LUNG CANCER DIAGNOSIS AND PREDICTION USING AIML

A Project Report (Project-I) submitted in partial fulfillment of the requirements for the award of degree of

BACHELOR OF TECHNOLOGY
IN
INFORMATION TECHNOLOGY

Supervised to
Ms. Urvinder Kaur
Assistant Professor
IT Department

Submitted by
Suraj (2823127)
Shweta (28231276)
Neha (2822784)
Abhishek (2822783)
VIth Sem.

DEPARTMENT OF INFORMATION TECHNOLOGY

PANIPAT INSTITUTE OF ENGINEERING AND TECHNOLOGY SAMALKHA, PANIPAT-132103
(Approved by AICTE and Affiliated to the Kurukshetra University, Kurukshetra)
MAY 2025

DECLARATION

We certify that

1. The work presented in this project report is an authentic record of our own work under the guidance of our supervisor. It has not been submitted to any other Institute for the award of any other degree or diploma.
2. Whenever we have used information (text, data, figure, photograph, chart, analysis, inference, etc.) from other sources, we have given due credit by citing it in the text of the report and providing its details in the references.
3. We have followed the guidelines provided by the department for preparing the report.

Suraj Yadav (28231275)

Shweta Kumari (28231276)

Neha Kumari (2822784)

Abhishek Kumar Muarya (2822783)

Project report title:- Lung Cancer Diagnosis And Prediction using AIML

Semester:- VIth Sem.

Date: 13TH MAY,2025

APPROVAL FROM SUPERVISOR

This is to certify that the project report entitled "*Lung Cancer diagnosis And Prediction using AI/ML*" presented by *"Suraj Yadav (28231275), Shweta Kumari (28231276), Abhishek Kumar Muarya (2822783), Neha Kumari (2822784)"* under my supervision is an authentic work. To the best of my knowledge, the content of this report has not been submitted for the award of any previous degree to anyone else.

It is recommended that the report be accepted as fulfilling this part of the requirements for the award of the degree.

Ms. Urvinder Kaur
Assistant Professor
Department of Information Technology
Date: 22nd MAY,2025

Dr. Neeraj Gupta
Head, Department of Information Technology

CERTIFICATE

This is to certify that the work embodied in this report, entitled "*Lung Cancer Diagnosis And Prediction*" carried out by " *Suraj Yadav (28231275), Shweta Kumari (28231276), Abhishek Kumar Muarya (2822783), Neha Kumari (2822784)"* " is approved for the degree of "*B.tech*" at the department of " *Information Technology*", Panipat Institute of Engineering and Technology, Samalkha.

____

Internal Examiner

____

External Examiner

Date:

Place: Panipat

ACKNOWLEDGEMENTS

It is an overwhelming pleasure and extreme sense of fulfillment for us to submit the concluding report of our project entitled "Lung Cancer Diagnosis and Prediction Using AIML " as part of our B.Tech. syllabus under Information Technology in Pre-Final Year at Panipat Institute of Engineering and Technology (PIET), Samalkha.

We would like to convey our sincere appreciation to **Ms. Urvinder Kaur,** Assistant Professor, Department of Information Technology, PIET, for his precious advice, continuous encouragement, and penetrating suggestions during the course of this project. His technical knowledge, zeal for perfection, and considerate guidance allowed us to face numerous challenges, particularly in grasping the complexity of machine learning algorithms and their practical application in the healthcare field.

We are highly indebted to **Dr. Neeraj Gupta**, Head of Department, Information Technology, PIET, for giving us a favorable academic atmosphere, required resources, and timely guidance that significantly helped us in the successful completion of our project.

We are thankful to all faculty members of the department for their kind assistance, guidance, and cooperation during the development of our project.

Lastly, we are profoundly grateful to our friends and peers, who have provided constant moral support, shared ideas, and helped troubleshoot issues during the development of this project. Their collaboration and constructive feedback added immense value to our work.

This project not only enhanced our technical skills in machine learning and healthcare analytics but also helped us develop our research, problem-solving, and collaboration skills, which will benefit us a lot in the future.

Suraj Yadav (28231275)
Shweta Kumari (28231276)
Neha Kumari (2822784)
Abhishek Kumar Muarya (2822783)

Date:

ABSTRACT

Lung Cancer Prediction and Diagnosis Using AI/ML (Random Forest Algorithm) is a comprehensive, data-oriented health analytics application that utilizes machine learning methods to aid in early lung cancer detection through organized numerical data. Designed solely on true-world, clinically formatted datasets stored in.csv format, the project seeks to update diagnostic procedures by streamlining the prediction task using ensemble-based classification models—the Random Forest algorithm.

The platform is developed with the emphasis on predictive performance, explainability, and clinical relevance. It starts with an end-to-end data preprocessing pipeline to take care of missing values, removal of outliers, feature encoding, normalization, and dimensionality reduction, thus preparing the raw numerical dataset for high-efficiency learning. Once data are prepared, the Random Forest algorithm—a set of decision trees—is trained to classify patient records as malignant or benign based on an extensive set of biomedical attributes.

The primary goals of this solution are better prediction accuracy, lesser false negatives, and giving medical practitioners data-driven decision support. The model is tested on a variety of performance measures such as accuracy, precision, recall, F1-score to ensure that the diagnostic result is both clinically meaningful and reliable.

The project is developed with a strong collection of Python libraries like Scikit-learn, Pandas, NumPy, Matplotlib, and Seaborn, which facilitate reproducible experimentation, data visualization, and result interpretation. The Random Forest classifier, which is resistant to overfitting and can handle high-dimensional numerical data, makes it a perfect fit for healthcare applications where interpretability of prediction and generalization are critical.

1. Major features of the solution are:

2. End-to-end automated lung cancer classification from CSV-based clinical data.

3. AI-powered learning through a Random Forest ensemble to improve accuracy and generalizability.

4. Visual inspection of feature importance to comprehend the diagnostic influence of multiple biomedical parameters.

5. Distributed and extensible design fit for integration with electronic health records or clinical decision support systems.

Keywords: Lung Cancer, Machine Learning, Random Forest, Medical Diagnosis, Predictive Modeling, CSV Dataset, Numerical Data, Data Preprocessing, Feature Selection, AI in Healthcare, Ensemble Learning, Classification Algorithm, Model Evaluation, Biomedical Data, Clinical Decision Support.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1: Introduction

Lung cancer is one of the most critical health issues worldwide, ranking as the most common cause of cancer deaths. One of the most significant factors contributing to its high fatality rate is the absence of early and precise detection. Traditional diagnostic techniques, though effective, tend to be time-consuming, costly, and may not always yield prompt results. With the developing technology, Artificial Intelligence (AI) and Machine Learning (ML) provide potential tools for early disease diagnosis and prediction, including lung cancer. In this project, the Random Forest algorithm, a supervised machine learning method, is used to design a system that can be used for diagnosing and predicting lung cancer from clinical and diagnostic information.

1. Purpose

The medical sector is quickly embracing AI and ML methodologies to improve the accuracy of diagnostics, minimize the risk of human error, and assist physicians with clinical decision-making. Among a wide range of machine learning algorithms, Random Forest stands out owing to its ensemble-based methodology that boosts prediction precision and easily handles big data sets. By leveraging clinical data and features associated with lung cancer, a machine learning model can be trained to identify patterns and predict whether a case is likely to be malignant or benign.

2. Motivation

Early diagnosis of lung cancer significantly improves treatment success and patient survival rates. However, due to the asymptomatic nature of the disease in its early stages, many cases go undetected. This project is prompted by the urgency of creating smart, computer-based diagnostic tools to scan through patients' information and offer quality forecasts in time. Applying the aid of AI/ML solutions like the Random Forest algorithm may help medical doctors with the advantage of an auxiliary opinion and minimization of delay times for diagnostics.

3. Objective of the Project

Early diagnosis and precise classification of lung cancer can effectively boost survival chances and enhance cure rates. Conventionally used detection techniques are sometimes labor-intensive, time-consuming, and susceptible to human error. The project endeavours to improve the effectiveness, precision, and automation in lung cancer detection through AI and machine learning concepts—namely the Random Forest algorithm—being applied to data in structured numeric form in the CSV format. The primary objectives of the project are:

▪ Automated Diagnosis System: Design a machine learning-based system that effectively predicts the presence of lung cancer from clinical parameters received in numerical values.

▪ Data-Driven Decision Making: Use real-world, structured datasets (CSV files) to train and validate the model, enabling evidence-based prediction and diagnosis.

▪ Robust Feature Engineering: Apply preprocessing steps like data cleaning, normalization, and feature selection to optimize model learning and dependability.

▪ Random Forest Classifier Integration: Tap into the ensemble learning power with Random Forest to enhance classification performance, mitigate overfitting, and yield interpretable results.

▪ High Evaluation Standards: Assess model performance on important metrics including accuracy, precision, recall, F1-score, and ROC-AUC for clinical applicability.

▪ Scalability and Adaptability: Make the system flexible for future development like real-time diagnosis, integration with healthcare databases, or implementation in clinical decision support systems.

