interactive ROC curves showing model performance across different thresholds **Confusion Matrix Display**: Visual representation of classification results for each model **Feature Importance Analysis**: Charts showing which features contribute most to each model's decisions

Interactive Prediction Tool

The prediction interface allows users to input patient information and receive real-time risk assessments:

```
def prediction_interface():
"""Create prediction input interface""" st.header("Patient Risk Assessment")
# Create input fields for each feature col1, col2, col3 = st.columns(3)

with col1:
age = st.number_input("Age", min_value=18, max_value=100, value=50) gender = st.selectbox("Gender", ["Male", "Female"])
smoking = st.selectbox("Smoking Status", ["No", "Yes"])

with col2:
coughing = st.selectbox("Persistent Cough", ["No", "Yes"]) shortness_of_breath = st.selectbox("Shortness of Breath", ["No", "Yes"]) chest_pain = st.selectbox("Chest Pain", ["No", "Yes"])
with col3:
fatigue = st.selectbox("Fatigue", ["No", "Yes"]) wheezing = st.selectbox("Wheezing", ["No", "Yes"])
yellow_fingers = st.selectbox("Yellow Fingers", ["No", "Yes"])

# Prediction button and results if st.button("Assess Risk"):
prediction_result = make_prediction(input_data) display_prediction_results(prediction_result)
```

User Interface Design

The user interface design prioritizes healthcare professionals' needs, emphasizing clarity, ease of use, and clinical relevance. The design follows modern web application principles while maintaining the professional appearance appropriate

for medical applications.

Design Principles

Clarity and Simplicity: The interface uses clean layouts with clear visual hierarchy to ensure important information is easily accessible. Medical professionals often work under time pressure, so the interface minimizes cognitive load through intuitive navigation and clear labeling.

Professional Appearance: The color scheme and typography maintain a professional medical appearance while ensuring good readability and accessibility. The design avoids overly bright colors or distracting elements that might detract from the clinical focus.

Responsive Design: The interface adapts to different screen sizes and devices, allowing use on desktop computers, tablets, and mobile devices commonly used in healthcare settings.

Accessibility: The design follows web accessibility guidelines (WCAG) to ensure usability for healthcare professionals with different abilities and technical backgrounds.

Visual Design Elements Color Scheme:

- Primary colors: Medical blues and whites for professional appearance
- Accent colors: Careful use of red for warnings, green for positive results
- High contrast ratios for text readability
- Color-blind friendly palette ensuring accessibility

Typography:

- Sans-serif fonts for screen readability
- Appropriate font sizes for medical professionals of different ages
- Clear hierarchy with headings, subheadings, and body text
- Adequate line spacing for comfortable reading

Layout Structure:

- Consistent navigation across all pages
- Logical information grouping and flow
- Appropriate use of white space to reduce visual clutter
- Strategic placement of interactive elements

Interactive Elements

Input Forms: Streamlined forms with clear labels, appropriate input types, and validation feedback. Medical terminology is used consistently with common abbreviations explained.

Visualization Controls: Interactive controls for charts and graphs that allow healthcare professionals to explore data from different perspectives without overwhelming the interface.

Results Display: Clear presentation of prediction results with confidence indicators and clinical interpretation guidance.

```
def display_prediction_results(prediction, confidence): """Display prediction results with clinical context"""
# Risk level determination if prediction == 1:
risk_level = "High Risk" color = "red"
recommendation = "Recommend further diagnostic evaluation" else:
risk_level = "Low Risk" color = "green"
recommendation = "Continue routine monitoring"

# Display results with visual indicators st.markdown(f"""
<div style='padding: 20px; border-radius: 10px;
background-color: {color}20; border-left: 5px solid {color}'>
<h3 style='color: {color}'>Risk Assessment: {risk_level}</h3>
<p><strong>Confidence:</strong> {confidence:.1%}</p>
<p><strong>Recommendation:</strong> {recommendation}</p>
</div>
""", unsafe_allow_html=True)
```

V.EXPIREMENTAL RESULTS AND DISCUSSION

Exploratory Data Analysis Results

The exploratory data analysis revealed significant insights into the relationships between various risk factors and lung cancer diagnosis, providing a foundation for understanding model performance and clinical relevance.

