LUNG CANCER PREDICTION USING SHAP AND ML TECHNIQUES

A PROJECT REPORT

submitted in partial fulfilment of the requirements for the degree of

Bachelor of Technology

in

COMPUTER ENGINEERING

Major Project I (01CE0716)

Submitted by

SHUBHAMKUMAR PANDIT

92200103155

NEEHARIKA JOSHI

92200103099

Faculty of Engineering & Technology

Marwadi University, Rajkot

August, 2025

Major Project I (01CE0716)

Department of Computer Engineering

Faculty of Engineering & Technology

Marwadi University

A.Y. 2025-26

CERTIFICATE

This is to certify that the project report submitted along with the project entitled **Lung Cancer Prediction Using SHAP and ML Techniques** has been carried out by **Shubhamkumar Pandit** (92200103155), **Neeharika Joshi** (92200103099), under my guidance in partial fulfilment for the degree of Bachelor of Technology in Computer Engineering, 7th Semester of Marwadi University, Rajkot during the academic year 2025-26.

Shailendrasinh Chauhan	Dr. Krunal Vaghela
Assistance Professor	Professor & Head
Department of Computer Engineering	Department of Computer Engineering

Major Project I (01CE0716)

Department of Computer Engineering

Faculty of Engineering & Technology

Marwadi University

A.Y. 2025-26

CERTIFICATE

This is to certify that the project report submitted along with the project entitled **Lung Cancer Prediction Using SHAP and ML Techniques** has been carried out by **Shubhamkumar Pandit** (92200103155) under my guidance in partial fulfilment for the degree of Bachelor of Technology in Computer Engineering, 7th Semester of Marwadi University, Rajkot during the academic year 2025-26.

Shailendrasinh Chauhan	Dr. Krunal Vaghela
Assistance Professor	Professor & Head
Department of Computer Engineering	Department of Computer Engineering

Major Project I (01CE0716)

Department of Computer Engineering

Faculty of Engineering & Technology

Marwadi University

A.Y. 2025-26

CERTIFICATE

This is to certify that the project report submitted along with the project entitled **Lung Cancer Prediction Using SHAP and ML Techniques has** been carried out by **Neeharika Joshi** (92200103099) under my guidance in partial fulfilment for the degree of Bachelor of Technology in Computer Engineering, 7th Semester of Marwadi University, Rajkot during the academic year 2025-26.

Shailendrasinh Chauhan	Dr. Krunal Vaghela
Assistance Professor	Professor & Head
Department of Computer Engineering	Department of Computer Engineering

Major Project (01CE0716)

Department of Computer Engineering

Faculty of Engineering & Technology

Marwadi University

A.Y. 2025-26

DECLARATION

We hereby declare that the **Major Project-I (01CE0716)** report submitted along with the Project entitled **Lung Cancer Prediction Using SHAP and ML Techniques** submitted in partial fulfilment for the degree of Bachelor of Technology in Computer Engineering to Marwadi University, Rajkot, is a Bonafide record of original project work carried out by me us at Marwadi University under the supervision of **Prof. Shailendrasinh Chauhan** and that no part of this report has been directly copied from any students' reports or taken from any other source, without providing due reference.

Sr. No	Student Name	Sign
1	Shubhamkumar Pandit (shubhamkumar.pandit119241@marwadiuniversity.ac.in)	
2	Neeharika Joshi (neeharika.joshi118492@marwadiuniversity.ac.in)	

Acknowledgement

Behind every successful project lies a blend of dedication, inspiration, and encouragement from several individuals whose support often goes beyond words. As we reach the culmination of our Major Project titled **"Lung Cancer Prediction Using SHAP and ML Techniques"** we take this opportunity to reflect with immense gratitude on the journey that brought us here.

First and foremost, we owe our deepest appreciation to **Prof. Shailendrasinh Chauhan**, Assistant Professor, Department of Computer Engineering, Marwadi University. His continuous support, insightful guidance, and constructive feedback were instrumental in shaping our project. His patience and belief in our potential encouraged us to explore new perspectives and push the boundaries of our understanding. He not only guided us technically but also instilled in us a research-oriented approach that will remain valuable throughout our academic and professional careers.

We extend our sincere thanks to **Dr. Krunal Vaghela**, Professor and Head, Department of Computer Engineering, for his constant encouragement and for providing an enriching academic environment. His leadership and commitment to quality education at Marwadi University have inspired us throughout our learning journey.