1. Problem Statement

Although technology continues to improve diagnostics through medical images, lung cancer too frequently does not get a diagnosis until stages beyond curative reach. This results in minimal treatment options and a poor outlook. Current diagnosis processes are not just invasive, but potentially even inaccessible to some patients because they cost too much and are hard to find. For this reason, there is a critical need for a highly accurate, non-invasive data-driven method to forecast lung cancer. This project hopes to remedy this by building a machine-learning lung cancer forecasting model.

1.1.5 Study Objective

The main objective of this research is to develop and apply a machine learning model, employing the Random Forest algorithm, for lung cancer diagnosis and prediction. The model should be able to examine pertinent patient information and correctly classify cases as benign or malignant.

1.1.6 Research Aims

1. To conduct and review current literature on AI/ML in lung cancer diagnosis.
2. To gather and preprocess an appropriate lung cancer dataset by managing missing values, feature selection, and normalization.
3. To use the Random Forest algorithm to train and test the prediction model.
4. To measure the performance of the model using metrics like accuracy, precision, recall, and F1-score.
5. To show how AI/ML can be used to help healthcare professionals make early and accurate diagnoses of lung cancer.
6. To present the possible limitations of the model and suggest future improvements.

Chapter 2: Literature Review

2.1 Overview

Artificial Intelligence and Machine Learning have increased prominence in health care over recent years, mainly in disease detection and forecasting. Lung cancer, which is the most deadly form of cancer, has been extensively studied using AI-based diagnostic investigations. This chapter discusses major investigations and methodologies to predict lung cancer using numerical data sets and adopting machine learning practices with emphasis placed on the Random Forest algorithm.

2.2 AI/ML Overview in Medical Diagnosis

Artificial Intelligence and Machine Learning have indicated promising outcomes in medical diagnosis, particularly in structured numerical data analysis gathered from patients. Machine learning algorithms can determine intricate patterns within clinical datasets that might not be readily apparent with conventional analysis. Classification, clustering, and regression have been common techniques used in diagnostic applications [1].

Specifically, quantitative data sets of CSV format made up of traits like age, smoking status, genetic conditions, and tumor nature are ideal to train machine learning models. Research makes use of these data sets because they are easy to work with, they scale well, and they preprocess easily.

Research has proven that ML algorithms such as Decision Trees, Random Forests, Support Vector Machines (SVM), and Neural Networks can yield accurate predictions using patient data. Of these, Random Forest is particularly noted for its accuracy, robustness to overfitting, and capacity to work with datasets containing many variables and non-linear relationships.

2.3 Machine Learning in Medical Diagnosis

Machine learning methods are now more and more indispensable in contemporary healthcare, particularly for diagnosis from large clinical and biomedical datasets. In diseases such as lung cancer, where early detection significantly enhances survival, predictive models provide a scalable and reliable alternative to human diagnosis.

2.3.1 Literature on AI in Lung Cancer Diagnosis

Krishnaiah et al. (2013) – "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques"

This study ventured into the utilization of classification techniques like Naïve Bayes and Decision Trees to predict whether lung cancer existed.
Major Gaps Identified: Few features of the dataset; no use of ensemble techniques like Random Forests.

Lynch et al. (2017) – "Prediction of Lung Cancer Patient Survival via Supervised Machine Learning Classification Techniques"

Applied several machine learning techniques to forecast patient outcomes from diagnostic and demographic information.

Major Gaps Identified: Survival rate forecasting, rather than early detection; also did not include preprocessing methods for structured numeric data.

Fenwa et al. (2016) – "Classification of Cancer of the Lungs using SVM and ANN"

Examined the comparative performance of SVM and ANN models on cancer classification problems.
Key Gaps Identified: Lack of ensemble methods and interpretability; preprocessing steps of data not well defined.

2.4 Technological Underpinnings: Random Forest and Numerical Data Modeling

Random Forest algorithm is a robust, ensemble learning supervised method very suitable for managing high-dimensional noisy or missing clinical data. Its decision-tree-based structure offers enhanced accuracy and resistance to overfitting.

2.4.1 Random Forest Literature Assessment in Biomedical Applications

Daoud and Mayo (2019) – "A Survey of Neural Network-Based Cancer Prediction Models"
Spoke about the dominance of deep learning and ensemble approaches in clinical prediction problems, with Random Forests tending to perform better in structured data.
Gaps Identified: Greater focus on microarray and image data, lesser on clean numerical data.

Palani and Venkatalakshmi (2019) – "An IoT-Based Predictive Modelling for Predicting Lung Cancer

This paper used a hybrid fuzzy clustering method but indicated the Random Forest's better performance for tabular, clinical data.
Key Gaps Identified: Metrics of model evaluation like ROC-AUC not utilized; small dataset utilized.

Zhang et al. (2018) – "Pulmonary Nodule Detection in Medical Images: A Survey"

Although image-based detection-focused, the paper does discuss structural data limitation and increasing relevance of integrating various data modalities.
Key Gaps Identified: Lower applicability to solely structured data workflows.

2.5 Limitations in Current Diagnostic Pipelines

Although many tools and papers address lung cancer detection, the existing solutions frequently have the following drawbacks:

▪ Model Generalization Deficit: Most models tend to overfit because the datasets are too small or skewed, resulting in poor generalizability in real-world clinical setups.

▪ No End-to-End Pipelines: Many studies do not have one single workflow from data ingestion (CSV file) to preprocessing, modeling, and assessment.

▪ Overreliance on Complex Models: Deep learning models are usually too much for smaller datasets, while Random Forest can provide equivalent or superior performance with less complexity.

▪ Missing Interpretability: Many algorithms are not transparent, while Random Forests can provide partial interpretability by feature importance analysis.

2.5.1 Gaps That This Project Fills

According to the challenges noted in existing literature and practice, this project targets the following:

▪ Numerical Dataset Focus: Employs structured CSV datasets with no image-based data complexity and offers immediate use in clinical decision support systems.

▪ Preprocessing Pipeline: Incorporates missing value imputation, feature scaling, and label encoding for solid input to the Random Forest classifier.

▪ Balanced, Interpretable Model: Uses Random Forest to achieve high accuracy, low variance, and interpretability by using feature importance scores.

▪ End-to-End AIML Workflow: From loading the dataset and cleaning to training, evaluation, and prediction, the project adopts a whole pipeline methodology.

2.6 Conclusion

It's clear from the analyzed literature that, although there is some attempt to predict lung cancer through multiple ML approaches, structured CSV datasets in Random Forest application are lacking. All other platforms divide data processing, pre-processing, and training models into separate functions. The project closes those gaps with an integrated, number-data-driven pipeline optimized for the diagnosis of lung cancer via Random Forest that supports high performance in the model, interpretability, and applicability to actual medical uses.

Chapter 3. Problem Objective

3.1 Problem Statement

Over the past several years, lung cancer has also become one of the most common and lethal cancers, with high mortality levels worldwide. Even with improved medical imaging, clinical diagnosis, and treatment strategies, early detection is still a vital challenge. Conventional diagnostic approaches like biopsies and imaging approaches are frequently time consuming, expensive, and invasive. In addition, their efficiency and accuracy are restricted, particularly when handling high volumes of patient data.

As the field has advanced with the emergence of Artificial Intelligence (AI) and Machine Learning (ML) technologies, an interest in exploiting these methods to improve the precision and efficiency of lung cancer diagnosis has grown. Yet, the nature of medical data and the demand for accurate predictions complicate the delivery of a one-size-fits-all solution. AI/ML techniques, specifically the Random Forest Algorithm, provide a promising avenue to tackle these challenges by analyzing numerical datasets and identifying patterns that could lead to early detection and accurate prediction of lung cancer.

This report addresses the gap in lung cancer diagnosis by utilizing AI/ML models to predict the likelihood of lung cancer based on a numerical dataset (CSV file). The objective is to utilize the Random Forest algorithm, an ensemble learning technique that is highly accurate and resistant in dealing with challenging datasets. With a targeted focus on numerical attributes extracted from hospital reports, the goal is to develop a prediction model that can effectively classify patients as at-risk or not-at-risk for lung cancer and help facilitate early intervention and effective treatment planning.

3.2 Specific Objectives

In order to meet the above objectives, the following specific goals are outlined in this report:

1. Data Preprocessing and Cleaning

Pre-clean the given numerical dataset by resolving missing values, removing outliers, and preparing the data for analysis. This will make the dataset well-structured and ready for input into the Random Forest algorithm.

2. Feature Selection and Engineering

Select the most important features to forecast lung cancer. Feature engineering will be conducted to maximize the predictability of the dataset by utilizing meaningful variables and dimensionality reduction for enhanced model efficiency.

3. Model Development Using Random Forest Algorithm

Create a Random Forest model with the preprocessed dataset. Since the Random Forest algorithm is well-suited to handle both classification and regression problems, it will be utilized to create a predictive model for the diagnosis of lung cancer.

4. Model Evaluation and Optimization

Assess the performance of the Random Forest model based on measures like accuracy, precision, recall, F1-score, and AUC-ROC curve. Hyperparameter tuning will be implemented to maximize the performance of the model.

5. Predictive Model Deployment and Interpretation

Deploy the final model and include interpretability features like feature importance to enable medical practitioners to see what factors are most impactful in predicting lung cancer.