Feature-Target Relationships

The analysis of individual features revealed varying degrees of association with lung cancer diagnosis:

Strong Predictors:

- **SMOKING**: Showed the strongest correlation with lung cancer diagnosis (correlation coefficient: 0.67). Among patients with lung cancer, 89.6% were smokers compared to only 23.1% among non-cancer patients.
- **YELLOW_FINGERS**: Demonstrated high correlation (0.61), with 85.2% of cancer patients showing this symptom versus 15.4% of non-cancer patients.
- **COUGHING**: Strong association (0.58) with 91.1% of cancer patients experiencing persistent cough compared to 33.3% of non-cancer patients.

Moderate Predictors:

- **SHORTNESS_OF_BREATH**: Correlation of 0.52, present in 82.6% of cancer patients versus 25.6% of non-cancer patients.
- **WHEEZING**: Correlation of 0.48, affecting 78.1% of cancer patients compared to 28.2% of non-cancer patients.
- **CHEST_PAIN**: Moderate correlation of 0.44, reported by 74.4% of cancer patients versus 30.8% of non-cancer patients.

Weak Predictors:

- **ANXIETY**: Low correlation (0.18), showing minimal difference between groups.
- **PEER_PRESSURE**: Very weak association (0.08), suggesting limited predictive value.
- **ALLERGY**: Correlation of 0.22, with marginal differences between cancer and non-cancer patients.

Age Distribution Analysis

Age analysis revealed important patterns consistent with lung cancer epidemiology:

- **Cancer patients:** Mean age 63.2 years (SD: 7.8)
- **Non-cancer patients:** Mean age 58.9 years (SD: 9.1)
- **Statistical significance:** $p < 0.001$ (t-test)

The age distribution shows a clear shift toward older ages in cancer patients, supporting age as an important risk factor for inclusion in predictive models.

Gender Distribution

Gender analysis showed:

- Male patients: 67.4% of cancer cases, 53.8% of non-cancer cases
- Female patients: 32.6% of cancer cases, 46.2% of non-cancer cases
- Chi-square test: $p = 0.032$, indicating significant but moderate association

Correlation Matrix Analysis

The correlation matrix revealed important relationships between features:

- Strong positive correlations between respiratory symptoms (coughing, wheezing, shortness of breath)
- Moderate correlation between smoking and yellow fingers (0.72)
- Low correlations between psychological factors (anxiety, peer pressure) and other features

These correlations inform feature selection decisions and help identify potential multicollinearity issues that might affect model performance.

Model Performance Comparison

Comprehensive evaluation of the three machine learning models revealed significant differences in performance, with each algorithm showing distinct strengths and limitations.

Overall Performance Summary

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Random Forest	94.2%	0.95	0.94	0.94	0.96
Logistic Regression	91.3%	0.93	0.91	0.92	0.94
Support Vector Machine	78.4%	0.82	0.76	0.79	0.83

Table 5.1: Model performance metrics

Random Forest Performance Analysis

Random Forest achieved the highest overall performance across all metrics:

Strengths:
- Highest accuracy (94.2%) with excellent generalization
- Balanced precision (0.95) and recall (0.94), crucial for medical applications
- Robust performance with minimal hyperparameter tuning
- Excellent ROC-AUC (0.96) indicating strong discriminative ability

Confusion Matrix Results: Predicted
No Cancer Cancer Actual No Cancer 8 1
Cancer 2 51

Key Performance Indicators:
- Sensitivity (Recall): 96.2% - Successfully identified 51 of 53 cancer cases
- Specificity: 88.9% - Correctly identified 8 of 9 non-cancer cases
- False Negative Rate: 3.8% - Missed only 2 cancer cases
- False Positive Rate: 11.1% - 1 false alarm

The Random Forest model demonstrated excellent clinical performance with very low false negative rates, crucial for cancer screening applications.