We would also like to express our gratitude to the entire faculty and staff of the **Department of Computer Engineering**, who have consistently supported us with their knowledge, resources, and timely assistance whenever required. Their dedication to student growth has played a significant role in our academic development.

A special thanks to **Marwadi University** for offering us a platform where curiosity meets opportunity. The university's culture of innovation, collaboration, and academic rigor has been crucial in transforming our ideas into a meaningful project.

Abstract

Lung cancer remains the leading cause of cancer-related mortality worldwide, primarily due to its late-stage diagnosis and limited treatment options at advanced stages. This project addresses the ***challenge of early detection*** *by integrating machine learning (ML) models with* ***explainable artificial intelligence (XAI)*** *techniques. Clinical text-based data comprising 5,000 patient records with 18 medical attributes, along with CT scan images, were analysed to classify cases into benign, malignant, or normal categories.*

Multiple ML models, including ***Random Forest and Logistic Regression****, were employed for textual data, while advanced convolutional neural network architectures such as* ***DenseNet121 and ResNet50*** *were applied to image data.* ***SHAP (SHapley Additive exPlanations) and Grad-CAM*** *were utilized to enhance interpretability, providing feature-level and visual explanations for predictions, thereby improving transparency and clinical trust.*

Experimental results demonstrate high accuracy, precision, and recall, with ***DenseNet121*** *achieving superior performance on CT scans. The findings highlight key risk factors such as smoking, throat discomfort, breathing issues, and environmental exposure, reinforcing their diagnostic significance. By combining predictive accuracy with interpretability, this work demonstrates the potential of AI-assisted systems to support early diagnosis, reduce diagnostic errors, and contribute toward more effective clinical decision-making in lung cancer care.*

List of Figures

List of Tables

Abbreviations

AI	Artificial Intelligence
AUC	Area Under the Curve
CNN	Convolutional Neural Network
CT	Computed Tomography
DL	Deep Learning
EDA	Exploratory Data Analysis
EHR	Electronic Health Record
F1-score	Harmonic Mean of Precision and Recall
FN	False Negative
FP	False Positive
Grad-CAM	Gradient-weighted Class Activation Mapping
IoT	Internet of Things
KNN	K-Nearest Neighbours
LR	Logistic Regression
ML	Machine Learning
MLP	Multi-Layer Perceptron
NB	Naïve Bayes
PLCO	Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial
RF	Random Forest
ROC	Receiver Operating Characteristic
ROC-AUC	Receiver Operating Characteristic – Area Under the Curve
SVM	Support Vector Machine
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Over-sampling Technique
TN	True Negative

TP	True Positive
XAI	Explainable Artificial Intelligence
XGB / XGBoost	Extreme Gradient Boosting

Table of Contents

CHAPTER 1

INTRODUCTION

1.1 Background and Motivation

One of the common and deadliest cancers is lung cancer, which causes more than 2.2 million new cases and 1.8 million deaths each year worldwide. It has a high mortality rate mainly because it is commonly diagnosed at an advanced stage where treatment is no longer an option and the probability of survival is at a very minimal level. Early identification is the most critical issue in patient outcome improvement but it is still a vital problem in clinical practice. Traditional methods of diagnosing, including imaging scans and tissue biopsies, are quite common, efficient, and have their own limitations. They can be invasive, time-consuming, expensive and also can be subject to human error or interpretation bias. These issues complicate the reliable detection of lung cancer at an early stage, which postpones timely intervention and aggravates the patient prognosis.

Over the last few years, new opportunities in resolving these issues have been created by the development of Artificial Intelligence (AI) and Machine Learning (ML). Through big data, AI-driven models will be able to find hidden patterns in patient data, imaging data, and lifestyle that would not be easily detected in traditional diagnosis techniques. When used with explainability tools, machine learning algorithms do not only increase accuracy, but also give us insight into factors contributing to predictions, thus increasing clinical trust. These new technologies can potentially screen lung cancer at an earlier, more accurate and less invasive stage than previous procedures. Therefore, AI-based solutions are gradually being acknowledged as potent tools to aid clinicians, decrease the time of diagnosis, and eventually enhance the survival rates and the quality of care available to patients with lung cancer.