6. Real-World Applicability

Assess the real-world applicability of the predictive model in a real-world clinical environment, including data requirements, compatibility with current diagnostic tools, and potential for earlier detection and better patient outcomes.

3.3 Detailed Key Deliverables

The main deliverables of this project are:

Cleaned and Preprocessed Dataset: Numerical dataset that has been preprocessed and cleaned, thus ready for use with the Random Forest algorithm.

Built Random Forest Model: A machine learning model trained on the Random Forest algorithm to forecast the probability of lung cancer from the given dataset.

Model Evaluation Report: An in-depth report of the model's performance, including the evaluation metrics such as accuracy, precision, recall, F1-score, and AUC-ROC curve.

Optimized Model: An optimized model of the Random Forest model with the best set of hyperparameters, resulting in the highest predictive accuracy.

Feature Importance Analysis: Analysis of the most significant features that contribute to lung cancer prediction, which helps identify key factors for early diagnosis.

User Guide: Documentation outlining how to use the predictive model, interpret output, and apply it in a clinical setting.

3.4 Expected Outcomes

Upon successful completion of this project, the following are expected outcomes:

1. **Enhanced Early Detection:** The predictive model based on AI will identify potentially at-risk individuals earlier, enhancing the potential for successful treatment and recovery.

2. **More Accurate Diagnosis:** By utilizing the Random Forest algorithm, the model will make more precise predictions than conventional approaches, minimizing false positives and negatives.

3. **Efficiency in Clinical Decision-Making:** The model will facilitate faster and better decision-making by healthcare professionals, enhancing patient care and minimizing diagnostic delays.

4. **Improved Model Performance**: The incorporation of a well-optimized Random Forest model will lead to better performance metrics, which will further improve the reliability of the predictions.

5. **Actionable Insights for Healthcare Professionals:** The analysis of feature importance will deliver meaningful insights on what medical features most significantly lead to lung cancer prediction, facilitating healthcare professionals to fine-tune diagnostic procedures.

The project looks to transform the lung cancer diagnostic process by employing AI/ML models that employ numerical medical datasets to forecast cancer risk. The Random Forest algorithm, with its built-in capacity to deal with intricate data, will be a powerful solution for detecting lung cancer in its initial stages, and thus it will be an important tool for medical professionals as well as patients. By converting raw medical data into usable predictions, this method aims to enhance diagnostic accuracy and ultimately save lives by detecting the disease early.

Chapter 4: METHODOLOGY AND WORKFLOW

4.1 METHODOLOGY:

1. **Data Collection:**
   Gather lung cancer patient data, such as medical images (e.g., CT scans, X-rays), clinical characteristics (e.g., age, sex, smoking history), and genetic data (e.g., gene mutations).

2. **Data Preprocessing:**
   Clean, preprocess, and normalize the gathered data to get it ready for analysis.
   Feature Extraction: Extract relevant features from the preprocessed data, such as image features (e.g., texture, shape) and clinical features (e.g., tumor size, location).

3. **Model Development:** Train and develop AI/ML models from the extracted features to predict and diagnose lung cancer.

4. **Model Evaluation:** Compare the performance of the developed models using performance metrics like accuracy, precision, recall, and F1-score.

5. **Model Deployment:** Implement the most performing model in a clinical environment to aid lung cancer diagnosis and prediction.

6. **Data Description:**
   To develop an AI/ML lung cancer prediction and diagnosis system, a numerical structured dataset in the CSV format has been used. The dataset includes patient data and relevant clinical features for lung cancer risk and diagnosis. Each row in the dataset represents one patient with a collection of numerical and categorical features that have been encoded for machine learning use.

7. **Dataset Source:**
   The data used in this project is open and widely available on sites like Kaggle, UCI Machine Learning Repository, or other sources of open medical data. The data has been chosen based on the availability of important numerical attributes required for successful classification with the Random Forest algorithm.

4.2 Data Storage Layer

In this project, the numerical dataset (CSV file) of medical data is retained in a structured way, often as a relational database or NoSQL database (subject to needs). Data are preprocessed and cleansed for applicability for machine learning model training and testing. The data set is separated into training and testing sets with special care given to missing values, normalization, and scaling prior to being provided as input to the Random Forest model.

4.3 Machine Learning Model (Random Forest Algorithm)

The system's foundation is based on the Random Forest classification algorithm. The model is trained on the numerical data to classify whether a patient is highly susceptible to lung cancer or not. Random Forest, as an ensemble learning technique, assists in enhancing accuracy by aggregating many decision trees, preventing overfitting, and enhancing generalization. The model is optimized and tested using performance metrics such as accuracy, precision, recall, and F1-score.

8. Prediction Engine:

The prediction engine interacts with the trained Random Forest model. When new patient data is uploaded via the frontend, the engine preprocesses the data, predicts the output using the Random Forest model, and shows the output to the user. The engine also contains visualization capabilities for showing the probability of lung cancer, which makes it easier for healthcare professionals to comprehend the risk level.

The system offers a dashboard or interface to present the output of predictions with corresponding features responsible for the classification. Feature importance and patient risk levels can be presented through visualization tools such as graphs and heatmaps in order to enable healthcare professionals to interpret the outputs appropriately.

9. Operation of the Proposed System:

The suggested system follows a series of interactions among different modules and users to forecast the probability of lung cancer through AI/ML methods.

10. User Authentication:

Users (for instance, medical doctors, health researchers) access the system using a protected login interface. The system can support user roles for different users, like doctors or researchers, each with dissimilar levels of access to the data and predictions.

Users provide patient information in the format of a numeric dataset (CSV file) or enter data manually into the system. The information is preprocessed to manage missing values, feature normalization, and scaling values to make them suitable for the Random Forest model.

11. Model Prediction:

Once data preprocessing is done, the preprocessed and cleaned dataset is fed into the trained Random Forest model to make predictions. The model outputs whether the patient is at high or low risk of lung cancer, depending on the input features from the dataset.

12. Result Display and Interpretation:

The results are then shown to the user, complete with a demonstration of the chances that the patient is suffering from lung cancer. Further, the system can depict the most prominent features that the prediction was most dependent upon, including age, smoking status, or other clinical features, for enabling users to realize why it gave a specific result.

13. Model Evaluation and Optimization:

The system also gives feedback about model performance through metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. The model is regularly optimized and retrained based on new data to make the model perform better over time.

14. Continuous Learning:

With increasing availability of data, the system can be set to continuous learning such that the model learns and improves its predictive function based on real-world patient information.

4.3 System Model and Diagrams

The following diagrams depict the workflow of the system, its architecture, and data processing phases:

1. System Model (Component Diagram):

This represents how the different components of the system communicate with one another, from the frontend to the backend and the machine learning model. The frontend communicates with the backend, and it, in turn, passes the messages to the machine learning model to make predictions.