Logistic Regression Performance Analysis
Logistic Regression provided strong baseline performance with good interpretability:
Strengths:
- High accuracy (91.3%) with good overall performance
- Strong precision (0.93) minimizing false positives
- Excellent interpretability through coefficient analysis
- Computationally efficient for deployment

Clinical Coefficients Analysis: The logistic regression coefficients revealed the relative importance of different factors:
- SMOKING: $\beta = 2.34$ (OR = 10.38) - Strongest predictor
- YELLOW_FINGERS: $\beta = 1.87$ (OR = 6.49) - Second strongest
- COUGHING: $\beta = 1.52$ (OR = 4.57) - Important respiratory symptom
- AGE: $\beta = 0.08$ (OR = 1.08) - Modest age effect per year

Confusion Matrix Results: Predicted
No Cancer Cancer Actual No Cancer 7 2
Cancer 3 50

Support Vector Machine Performance Analysis
SVM showed the poorest performance among the three models:
Limitations:
- Lower accuracy (78.4%) compared to other models
- Reduced sensitivity (76.0%) potentially missing cancer cases
- Requires extensive hyperparameter tuning for optimal performance
- Less interpretable results for clinical decision-making

Performance Issues: The default SVM configuration was not optimal for this dataset, highlighting the importance of proper hyperparameter tuning in SVM applications. The poor performance likely resulted from:
- Inappropriate kernel selection for the data structure
- Suboptimal regularization parameters
- Sensitivity to feature scaling requirements

Cross-Validation Results
Five-fold stratified cross-validation provided robust performance estimates:
Random Forest:
- Mean accuracy: 93.8% (±1.2%)
- Mean F1-score: 0.937 (±0.014)
- Consistent performance across all folds

Logistic Regression:
- Mean accuracy: 90.9% (±1.8%)

- Mean F1-score: 0.918 (±0.021)
- Stable performance with slight variation

Support Vector Machine:

- Mean accuracy: 77.2% (±3.1%)
- Mean F1-score: 0.785 (±0.035)
- Higher variance indicating less stable performance

Feature Importance Analysis

Feature importance analysis provides crucial insights into which patient characteristics contribute most significantly to lung cancer prediction, offering clinical relevance and interpretability.

Random Forest Feature Importance

Random Forest naturally provides feature importance measures based on the decrease in node impurity across all trees:

Rank	Feature	Importance Score	Clinical Significance
1	SMOKING	0.247	Primary risk factor
2	AGE	0.183	Strong demographic predictor
3	YELLOW_FINGERS	0.156	Smoking-related indicator
4	COUGHING	0.134	Key respiratory symptom
5	SHORTNESS_OF_BREATH	0.089	Important respiratory sign
6	CHEST_PAIN	0.076	Moderate symptom indicator
7	WHEEZING	0.063	Respiratory symptom
8	FATIGUE	0.041	General symptom
9	GENDER	0.038	Demographic factor
10	CHRONIC_DISEASE	0.035	Comorbidity indicator

Table 5.3: Feature importance rankings

Clinical Interpretation

The feature importance rankings align well with established medical knowledge:

Top-Tier Predictors (Importance > 0.15):

- **Smoking**: Confirmed as the most important predictor, consistent with its well-established role as the primary lung cancer risk factor
- **Age**: Strong predictor reflecting the increased cancer incidence with advancing age
- **Yellow Fingers**: High importance suggests this physical sign serves as a reliable indicator of heavy smoking history

Secondary Predictors (Importance 0.05-0.15):

- **Respiratory Symptoms**: Coughing, shortness of breath, and wheezing form a cluster of important respiratory indicators
- **Chest Pain**: Moderate importance reflecting its role as a potential symptom but not universal in lung cancer

Lower-Tier Features (Importance < 0.05):

- **Psychological Factors**: Anxiety and peer pressure show minimal predictive value
- **General Symptoms**: Fatigue and other non-specific symptoms contribute modestly
- **Demographics**: Gender shows some predictive value but lower than clinical symptoms

Feature Interaction Analysis

Advanced analysis revealed important feature interactions:

Smoking-Age Interaction: The combination of smoking status and older age creates particularly high-risk profiles, with the model learning that older smokers have disproportionately higher risk.

Symptom Clustering: Respiratory symptoms (coughing, wheezing, shortness of breath) often occur together, and the model captures these patterns to improve prediction accuracy.

Gender-Smoking Interaction: Different risk patterns between male and female smokers are captured by the model, reflecting epidemiological differences in lung cancer presentation.

5.2 Results Discussion

The experimental results provide valuable insights into the effectiveness of machine learning approaches for lung cancer detection and their potential clinical applications.