1.2Problem Statement

In spite of the modern technology in medical imaging and diagnostic, lung cancer is still being diagnosed at later stages of the disease when there is less treatment options available and the survival rates become very low. Clinically competent but commonly costly, invasive

and highly reliant on manual interpretation by radiologists, are traditional methods of diagnosis such as CT scans and biopsies which are susceptible to human error. Moreover, numerous of the prevailing machine learning and deep learning algorithms to predict lung cancer have high accuracy yet become black-box models, providing minimal to no interpretability. This is not a transparent setup, thus making adoption in clinical settings a barrier, as explainability and trust are key to medical decision-making.

Also, the majority of studies have used either clinical data or imaging data individually thus limiting the predictive power of models because the models have not utilized the complementary information conveyed by multimodal data. These shortcomings underscore the urgency to provide a holistic, interpretable, and precise system capable of anticipating lung cancer at its early stage through the use of both clinical and imaging data with the ability to explain to its predictions in a manner that can be relied upon in making clinical decisions.

1.3 Objectives of the Study

The primary objective of this project is to design and develop an intelligent and explainable system that will be able to predict lung cancer at an early stage through the application of both machine learning and advanced explainable artificial intelligence (XAI) architectures. In contrast to the time-consuming and invasive nature of other forms of traditional diagnostic testing, the system allows the integration of various kinds of data, such as clinical records, text-based datasets, and CT scan images, to create predictive models that are capable of categorizing cases with high accuracy (benign, malignant, or normal) into one or more categories. The framework will increase diagnostic accuracy by introducing a variety of input sources to help deal with the multidimensionality of detecting lung cancer.

In addition to predictive performance, another major area of concern in this project is enhancing interpretability, which is currently a significant obstacle to AI system adoption in clinical practice. To address this, this study uses SHAP to analyse clinical and textual data and Grad-CAM to interpret deep learning networks used on CT scans. These tools are transparent in that they recognize the strongest risk factors, including smoking history and age, along with respiratory symptoms, and specific areas in CT images that lead the model to its predictions. This interpretability builds trust in healthcare professionals, since they can not only know what is predicted by the model, but also understand why a specific decision has been taken.

In the end, this project shows that by ensuring accuracy and interpretability, diagnostic errors can be greatly minimized, earlier intervention may be provided, and patient survival rates can be enhanced. The research aims to close the divide between the state-of-the-art AI technology and the safe and reliable implementation of AI in healthcare by developing a powerful, yet transparent system. By doing so, it will help clinicians make informed decisions, build more confidence in predictions made by AI and help in the general fight against lung cancer.

1.4 Scope of the Project

This project would be extended to design and development of a predictive framework which uses both clinical and imaging data to detect lung cancer in its early stages. In the study, a structured clinical dataset of 5,000 patient records is used with 18 medical attributes, including age, gender, smoking history, oxygen saturation, and family history, as well as a set of CT scan images, benign, malignant, and normal.

Machine learning, e.g., Random Forest and Logistic Regression on the clinical dataset, and deep learning, e.g., DenseNet121, ResNet50, and EfficientNet on the imaging dataset, are used to predict patterns and correlations between risk factors and extract and classify visual features of lung abnormalities respectively. In order to guarantee the reliability of findings, the project focuses on the interpretability of SHAP with feature-level explanations and Grad-CAM with highlighting the relevant areas in CT scans.

The comparative analysis of various models in terms of accuracy, precision, recall and F1-score to find out the most effective methods is also in the scope. Although the present work has an orientation on the development and assessment of models, the larger vision is to be incorporated into a wider clinical decision support system, which can help physicians with real-time diagnoses.

CHAPTER 2

LITERATURE REVIEW

2.1 Existing Approaches for Lung Cancer Prediction

The necessity to diagnose lung cancer at earlier and more reliable stages has caused the explosion of interest in prediction and classification of this type of cancer in the research community during the past several years. The application of machine learning algorithms, such as Logistic Regression, Support Vector Machines (SVM), Decision Trees, and Random Forests, to clinical data sets has been studied by many researchers. These datasets could include valuable patient information such as age, smoking, lifestyle and medical history, which are significant risk factors of lung cancer.

The algorithms mentioned above have already yielded promising outcomes with regard to identifying high-risk patients and classifying the cases as cancer-inducing or non-cancer-inducing with a high degree of accuracy. Their ability to work with structured tabular data and highlight underlying associations has made them useful risk assessment and clinical decision support tools. Their performance, however, is strongly dependent on the size, quality and diversity of the data and it is therefore necessary to test them on larger and more representative samples.