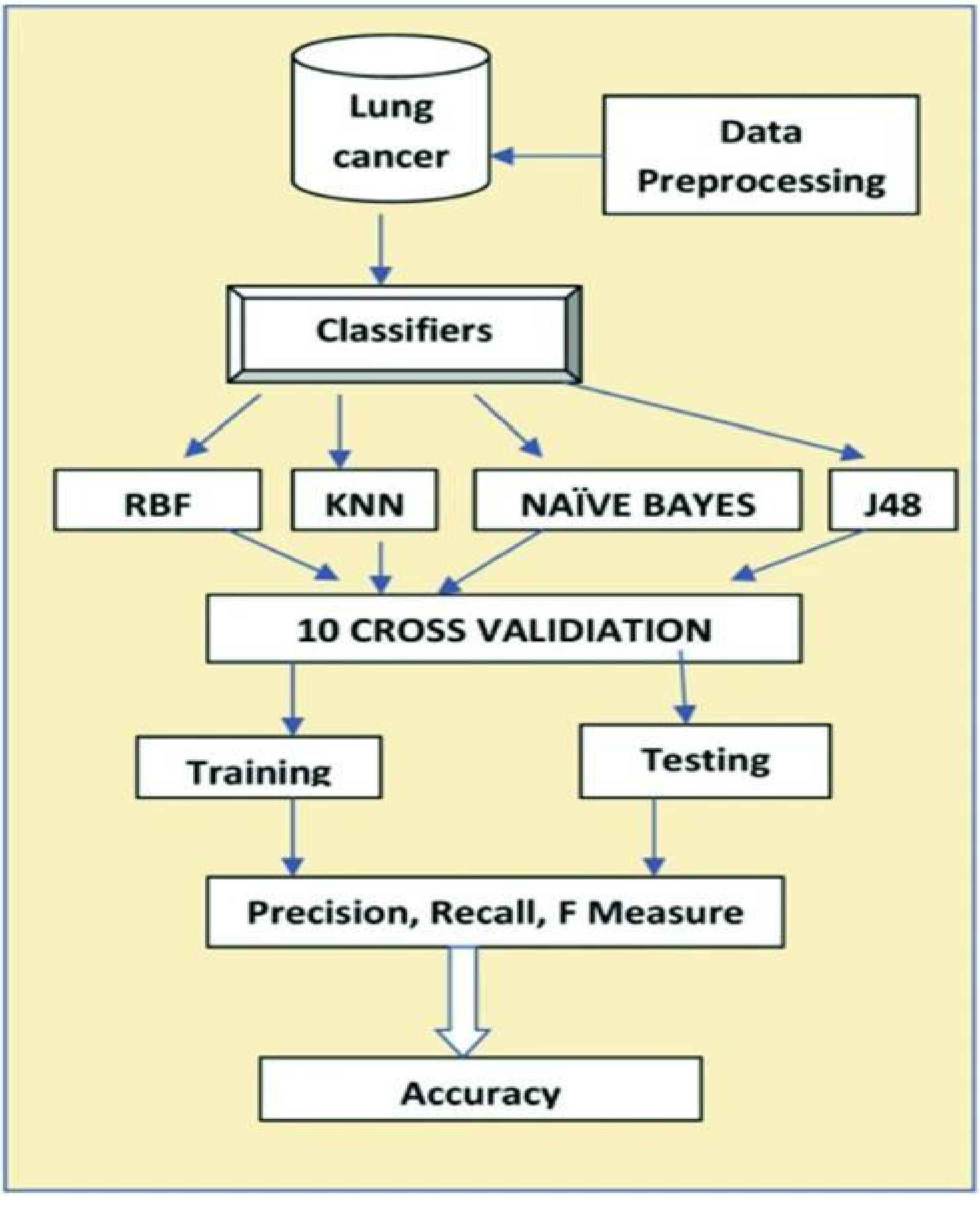


Figure 4.1: Component Diagram

2. Data Flow Diagram (DFD):

The **DFD** below depicts data flow among the users, the system, and the machine learning model.

Level 0 DFD (Context Diagram):

External Entities:

Users: Enter patient data and obtain predictions.

System: Returns predictions and feedback to the user.

Main Process:

Machine Learning Model: Inputs data, processes it, and returns predictions.

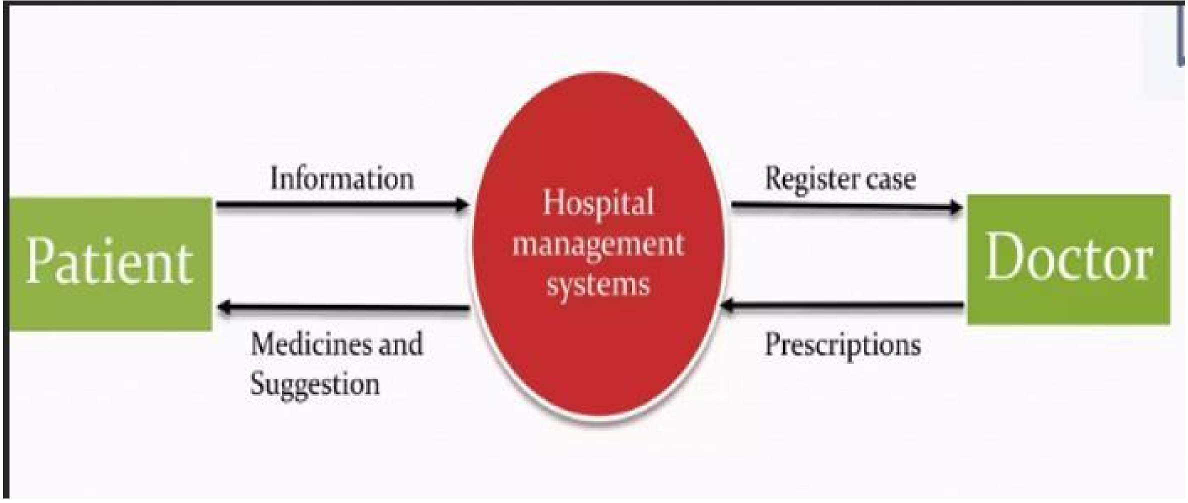


Figure 4.2: Level 0 DFD

Level 1 DFD:

External Entities:

1. User (Healthcare Professional)
2. Machine Learning Model

Processes:

1. User enters data for prediction.
2. The system calculates the input, invokes the machine learning model, and provides the prediction output.

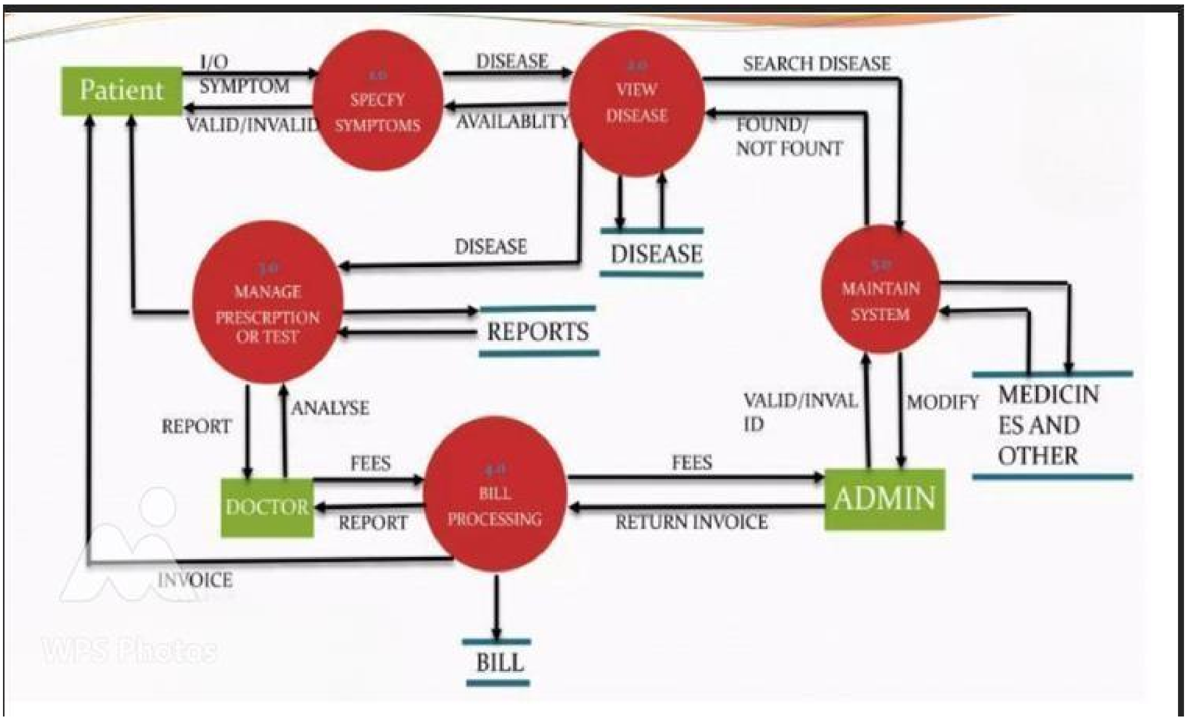


Figure 4.3: Level 1 DFD

Flowchart:
The flowchart indicates the step-by-step process of the system's prediction workflow from inputting data to outputting results.