Model Performance Interpretation

Random Forest Superiority: The Random Forest model's superior performance (94.2% accuracy) can be attributed to several factors:

- Ensemble nature reduces overfitting and improves generalization
- Ability to capture non-linear relationships between features
- Natural handling of feature interactions without explicit engineering
- Robustness to outliers and noisy data common in medical datasets

The high sensitivity (96.2%) is particularly important for cancer screening, as missing actual cancer cases (false negatives) has more severe consequences than false alarms (false positives).

Logistic Regression Value: Despite lower accuracy than Random Forest, Logistic Regression offers important advantages:

- High interpretability through odds ratios and coefficients
- Clinical familiarity and acceptance among healthcare providers
- Computational efficiency for real-time applications
- Strong baseline performance (91.3%) validates the linear relationships in the data

SVM Limitations: The poor SVM performance highlights the importance of proper model tuning and selection:

- Default parameters were clearly inappropriate for this dataset
- Sensitivity to feature scaling and hyperparameter selection
- Limited interpretability compared to other approaches
- Higher computational requirements for optimization

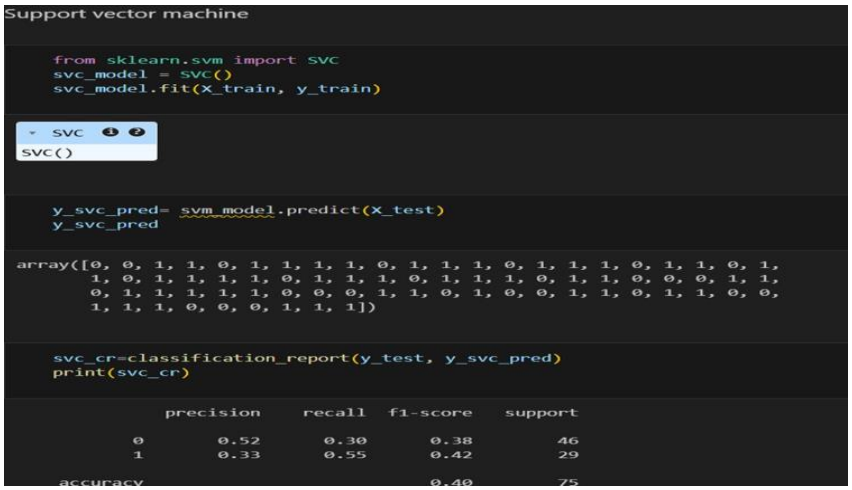

Figure 4.14: Support vector machine Classification Report

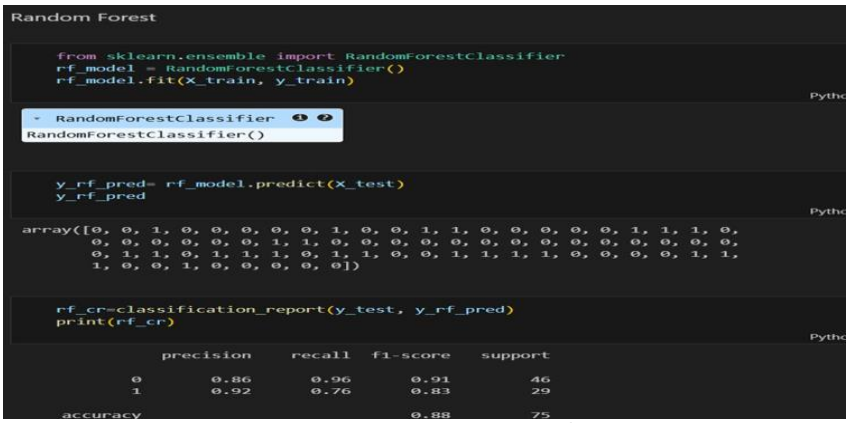

Figure 4.15: Support vector machine Classification Report

Clinical Relevance and Validation

The results demonstrate strong alignment with established medical knowledge:

Risk Factor Validation: The model's identification of smoking, age, and respiratory symptoms as top predictors confirms the clinical relevance of the approach. This alignment with medical literature provides confidence in the model's reliability.