Meanwhile, deep learning has made a significant leap in the medical image analysis, especially CT scan. Convolutional Neural Networks (CNNs) and their enhanced forms, such as DenseNet, ResNet, and EfficientNet have demonstrated good performances as they are trained to learn hierarchical features of images automatically. These models can be used to detect and categorise the nodules in the lungs as benign, malignant or normal with high accuracy compared to the traditional methods. They are also able to identify small trends in pictures that human experts might overlook.

In addition, ensemble learning algorithms, including XGBoost and gradient boosting, are now employed to aggregate forecasts of multiple classifiers to enhance resilience and predictive power further. In sum, these developments show the potential of artificial intelligence to revolutionize traditional methods of diagnostics, reduce the rate of diagnostic errors, and finally help to save more lives by making the process of identifying lung cancer faster, more precise, and non-invasive.

Table 2.1 Literature Review

Ref. no	Dataset	Result	Methods Used	Limitations
[1]	LIDC-IDRI (Lung Image Database Consortium and Image Database Resource Initiative) LIDC-IDRI-DA-RAD (a variant of LIDC-IDRI with additional radiologist annotations)	Accuracy: 98% AUC (SVM): 0.9875 Training Accuracy (KNN): Lower than 98% (exact not given)	Used ML on 309 samples (16 features); removed age, gender, anxiety after correlation. Trained LR, SVM, KNN; evaluated using accuracy and AUC.	Limitations include small self-reported dataset which may introduce bias and limit generalizability. The study uses binary classification without considering cancer stages. Important factors such as genetics and environmental exposure are not included. KNN model shows overfitting which affects reliability.
[2]	Patient symptom and medical history data collected through questionnaires covering symptoms like cough, wheezing, yellow fingers, fatigue, etc. CT scan images obtained for patients classified as medium or high risk based on questionnaire analysis.	Random Forest Accuracy: 96.11% XGBoost Accuracy: 95.92% AdaBoost Accuracy: 95.74%	Used Random Forest, AdaBoost, XGBoost for lung cancer prediction. Handled imbalance with SMOTE and improved performance using Bayesian hyperparameter tuning.	Results depend on dataset quality and diversity as limited data may affect generalizability. Validation in real-world clinical settings is still required. Some important patient factors such as genetics and environment may be missing from the dataset.

[3]	BIR Lung Dataset LIDC-IDRI (Lung Image Database Consortium and Image Database Resource Initiative)	XGBoost Accuracy: 96.92% LightGBM Accuracy: ~93.5% AdaBoost Accuracy: ~92.3% Logistic Regression Accuracy: 67.41%	Collected data from 5000 patients (25 features, 5 hospitals). Applied preprocessing, feature selection (RapidMiner), 10-fold CV. Evaluated XGBoost, LightGBM, AdaBoost, LR, and SVM.	Data was manually collected from a specific region which may limit generalizability. Only structured tabular data was used without imaging or genomic inputs. Deep learning techniques such as CNNs were not applied. Real-time clinical deployment was not evaluated.
[4]	Lung Cancer Dataset from the UCI Machine Learning Repository LIDC-IDRI (Lung Image Database Consortium and Image Database Resource Initiative)	KNN Accuracy: 92.86% Bernoulli NB Accuracy: 91.07% Gaussian NB Accuracy: 91.07% Other Models (SVM, RF, XGB, MLP): Accuracy between 85% – 89%	Used 310 patient records (16 features) from Kaggle. Pre-processed data (deduplication, encoding, gender-wise analysis). Applied 10-fold CV (80/20 split). Compared LR, NB (Gaussian/Bernoulli), SVM, RF, KNN, XGBoost, Extra Tree, AdaBoost, MLP, and two ensembles (XGB+ADA, Voting).	The dataset size was limited to 310 samples which affects generalizability. Only tabular data was used without imaging or genomic features. Attribute correlations were moderate and results may differ on real-world EHR data. Future work should explore larger datasets and advanced ensemble methods for validation.
[5]	Cleveland Hospital Records National Lung Screening Trial (NLST) Prostate, Lung, Colorectal, and	Logistic Regression: Accuracy: 52% Precision: 52% Recall: 57% F1 Score: 54%	Used 3000 samples from Cleveland, public health, and NLST/PLCO datasets. Features: demographic, lifestyle, clinical	Models showed low accuracy due to a small and noisy dataset. Important features may be missing and class