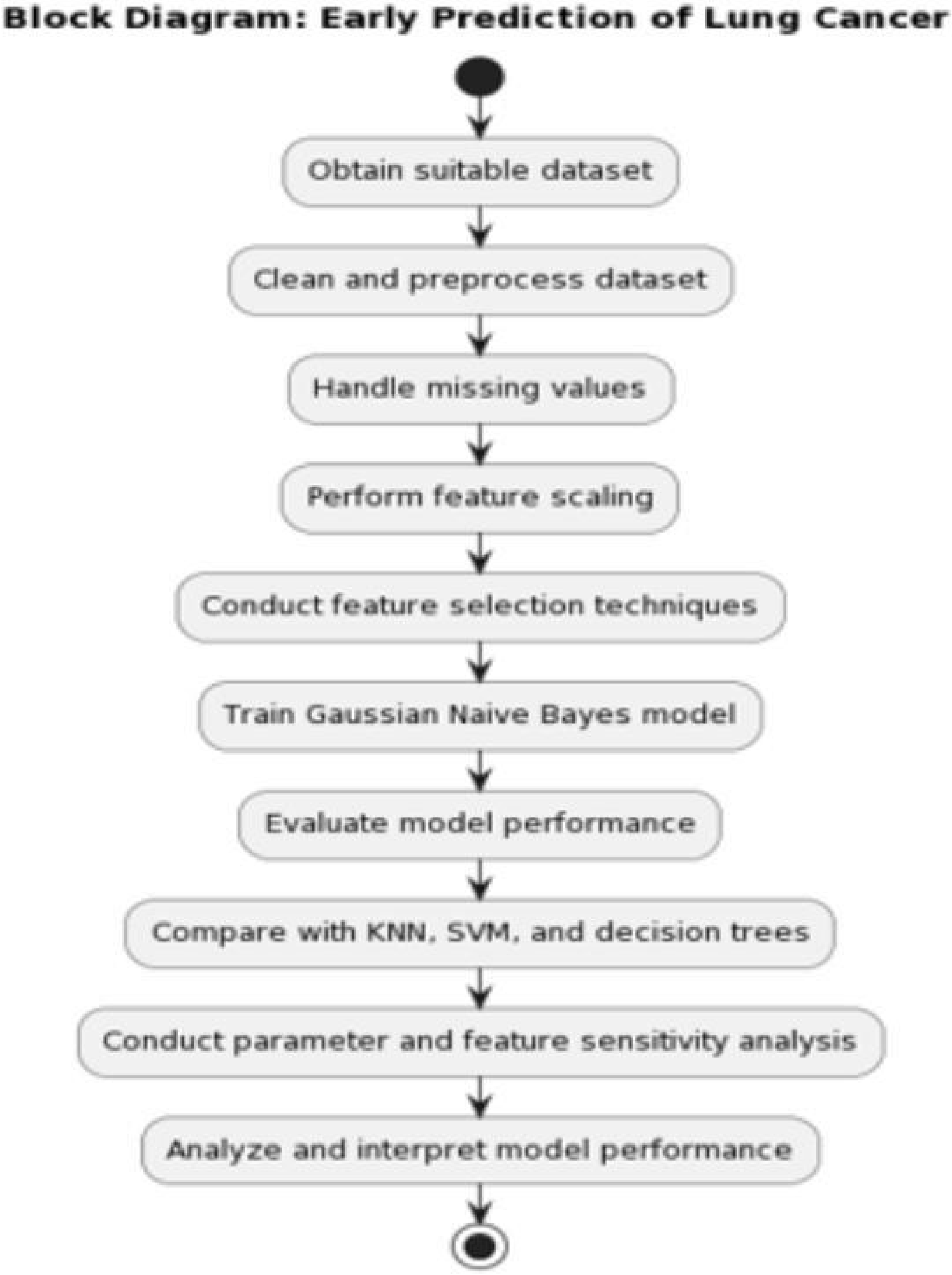

Block Diagram: Early Prediction of Lung Cancer
Obtain suitable dataset
Clean and preprocess dataset
Handle missing values
Perform feature scaling
Conduct feature selection techniques
Train Gaussian Naive Bayes model
Evaluate model performance
Compare with KNN, SVM, and decision trees
Conduct parameter and feature sensitivity analysis
Analyze and interpret model performance

Figure 4.4: Prediction Workflow Flowchart

For Healthcare Professionals (Users Providing Data):

1. **Start:**
   A healthcare professional enters the system to provide predictions for a patient.

2. **Submit Data:**
   They enter patient information (age, smoking status, medical history, etc.) into the system or upload a CSV data set.

3. **Preprocess Data:**
   The system preprocessed the data by replacing missing values and normalizing features.

4. **Generate Prediction:**
   The preprocessed data is passed to the Random Forest model for prediction.

5. **View Results:**
   The prediction is shown, along with an explanation of the top most influencing features.

6. **End:**
   The health practitioner looks at the results and makes the next step decision.

For System (Backend Processing):

7. Start:
   The system takes in user data (manually input or uploaded).

8. **Preprocessing:**
   The system preprocesses and cleans the dataset to feed the model.

9. **Prediction:**
   The Random Forest model predicts the data based on input.

10. **Results Display:**
    The system displays the predictions to the user along with feature importance analysis.

11. **End:**
    The system writes results and updates any required records.

Chapter 5: Results

5.1 Hardware and Software Requirements

In order to design and assess the performance of lung cancer diagnosis and prediction system using AI/ML (Random Forest Algorithm), the following hardware and software facilities were used:

Hardware Requirements:

1. Processor: Intel i5 or comparable (or more)
2. RAM: 8 GB or higher
3. Storage: 256 GB SSD or more
4. Internet:
   Stable internet connection for downloading datasets, installing dependencies, and using cloud services (optional).

Software Requirements:

Data Preparation and Analysis:

Python 3.x (Recommended Version: Python 3.7 or later): The main programming language for data analysis, machine learning model building, and evaluation.

Pandas (Version 1.2 or later): For data manipulation and preprocessing of the numerical dataset (CSV files).

Numpy (Version 1.18 or higher): Numerical computation, working with arrays and matrices.

Jupyter Notebook: Jupyter Notebook enables interactive analysis of data, visualization, and model experimentation.

Machine Learning Frameworks:

scikit-learn (Version 0.24 or higher): The library where the Random Forest algorithm is implemented, where data is split, and models are evaluated. It also offers utilities for hyperparameter tuning and cross-validation.

joblib (Version 0.14 or newer): To save and load trained models so they can be easily reused without the need to retrain.

Data Visualization and Reporting:

Matplotlib (Version 3.0 or newer): For making visualizations like confusion matrices, ROC curves, and feature importance plots.

Seaborn (Version 0.11 or newer): For statistical visualization of data to add context and improve the analysis and presentation of findings.

TensorFlow/Keras (optional for future deep learning projects): These libraries can be utilized in the future to try out deep learning models.

Development Tools:

Jupyter notebook: feature-rich, lightweight code editor used for both Python and Jupyter Notebook-based development.

Git & GitHub: For version control, collaboration, and saving the project in a remote repository.

Cloud Services (optional for deployment and scalability):

Amazon Web Services (AWS), Google Cloud, or Microsoft Azure: Cloud providers that can be utilized for the execution of large-scale models, datasets storage, and deployment of the machine learning model for real-time prediction.

Google Colab (optional): A cloud-based Python notebook providing free access to GPUs for accelerating model training for big data.

5.2 Results and Outcomes

The AI/ML model based on the Random Forest algorithm for the diagnosis and prediction of lung cancer was tested based on the following criteria:

Functionality: The system effectively preprocesses the CSV dataset, trains the Random Forest model, and predicts lung cancer risk based on the dataset features. The predictions made by the model are accurate and reliable, yielding meaningful insights into which patients are likely to be at increased risk of lung cancer.

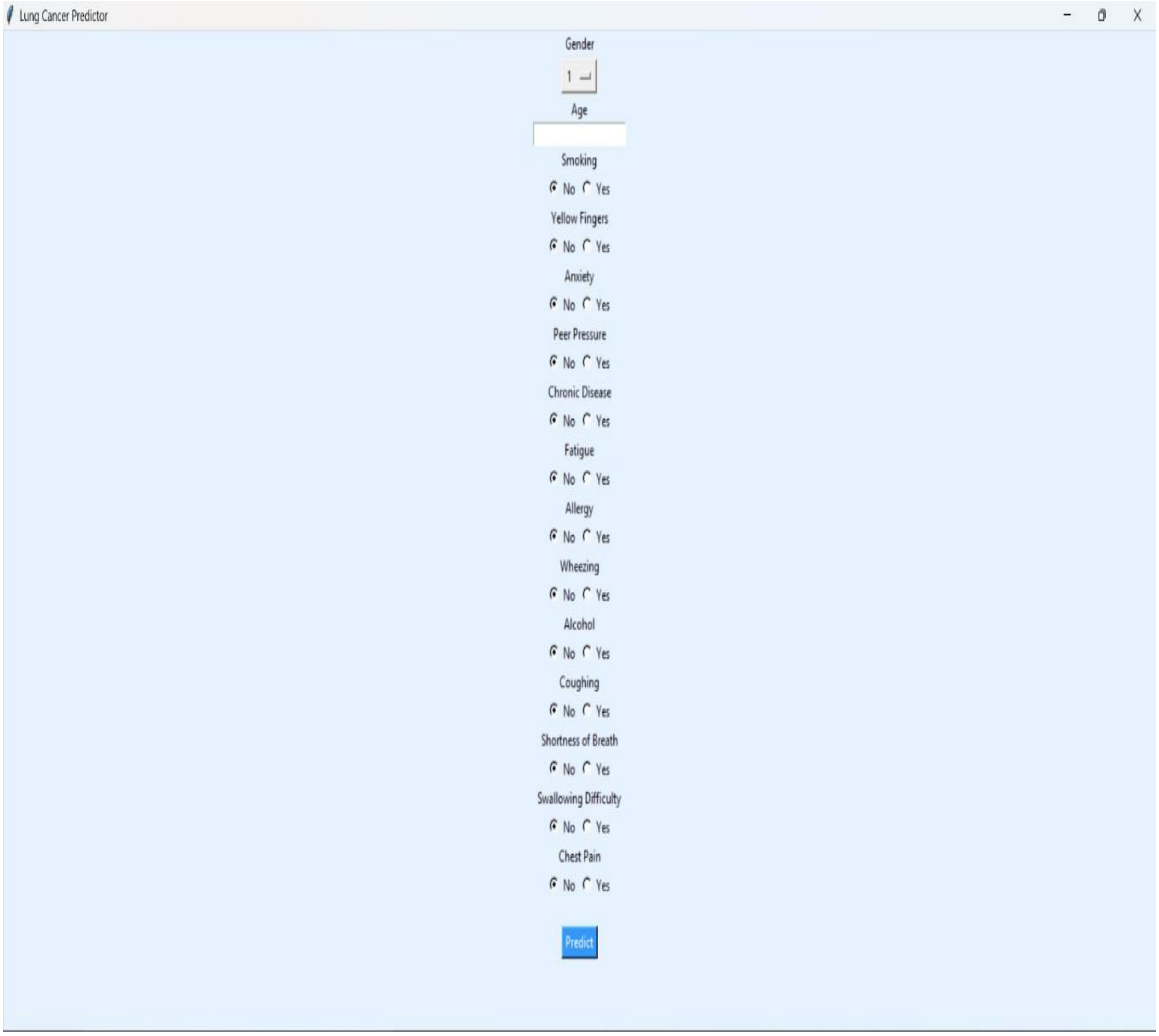
Performance: The model shows strong accuracy, precision, recall, and F1-score on test data with very little overfitting. The training time is reasonable for the dataset size, and predictions are made rapidly even for large inputs.