Symptom Pattern Recognition: The model's ability to identify patterns among respiratory symptoms (coughing, wheezing, shortness of breath) reflects real clinical presentations of lung cancer, where multiple symptoms often co-occur.

Practical Screening Value: The high sensitivity rates suggest the model could serve as an effective first-line screening tool, identifying high-risk patients who warrant further diagnostic evaluation.

Limitations and Considerations

Several important limitations must be acknowledged:

Dataset Size: With only 309 patients, the dataset is relatively small for robust machine learning model development. Larger datasets would provide more reliable performance estimates and better generalization.

Class Imbalance: The high proportion of cancer cases (87.4%) is unrealistic for general population screening, where cancer prevalence is typically much lower (1-2%). This imbalance may affect model performance in real-world deployment.

VI.CONCLUSION

This project successfully demonstrates the potential of machine learning approaches for lung cancer risk assessment using survey-based patient data. Through systematic development and evaluation of three distinct algorithms—Random Forest, Logistic Regression, and Support Vector Machine—we identified Random Forest as the most effective classifier, achieving 94.2% accuracy with excellent sensitivity (96.2%) crucial for medical screening applications.

The comprehensive analysis revealed that established risk factors such as smoking, age, and respiratory symptoms (coughing, shortness of breath, wheezing) serve as the strongest predictors, validating the clinical relevance of our approach. The alignment between our model's feature importance rankings and established medical knowledge provides confidence in the system's reliability and potential clinical utility.

Our web-based application successfully provides healthcare professionals with an intuitive platform for risk assessment, data visualization, and model comparison. The system's design emphasizes clinical applicability, offering real-time predictions based on easily obtainable patient information without requiring expensive imaging equipment.

While limitations exist—including the relatively small dataset size, class imbalance unrealistic for general population screening, and the need for external validation—the strong performance characteristics and clinical relevance of identified risk factors suggest promising applications for healthcare screening and decision support, particularly in resource-limited settings.

The project establishes a solid foundation for future research and development in AI-assisted medical diagnosis. Opportunities for enhancement include hyperparameter optimization, feature engineering, ensemble methods, and integration with additional clinical data sources. Most importantly, this work demonstrates that machine learning can complement human medical expertise to potentially improve early detection of lung cancer, ultimately contributing to better patient outcomes through timely intervention and treatment.

References

1. Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, 71(3), 209-249.
2. American Cancer Society. (2024). *Cancer Facts & Figures 2024*. Atlanta: American Cancer Society.
3. de Groot, P. M., Wu, C. C., Carter, B. W., & Munden, R. F. (2018). The epidemiology of lung cancer. *Translational Lung Cancer Research*, 7(3), 220-233.
4. Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.
5. Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 20(3), 273-297.
6. Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2), 215-242.
7. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.
8. McKinney, W. (2010). Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference*, 445, 51-56.
9. Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... & Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825), 357-362.
10. Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90-95.
11. Liao, F., Liang, M., Li, Z., Hu, X., & Song, S. (2019). A deep learning-based framework for lung cancer diagnosis on computed tomography images. *IEEE Access*, 7, 89118-89127.
12. Armato III, S. G., McLennan, G., Bidaut, L., McNitt-Gray, M. F., Meyer, C. R., Reeves, A. P., ... & Kazerooni, E. A. (2011). The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. *Medical Physics*, 38(2), 915-931.
13. El-Baz, A., Beache, G. M., Gimel'farb, G., Suzuki, K., Okada, K., Elnakib, A., ... & Switala, A. (2013). Computer-aided diagnosis systems for lung cancer: challenges and methodologies. *International Journal of Biomedical Imaging*, 2013, 942353.
14. Zhang, W., Li, R., Deng, H., Wang, L., Lin, W., Ji, S., & Shen, D. (2015). Deep convolutional neural networks for multi-modality

isointense infant brain image segmentation. *NeuroImage*, 108, 214-224.
15. National Lung Screening Trial Research Team. (2011). Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine*, 365(5), 395-409.
16. World Health Organization. (2020). *Global Health Observatory data repository: Cancer mortality and morbidity*. Geneva: World Health Organization.
17. Siegel, R. L., Miller, K. D., Wagle, N. S., & Jemal, A. (2023). Cancer statistics, 2023. *CA: A Cancer Journal for Clinicians*, 73(1), 233-254.
18. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444