	Ovarian (PLCO) Cancer Screening Trial	Random Forest Accuracy: 50% XGBoost Accuracy: 49%	symptoms. Applied LR, RF, XGBoost with preprocessing and encoding. Evaluated via accuracy, precision, recall, and F1 score.	imbalance could be present. Advanced models underperformed likely because of limited data quality. Imaging and genetic data were not included.
[6]	The study used a proprietary dataset of 10,000 patient records with clinical and radiographic data for lung cancer prediction.	KNN: Accuracy: 95.0% Precision: 90.5% Recall: 93.8% F1-Score: 92.1% Decision Tree Accuracy: 87.8% Random Forest Accuracy: 89.2% SVM Accuracy: 90.4%	Used 10,000 patient records with demographic, lifestyle, medical, and radiographic features. Applied imputation, IQR outlier removal, Z-score normalization. Split: 50% train, 15% val, 15% test. Used KNN (k=5, Euclidean).	KNN is computationally intensive on large datasets, sensitive to irrelevant features, and its performance can degrade in high-dimensional spaces. Its effectiveness relies heavily on proper feature selection.
[7]	National Lung Screening Trial (NLST) Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial LIDC-IDRI	KNN Accuracy: 99% SVM Accuracy: 96.3%	Used 1000 records with 22 features (demographic, lifestyle, medical). Applied SVM, KNN, DT, RF with Grid Search tuning. Used XAI methods: decision boundaries, LIME, and tree visualization.	Model performance depends on dataset quality and available features. Bias may arise from underrepresented demographics. High accuracy could indicate overfitting. There is a risk of false positives or negatives. Age and gender showed low feature importance.
[8]	SEER (Surveillance, Epidemiology, and End Results) dataset –	XGBoost: Accuracy: 94.42% Precision: 95.66%	Applied XGBoost, LightGBM, AdaBoost, and Bagging on Kaggle dataset (309 samples,	The small dataset size of 309 samples may limit generalizability. Reliance on survey

	comprising 1,000 patient records. Data World dataset – comprising 1,000 patient records. Lanzhou University dataset – comprising 277 patient records. HRQOL (Health-Related Quality of Life) dataset – comprising 809 patient records. UCI Lung Cancer dataset – comprising 32 patient records.	Recall: 94.46% F1-Score: 94.74% AUC: 98.14% LightGBM Accuracy: 92.56% LightGBM AUC: >97% AdaBoost Accuracy: 90.70% AdaBoost AUC: >97% Bagging Accuracy: 89.76% Bagging AUC: 95.30%	16 features).Performed preprocessing, handled imbalance with SMOTE, and used 10-fold CV. Evaluated with accuracy, precision, recall, F1-score, and AUC.	data introduces potential bias or inaccuracies. Only symptom-based features were used, excluding key factors like genetics and environmental exposures. Clinical validation is needed to confirm effectiveness in real-world settings. Results may not generalize to other populations without additional studies
[9]	A Kaggle-based lung cancer dataset containing patient symptom and lifestyle information	Random Forest: Accuracy: 92.3% Precision: 97% F1-Score: 95% Decision Tree Accuracy: 91% Logistic Regression Accuracy: 89.7% SVM: Accuracy: 88.5% Recall: 95.6%	Used lung cancer dataset (13 features) from Kaggle. Implemented four machine learning algorithms Random Forest, SVM, Logistic Regression, and Decision Tree on Google Collab using Python. Evaluation was based on accuracy, precision, recall, and F1-score (harmonic mean).	The dataset lacked size and diversity. Only traditional machine learning models on tabular data were applied without deep learning or imaging-based methods. SVM performed poorly with noisy inputs and logistic regression was limited by its linearity assumption. External validation or cross-validation was not reported
[10]	LIDC-IDRI (The Lung Image Database Consortium image collection)	CNN Accuracy: up to 97.2% ANN Accuracy: up to 96.67% SVM Accuracy: up to 96.7%	Review paper summarizing ML/DL methods for lung cancer detection. Covers preprocessing (filtering, CLAHE),	The study is review-based and does not include new dataset evaluation . Most referenced

	TCIA datasets (e.g., “RIDER-LUNGCT-SEG”, “QIN-LUNGCT-SEG”) Kaggle lung cancer datasets (CT or clinical) LC25000 for histopathology images		segmentation (watershed, k-means), feature extraction (GLCM, PCA, CNN), and classification (SVM, ANN, CNN). Uses datasets: LIDC-IDRI, LUNA16, Kaggle DSB.	methods rely on large labelled datasets and may suffer from overfitting. Model comparisons lack unified benchmarking, limiting consistent performance assessment.