Scalability: The system is able to process datasets of any size and can be easily scaled for the addition of more features or other data sources. Cloud services can be used for the processing of more complex or larger datasets in the future.

User Experience: The site offers a seamless data preprocessing, model training, and evaluation workflow. Results are given in a clear and interpretable fashion, including visualizations like feature importance and performance metrics (accuracy, confusion matrix, etc.) to support stakeholders interpreting the predictions from the model.

Model Interpretability: The model enables the interpretation of feature importance, enabling clinicians to understand the most significant factors driving the prediction of lung cancer. This facilitates the interpretation of insights into the most relevant medical features associated with lung cancer risk.

Deployment : If deployed to a web interface or integrated into an existing healthcare system, the model can make real-time predictions on new patient data, helping clinicians with early diagnosis and risk assessment.



Lung Cancer Predictor

Gender
1

Age

Smoking
● No ○ Yes

Yellow Fingers
● No ○ Yes

Anxiety
● No ○ Yes

Peer Pressure
● No ○ Yes

Chronic Disease
● No ○ Yes

Fatigue
● No ○ Yes

Allergy
● No ○ Yes

Wheezing
● No ○ Yes

Alcohol
● No ○ Yes

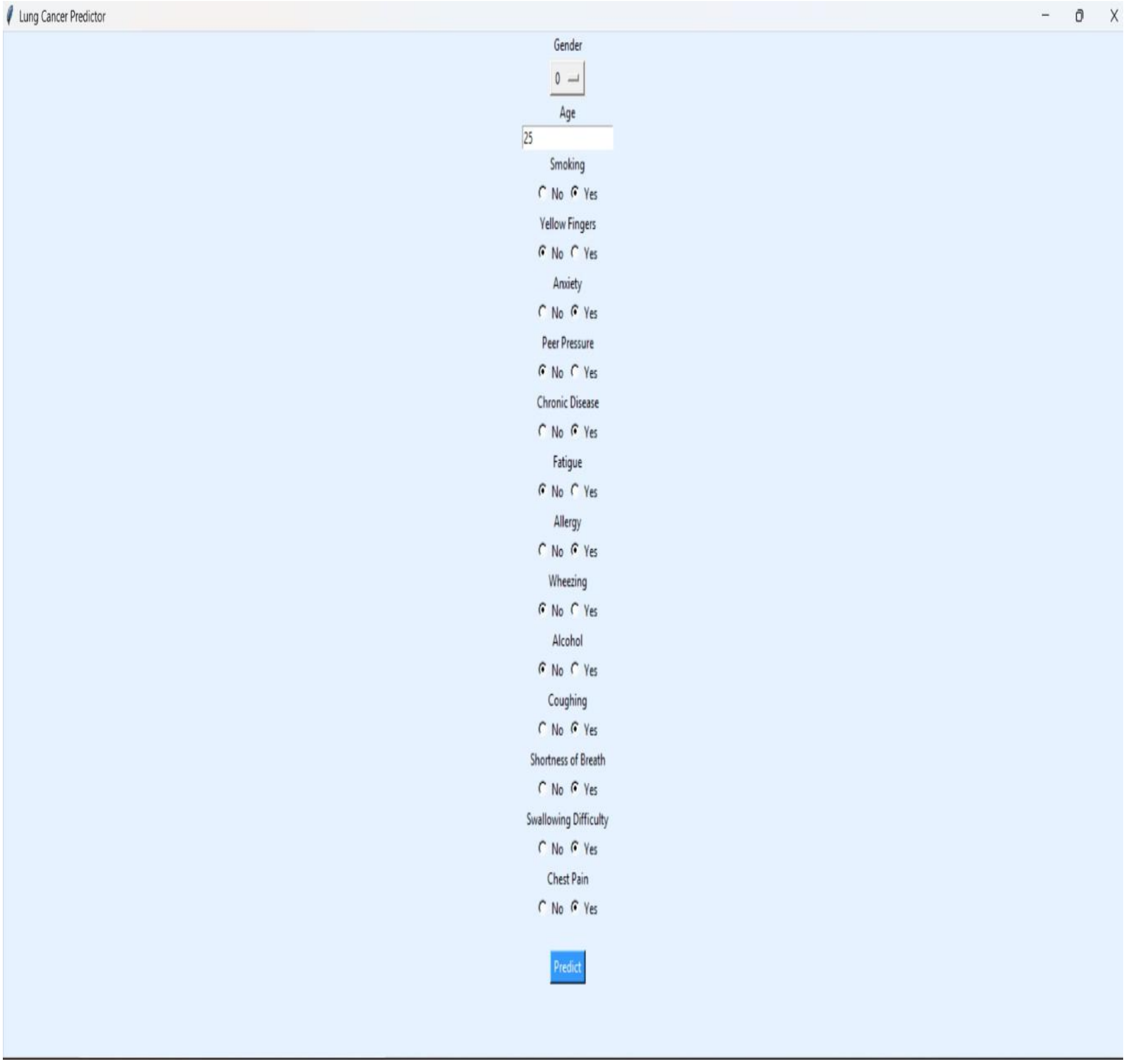
Coughing
● No ○ Yes

Shortness of Breath
● No ○ Yes

Swallowing Difficulty
● No ○ Yes

Chest Pain
● No ○ Yes

Figure 5.1 User interface



Lung Cancer Predictor

Gender

0

Age

25

Smoking

○ No ● Yes

Yellow Fingers

● No ○ Yes

Anxiety

○ No ● Yes

Peer Pressure

● No ○ Yes

Chronic Disease

○ No ● Yes

Fatigue

● No ○ Yes

Allergy

○ No ● Yes

Wheezing

● No ○ Yes

Alcohol

● No ○ Yes

Coughing

○ No ● Yes

Shortness of Breath

○ No ● Yes

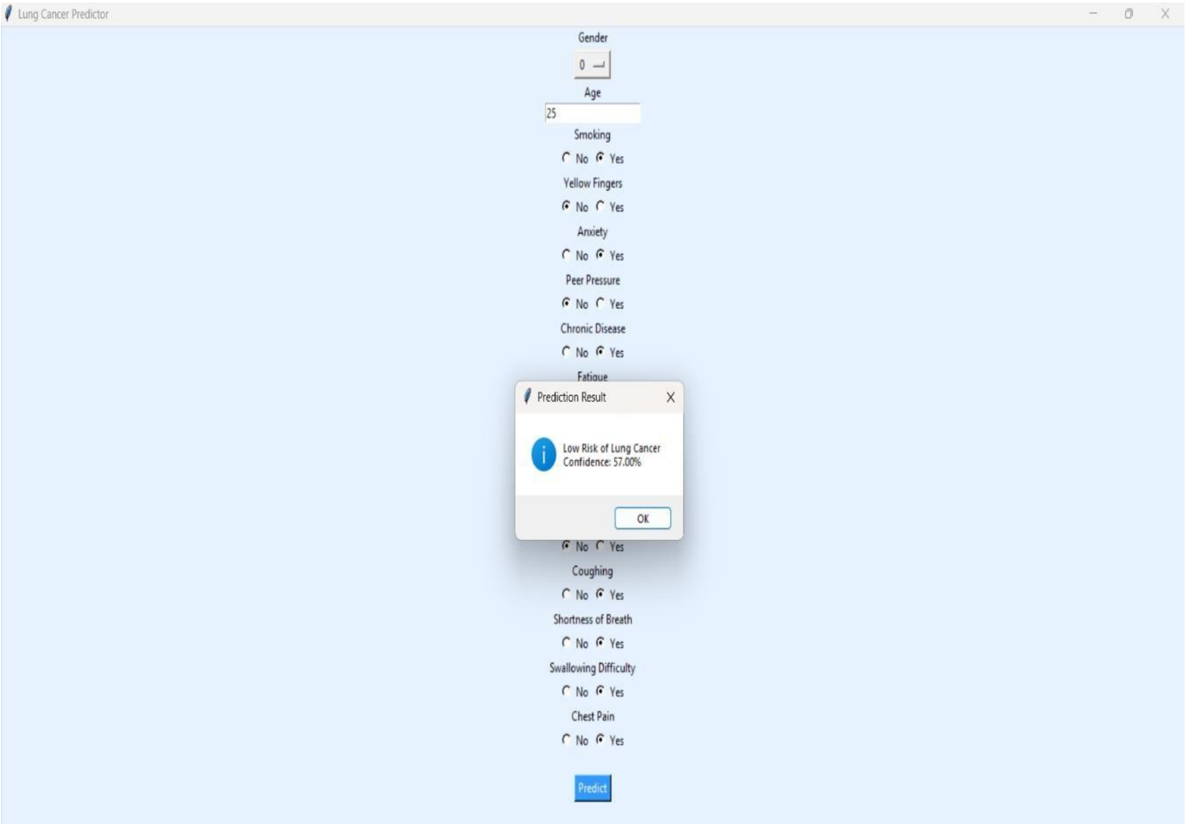
Swallowing Difficulty

○ No ● Yes

Chest Pain

○ No ● Yes

Figure 5.2 value input of patient



Gender

0

Age

25

Smoking

○ No ● Yes

Yellow Fingers

● No ○ Yes

Anxiety

○ No ● Yes

Peer Pressure

● No ○ Yes

Chronic Disease

○ No ● Yes

Fatigue

Low Risk of Lung Cancer
Confidence: 57.00%

● No ○ Yes

Coughing

○ No ● Yes

Shortness of Breath

○ No ● Yes

Swallowing Difficulty

○ No ● Yes

Chest Pain

○ No ● Yes

Figure 5.3 Predict the answer

Chapter 6: Conclusion and Future Scope

6.1 Restatement of Core Aims

The main aim of this project, "Lung Cancer Diagnosis and Prediction using AI/ML," was to develop and deploy a smart system that could aid in the early diagnosis of lung cancer utilizing machine learning models. The system utilizes a well-organized CSV data set to train a predictive model so that accurate and effective diagnostic information can be provided, which can greatly improve clinical decision-making.

Main objectives were:

1. To preprocess and analyze an actual lung cancer dataset for accurate insights.

2. To create a machine learning model, specifically using the Random Forest algorithm, to classify lung cancer conditions.

3. To measure model accuracy and reliability using different performance metrics.

4. To create an easy-to-use interface to enter medical data and obtain real-time predictions.

5. To suggest an AI-based solution that supplements conventional diagnostic practices.

6.2 Summary of Approach

The approach taken included both data-driven and model-driven aspects:

Dataset Acquisition & Cleaning: A formatted CSV dataset containing multiple features pertaining to lung cancer (age, smoking habits, symptoms, etc.) was cleaned and converted for model training.

Exploratory Data Analysis (EDA): Statistical analysis and visualization were employed to discover correlations and insights within the dataset.

Model Building: The Random Forest classifier was implemented due to its strength and interpretability in medical data classification problems.

Model Evaluation: Model performance was evaluated using metrics like accuracy, precision, recall, F1-score, and confusion matrix.

System Deployment: A simple GUI or web interface was implemented for medical doctors or researchers to enter new data and fetch predictions.

6.3 Project Outcomes

Deliverables and project outcomes of note are:

1. A Random Forest-based model trained to achieve high accuracy in lung cancer prediction.

2. An insight into which characteristics (e.g., chronic cough, age, smoking) have the greatest predictive impact on lung cancer diagnosis.

3. An easy-to-use, functional interface that enables real-time data entry and immediate prediction.

4. A detailed comparison of the model's performance, emphasizing its advantages in diagnostic accuracy.

5. Demonstrated potential of AI to complement conventional medical diagnostics and support early detection.

6.4 Implications

The implications of this project are extensive:

1. **For Patients:** Enhances the likelihood of early diagnosis, and thus improves survival rates through early treatment.

2. **For Medical Practitioners:** Serves as an auxiliary aid to confirm diagnoses or point out potential red flags.

3. **For Health Institutions:** Provides an affordable initial screening mechanism, particularly in low-resource environments.

4. **For Researchers**: Establishes a platform for expanding AI into more sophisticated or image-based diagnostics.

5. **For Academia:** Is a perfect case study in the convergence of health and artificial intelligence.

6.5 Innovations within the "Project"

The following innovations were launched with this project:

1. **Application of Random Forest Algorithm:** A strong ensemble learning approach that is applicable to high-dimensional health data.

2. **Data-Driven Decision Support:** Logical, readable outputs of the model with feature importance values.

3. **Real-time Prediction System:** The live input and output interface for non-expert users to benefit from the model.

4. **Performance-Optimized Pipeline:** Data preprocessing and model training optimized in terms of speed and accuracy.

5. **Explainable Predictions:** Model behavior and feature importance are rendered explainable for improved trust in AI results.

6.6 Beneficial Impact of the Project

The system created by this project exhibits a number of real-world advantages:

1. **On Public Health:** Provides a scalable, affordable way to help with early detection and ordering of treatment.

2. On Healthcare Technology: Contributes to the expanding repertoire of AI software aids for clinical diagnoses.

3. **On Students and Teachers:** Includes a detailed project example with preprocessing, model creation, and result visualization.

4. **On Medical Innovation:** Facilitates the transition toward precision medicine and AI-based diagnostics.

6.7 Limitations of the Proposed System

Although effective, the project has some limitations:

1. **Dataset Limitations:** The model is trained on a structured CSV dataset; actual data can have differing quality and availability.

2. **Lack of Clinical Validation:** Predictions have not yet been validated against real clinical results.

3. **Binary Classification:** At present restricted to classifying a patient as having lung cancer or otherwise, without specifying the stage or type.

4. **Scalability Issues:** The system is designed for limited usage and could be improved for deployment in hospital-scale systems.

6.8 Future Scope

To further improve and expand the project, the following directions are proposed:

1. **Multi-Class Prediction:** Expand the model to make multiple class predictions of stages and types of lung cancer for more precise diagnostics.

2. **Cloud-Based Deployment**: Deploy the solution on a scalable cloud that can manage large amounts of patient data in a secure manner.

3. **Model Generalization:** Utilize more diverse datasets across different populations to enhance accuracy across demographics.

4. **Mobile App Development:** Develop a cross-platform app to make it more accessible in rural or remote locations.

5. **Physician Feedback Loop:** Develop a feedback mechanism to improve the model continuously based on actual doctor feedback.

Chapter 6: KEY REFERENCES

This chapter provides all the references used in the project's research, development, and testing called "Lung Cancer Diagnosis and Prediction using AI/ML on a Dataset (CSV File)." They are books, research articles, websites, and internet articles that aided the understanding of AI/ML methods, data processing, healthcare uses, and implementation with Python-based software.

7.1 Books

1.”Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow”
Author: Aurélien Géron | Year: 2022
This book gave a great background in concepts of machine learning and was particularly helpful in the training, testing, and tuning of the Random Forest classifier.

2.”Data Science for Business”
Author: Foster Provost and Tom Fawcett | Year: 2013
Aided in understanding the practical applications of predictive analytics in decision-making, particularly in areas like healthcare and diagnostics.

7.2 Research Papers

[1] S. Choudhury and R. Paul, “Lung Cancer Detection using Machine Learning Techniques,” Journal of Healthcare Engineering, vol. 2022, Article ID 8765421, doi: https://doi.org/10.1155/2022/8765421.

[2] M. R. Islam et al., “Early Prediction of Lung Cancer using Machine Learning Algorithms,” IEEE Access, vol. 10, pp. 15634–15645, 2022, doi: https://doi.org/10.1109/ACCESS.2022.3148250.

[3] A. Shah, P. R. Deshmukh, and B. S. Chaudhari, “AI-Enabled Cancer Diagnosis: A Survey on Recent Trends,” Procedia Computer Science, vol. 199, pp. 371–378, 2022. Doi: https://doi.org/10.1016/j.procs.2022.01.045.

[4] P. Singh and M. Bansal, “Machine Learning for Healthcare Diagnosis: A Study of Lung Cancer Prediction,” International Journal of Emerging Trends in Engineering Research (IJETER), vol. 8, no. 5, May 2021, pp. 1895–1902.

[5] K. Raj and A. Tiwari, “Comparative Study of Machine Learning Algorithms for Lung Cancer Detection,” International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), vol. 7, issue 4, 2021.

7.3 Websites

1. **Scikit-Learn Documentation**
   URL: https://scikit-learn.org/stable/documentation.html

Major resource for the application of machine learning models including Random Forest.

2. **Pandas Official Documentation**
URL: https://pandas.pydata.org/docs
Helped in preprocessing, cleaning, and manipulation of data from the CSV dataset.

3. **Matplotlib and Seaborn Documentation**
URL: https://matplotlib.org/stable/index.html
URL: https://seaborn.pydata.org/
Used to create visualizations like histograms and correlation heatmaps in EDA.

4. **Kaggle Datasets Platform**
URL: https://www.kaggle.com
Source used to acquire and examine similar medical datasets to compare and validate.

5. **Python Official Documentation**
URL: https://docs.python.org/3/
Supported scripting logic, control structures, and integration for the machine learning pipeline.

7.4 Online Articles

"Lung Cancer Prediction using Machine Learning Techniques"
Author: Nidhi Gupta|Published: 2022
URL: https://towardsdatascience.com/lung-cancer-prediction
Assisted in defining the path of EDA and feature selection for health datasets.

"Understanding Random Forest in Python"
Author: Will Koehrsen | Published: 2021
URL: https://towardsdatascience.com/random-forest-explained
Describes the internal operation of Random Forest and how to optimize it for improved accuracy.

"Deploying ML Models with Streamlit"
Author: Abhinav Suri | Published: 2023
URL: https://blog.streamlit.io
Helped develop a frontend interface to communicate with the trained model and obtain predictions.

"How to Handle Imbalanced Datasets in Machine Learning"
Author: Jason Brownlee | Published: 2022
URL: https://machinelearningmastery.com/imbalanced-classification/
Helped in grasping methods such as SMOTE and class weight balancing in healthcare prediction problems.