Description: The table provides a summary of lung cancer prediction studies with accuracies ranging between 50-99% across all ML/DL techniques with the weaknesses largely attributed to small data and little validation.

2.2 Identified Limitations in Prior Work

Although the application of machine learning and deep learning methods in lung cancer prediction has made several strides, existing research has a number of limitations. The interpretability of most research is also a significant weakness since it does not provide information about how it is performed to make predictions. This is a hindrance to clinical uptake as medical practitioners need transparency and explanations behind AI-aided diagnoses.

A large portion of studies also use small or unbalanced data and restrict the applicability and strength of the models to a wide range of patients. Overfitting is a serious threat in the situation of deep learning models, particularly when the size of datasets is inadequate.

Although certain methods can produce high accuracy, they do not balance the performance measures like precision and recall and as a result, make errors to detect malignant cases, which is important in early-stage diagnosis.

Also, little focus has been on multimodal integration where most publications have focused on either textual clinical data or imaging data and not a combination of both, which would enable more comprehensive prediction. Ensemble methods, though potentially effective, are not yet the focus of much research, and those techniques of feature selection used in past reports tend to be simplistic, which can easily cause overlook of significant predictors. All these limitations point to the gaps, which need to be sealed in order to enhance the reliability and clinical applicability of AI-based lung cancer prediction systems.

2.3 Research Gap

The literature review reveals the definite gap in the progress of predictive systems that can be characterized by the combination of the three factors: the accuracy, interpretability, and multimodal analysis. Although there are high-performance lung cancer prediction models, they are black-box, making them less credible in a clinical setting. Equally, this application of either the clinical datasets or the imaging datasets alone inhibits the breadth of diagnosis, as the single type of data would not yield every form of complexity of lung cancer risks and manifestations.

In addition, there are a limited number of studies that combined explainable AI models like SHAP or Grad-CAM to give both feature-based and visual explanations of predictions. Such a lack of interpretation limits the possibility of AI systems to be embraced and embraced by healthcare providers. The other gap is the lack of successful use of ensemble and hybrid methods that would enhance the strength and accuracy of forecasts by capitalizing on the strengths of various models.

To fill in these gaps, a holistic system is needed, not only to deliver high predictive performance but to provide transparency and reliability through explainable methods and using a combination of text and imaging data.

CHAPTER 3

SYSTEM DESIGN AND METHODOLOGY

3.1 Workflow of the Proposed System

The suggested lung cancer prediction system will be structured as a multi-step workflow that will combine both the clinical text-based information and CT scan image data to obtain reliable and interpretable outcomes. The workflow starts with the process of data acquisition, when structured clinical records and medical images are obtained on the basis of publicly available datasets.

This is then succeeded by preprocessing which entails cleaning and normalization followed by feature extraction in order to make the data suitable to be analysed. In the case of a clinical dataset, random forest or logistic regression machine models are trained to both learn the important trends and associations between risk factors.

In the case of CT scan images, the DenseNet121, ResNet50 and EfficientNet convolutional neural network architectures are used to automatically train deep visual features that could differentiate between benign and malignant and normal cases. After the models are trained, their performance is evaluated based on performance measures such as accuracy, precision, recall, F1-score and ROC-AUC to ascertain how effective the models are.

In order to increase trust and interpretability, explainability methods are incorporated into the workflow. SHAP is run over the clinical data models to describe the contribution of the individual attributes of smoking, age, and throat discomfort, and Grad-CAM is applied to CNN models to produce heatmaps that indicate important areas in CT scans that affect the decision. This workflow makes the system not only accurate but also transparent and thus meets the requirements of being used in the clinical settings.