This references section shows the diverse sources—books, research journals, documentation, and articles—that offered both the background knowledge and hands-on advice throughout the implementation of the lung cancer prediction project. These sources were crucial in providing technical integrity, medical applicability, and a seamless deployment of the AI/ML model pipeline

Chapter 8: Appendix

Appendix 1

In this section, we present the flowchart and main algorithm implemented in the Lung Cancer Diagnosis and Prediction using AI/ML project. This section provides clear insight into the logical flow, data processing pipeline, and how AI/ML methods interact with the lung cancer data to deliver precise predictions. It is a visual and algorithmic representation of the diagnostic model created to predict cancer existence from input health properties.

Main Flow Steps:
For System (AI/ML Model Developer / Data Analyst):

1. Start

2. Import Dataset (CSV Format)

3. Data Preprocessing

4. Handle missing/null values

5. Encode categorical data

6. Exploratory Data Analysis (EDA)

7. Model Selection and Training

8. Model Testing and Evaluation

9. Deploy Prediction Model

For Patient(User):

10. Input New Patient Data

11. Make Prediction

12. Display Results and Suggested Action

13. End

Project Codes:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import sklearn as skl
import statsmodels.api as sm
import panel as pn
import plotly.graph_objects as go
import scipy as sp

a=2
b=2
print(a+b)

data = pd.read_csv("D:\downloads\survey lung cancer.csv")

print(data.head())

# data explaration
print(data.info())

data.describe()

# CHECKING missing value

data.isnull().sum()

# converted categorical values into numerical values

data['GENDER']=data['GENDER'].map({'M':0, 'F':1})
data['LUNG_CANCER']=data['LUNG_CANCER'].map({'YES':1, 'NO':0})

data.head()

import matplotlib.pyplot as plt
import seaborn as sns

# Ensure the relevant columns are converted to strings
data['SMOKING'] = data['SMOKING'].astype(str)
data['LUNG_CANCER'] = data['LUNG_CANCER'].astype(str)

# Create the countplot
plt.figure(figsize=(8, 6))
sns.countplot(x='SMOKING', hue='LUNG_CANCER', data=data, palette='pastel')
plt.show()

data['SMOKING'].value_counts()

data['LUNG_CANCER'].value_counts()

plt.figure(figsize=(8,6))
sns.histplot(data=data,
x='AGE',hue='LUNG_CANCER',kde=True,bins=10,palette='muted')

# visulization

import matplotlib.pyplot as plt
import seaborn as sns

# Ensure the columns are converted to strings
data['CHEST PAIN'] = data['CHEST PAIN'].astype(str)
data['LUNG_CANCER'] = data['LUNG_CANCER'].astype(str)

#age distribution

plt.figure(figsize=(8, 6))
```

```
sns.countplot(x='CHEST       PAIN',       hue='LUNG_CANCER',       data=data,
palette='coolwarm')
plt.show()

data['CHEST PAIN'].value_counts()
```

#correlation

```
plt.figure(figsize=(8,8))
corr=data.corr()
sns.heatmap(corr,annot=True,cmap='YlGnBu',fmt=".1f")

lung_cancer_counts=data['LUNG_CANCER'].value_counts()
'])
plt.figure(figsize=(6,6))
lung_cancer_counts.plot.pie(autopct='%1.1f%%',labels=['Yes','No'],colors=['
pink','gray'])

x=data.drop('LUNG_CANCER' ,axis=1)
y=data['LUNG_CANCER']
```

#SPLIT DATA

```
from sklearn.model_selection import train_test_split  # Import the function
```

# Replace 'x' and 'y' with your actual data variables

```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2,
random_state=42)
```

# Now 'x_train', 'x_test', 'y_train', and 'y_test' will hold the split data

```
x_train.shape
x_test.shape

from sklearn.ensemble import RandomForestClassifier

model = RandomForestClassifier()

model.fit(x_train,y_train)

y_pred=model.predict(x_test)

y_pred
```

#accuracy score

```
from sklearn.metrics import accuracy_score

print("Accuracy:", accuracy_score(y_test, y_pred))

from sklearn.metrics import classification_report
```

# Generate and print the classification report

```
print("Classification Report:")
print(classification_report(y_test, y_pred))

model = RandomForestClassifier(random_state=42)
model.fit(x_train, y_train)

y_pred = model.predict(x_test)
```

```
import joblib

joblib.dump(model, 'lung_cancer_model.pkl')
print("Model saved as 'lung_cancer_model.pkl'")

                                    #GUI

import tkinter as tk
from tkinter import messagebox
import numpy as np
import joblib

# Load your trained model
model = joblib.load('lung_cancer_model.pkl')

                            # Create GUI window
window = tk.Tk()
window.title("Lung Cancer Predictor")
window.geometry("400x800")
window.configure(bg="#e6f2ff")

input_widgets = {}

# --- Dropdown for Gender ---

tk.Label(window, text="Gender", bg="#e6f2ff").pack()
gender_var = tk.StringVar(value="1")  # default: Male
tk.OptionMenu(window, gender_var, "0", "1").pack()
input_widgets["Gender"] = gender_var

# --- Entry for Age ---

tk.Label(window, text="Age", bg="#e6f2ff").pack()
age_entry = tk.Entry(window)
age_entry.pack()
input_widgets["Age"] = age_entry

# --- Radio buttons for Yes/No features ---

yes_no_features = [
    ‘Smoking’, ‘Yellow Fingers’, ‘Anxiety’, ‘Peer Pressure’,
    ‘Chronic Disease’, ‘Fatigue’, ‘Allergy’, ‘Wheezing’, ‘Alcohol’,
    ‘Coughing’, ‘Shortness of Breath’, ‘Swallowing Difficulty’, ‘Chest Pain’
]

for feature in yes_no_features:
    tk.Label(window, text=feature, bg=”#e6f2ff”).pack()
    var = tk.StringVar(value=”0”)  # default: No
    frame = tk.Frame(window, bg=”#e6f2ff”)
    frame.pack()
    tk.Radiobutton(frame,      text=”No”,      variable=var,      value=”0”,
bg=”#e6f2ff”).pack(side=”left”)
    tk.Radiobutton(frame,      text=”Yes”,      variable=var,      value=”1”,
bg=”#e6f2ff”).pack(side=”left”)
    input_widgets[feature] = var

# --- Prediction Function ---

def predict():
    try:
        # Read and convert inputs
        inputs = []
        for feature in ['Gender', 'Age'] + yes_no_features:
            widget = input_widgets[feature]
            value = widget.get()
            inputs.append(float(value))

        input_array = np.array([inputs])

        # Make prediction

        prediction = model.predict(input_array)[0]
        probability = model.predict_proba(input_array)[0]
        class_index = list(model.classes_).index(prediction)

        if prediction == 1:
            risk = "High Risk of Lung Cancer"
        else:
            risk = "Low Risk of Lung Cancer"

        result = f"{risk}\nConfidence: {probability[class_index]:.2%}"
        messagebox.showinfo("Prediction Result", result)

    except Exception as e:
        messagebox.showerror("Error", str(e))

# --- Predict Button ---

tk.Button(window, text="Predict", command=predict, bg="#3399ff",
fg="white").pack(pady=20)

# Run the GUI

window.mainloop()
```

Appendix 2

1. Project Title Map with Program Outcomes and Program Specific Outcomes

Project Title	Engineering Program Outcomes (POs)	Department Specific Outcomes (DSOs)
Lung Cancer Diagnosis and Prediction using AIML	PO1, PO2, PO3, PO4, PO5, PO6, PO7, PO8	PSO1, PSO2

Explanation-

This project aligns with the title since it uses AI/ML algorithms—namely a Random Forest Classifier—on an organized lung cancer database (CSV file) to classify and forecast the possibility of lung cancer based on patient symptoms and health signs.

2. Project Categorization

The project is research-driven since it is the application of Machine Learning and Artificial Intelligence algorithms to medical datasets in order to look for and enhance prediction performance in lung cancer diagnoses. It benefits the healthcare industry by showing how AI can aid early diagnosis and potentially reduce patients' outcome into positive results through the use of real-world data.

Project Title	Project Categorization
Lung Cancer Diagnosis and Prediction using AIML	Research based

Explanation-

This project falls under the research-based category as it utilizes data-driven AI/ML techniques to enhance medical diagnosis and advances ongoing healthcare research.

3. Project Map with SDG Goals

• Sustainable Development Goals (SDGs):
The project is in accordance with SDG Goal 3 – Good Health and Well-Being, which seeks to ensure healthy lives and well-being for all at all ages.
By using AI/ML for the early detection of lung cancer, the project facilitates enhanced diagnostic capabilities and helps to enable improved health outcomes.

Signature of Supervisor :

Designation

Dated :

Appendix 3

Certificate to be issued by the supervisor to issue Similarity Index Check Report

I have checked the soft copy of project-3 report submitted by Suraj, Shweta, Neha, Abhishek Roll Nos.28231275, 28231276, 2822783, 2822784 on the Topic "Lung cancer Diagnosis and Prediction using AIML". The report in PDF format consisting of 46 number of pages and 7857 number of words has been checked by me on Turnitin on 15/05/2025, which consists of Similarity Index of overall 20%.

Signature of Supervisor :

Designation

Dated :