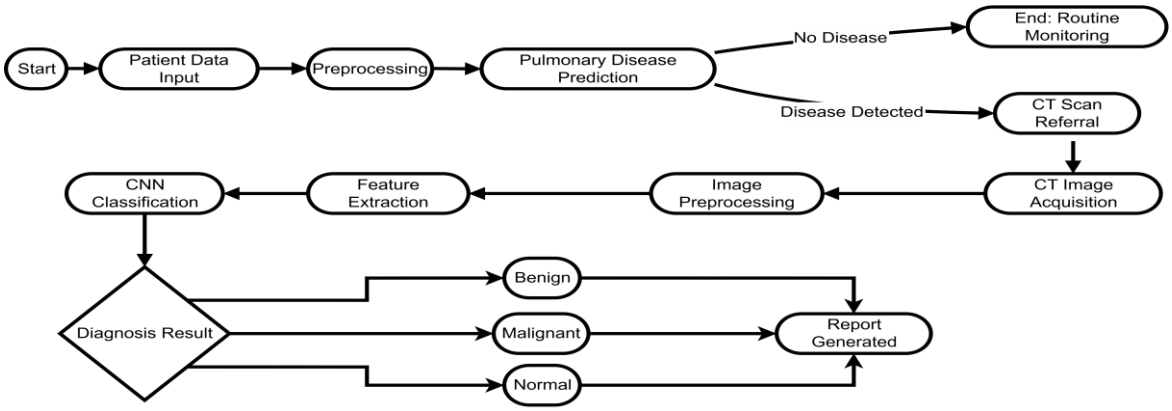

Fig 3.1 Workflow Diagram

Description: The figure is a presentation of a pulmonary disease prediction system. Patient data is pre-processed to predict disease; when a disease is identified, CT scans are pre-processed, features are extracted and classified using CNN to benign, malignant or normal and a report is generated.

3.2 Dataset Description

The research employs two different datasets in order to view the lung cancer diagnosis in both textual and the imaging point of view. The primary dataset is a table, text-based and 5000 patients with 18 medical attributes. These characteristics are demographic, e.g. age and gender, lifestyle, e.g. smoking history and alcohol consumption, and clinical, e.g. oxygen saturation, breathing problems, discomfort in the throat, and family medical history.

The data set offers a holistic perspective of the clinical risk factors that are related to lung cancer. The second dataset comprises CT scan images of three categories; benign, malignant, and normal. These images provide visual data about the abnormalities in the lungs and they can be regarded as a critical element that aids in the identification of cancer at the structural level.

The textual and imaging datasets complement each other and offer a chance to use a multimodal analysis, which makes the system stronger in prediction.

3.3 Data Preprocessing and Visualization

Data preprocessing is a critical step to ensure the quality and reliability of the inputs used in model training. For the textual dataset, preprocessing involves handling missing values, normalizing numerical attributes, encoding categorical variables, and balancing the dataset to reduce bias.

Exploratory data analysis (EDA) is performed to visualize patterns and correlations within the dataset. For example, visualization of smoking history against pulmonary disease prevalence reveals a strong positive correlation, indicating smoking as a major risk factor. Similarly, age group analysis shows higher disease occurrence among individuals in the 30, 50–60, and 80+ age ranges.

Other visualizations highlight the prevalence of risk factors such as pollution exposure, mental stress, and long-term illness. For CT scan images, preprocessing includes resizing, normalization, and augmentation techniques such as rotation and flipping to increase dataset diversity and improve the robustness of deep learning models. Visualization tools such as

Grad-CAM are later applied to CT images to highlight regions of clinical importance, reinforcing interpretability. The system performance is critical in the selection of models. In the case of clinical data, classical machine learning algorithms, which include Random Forest, Logistic Regression, Gradient Boosting, AdaBoost, Support Vector Machines and K-Nearest Neighbours, were tested. Random Forest proved to be the most successful model because it can address non-linear relationship, overfitting is mitigated, and its results are balanced in terms of accuracy, precision, and recall.

In regard to image-based data, we chose deep-learning models, namely, DenseNet121, ResNet50, and EfficientNet-B0 CNN architectures. DenseNet121, with its masses of densely connected layers, was the most accurate and stable in the classification of CT scans into benign, malignant, as well as normal scans.

The training of the model was performed using TensorFlow and Keras systems, and hyperparameter tuning was used to make the model more efficient. To achieve robustness, cross-validation methods were utilized, and classification performance was measured with respect to different classes by use of confusion matrices.

3.4 Model Selection and Training

Model selection is an important factor in ensuring high system performance. In the case of clinical data, a variety of traditional machine learning models were investigated, amongst which are Random Forest, Logistic Regression, Gradient Boosting, AdaBoost, Support Vector Machines and K-Nearest Neighbours. Random Forest was the best of these because it was able to capture non-linear relationships, it did not overfit, and it produced balanced results in terms of accuracy, precision, and recall. Its ensemble architecture enabled it to provide consistent and reliable predictions across structured clinical data, qualifying it as a potent candidate to detect early lung cancer.

In the case of image-based data, deep learning models have been favoured with DenseNet121, ResNet50 and EfficientNet-B0 CNN architectures being tested. DenseNet121 has been shown to be the most accurate and stable model in classifying CT scans as benign, malignant, and normal due to its fully connected layers, which encourage reuse of features and efficient gradient flow.

Training was carried out using TensorFlow and Keras frameworks and hyperparameter tuning was performed to ensure maximum performance. To achieve a robustness, cross-validation methods were used, and the confusion matrices were examined to give extensive

information concerning the classes-wise performance, which additionally contributes to the credibility of the predictive system.

Table 3.1 Model Evaluation for Textual Dataset

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	CV Mean
Random Forest	**90.80%**	**90.78%**	**90.80%**	**90.79%**	**92.20%**	**90.50%**
Gradient Boosting	89.90%	89.88%	89.90%	89.89%	92.37%	90.08%
AdaBoost	88.90%	88.93%	88.90%	88.91%	91.92%	88.80%
SVM (Optimized)	88.30%	88.59%	88.30%	88.36%	92.53%	88.07%
Logistic Regression	87.50%	88.21%	87.50%	87.59%	92.63%	87.90%
Neural Network	86.50%	86.46%	86.50%	86.43%	89.71%	86.05%
Naive Bayes	85.30%	85.77%	85.30%	85.39%	88.74%	86.20%
K-Nearest Neighbors	85.10%	85.24%	85.10%	85.14%	90.07%	85.27%
Decision Tree	80.20%	80.16%	80.20%	80.18%	79.37%	83.47%

Description: Table 3.0 shows the analysis of the lung cancer prediction models on textual data and outlines the accuracy and limitations of various ML/DL methods.

Table 3.2 Model Evaluation for CT-Scan Dataset

Model	Accuracy	Precision	Recall	F1-Score
DenseNet121	**99.00%**	**99.45%**	**99.44%**	**99.44%**
ResNet50	**99.00%**	98.92%	98.89%	98.89%
EfficientNet-B0	97.00%	97.55%	97.50%	97.50%

Description: Table 3.1 indicates that DenseNet121 and ResNet50 have the highest accuracy of 99 percent and the third model EfficientNet-B0 has a slightly lower accuracy of 97 percent.

3.5 Explainability using SHAP and Grad-CAM

Interpretability and explainability are fundamental aspects of this project, as they ensure that the predictive models operate based on meaningful, clinically relevant information rather than random or spurious correlations, which is essential for medical applications where trust and transparency are critical. For the machine learning models trained on clinical data, SHAP (SHapley Additive exPlanations) was employed as a powerful tool to quantify the contribution of each feature to the model's predictions. SHAP, grounded in cooperative

game theory, assigns each feature a Shapley value that represents its marginal contribution to the prediction, allowing for both local interpretability—explaining why the model predicted a particular outcome for an individual patient—and global interpretability—understanding which features are generally the most influential across the entire dataset. The analysis revealed that smoking history, the presence of throat discomfort, breathing difficulties, and exposure to environmental pollution were the most critical determinants of lung cancer risk, consistently showing strong positive contributions to malignancy predictions. In contrast, variables such as gender and alcohol consumption exhibited minimal influence on the predictions, suggesting that these factors were less significant in the context of this dataset. By providing these insights, SHAP not only validated known clinical risk factors but also allowed for a quantitative assessment of their relative importance, enhancing the transparency and trustworthiness of the clinical data models.
For the deep learning models trained on CT scan images, which are inherently more complex and difficult to interpret due to the high-dimensional nature of image data, Grad-CAM (Gradient-weighted Class Activation Mapping) was used to provide visual explanations. Grad-CAM generates heatmaps by computing the gradients of the target class with respect to the final convolutional layers, thereby highlighting the regions of the image that most strongly influenced the model's prediction. The resulting heatmaps consistently focused on areas of abnormal tissue, nodules, or lesions in the lungs in cases of malignancy, demonstrating that the network learned to recognize medically meaningful structures rather than relying on irrelevant or noisy patterns. This spatial visualization not only provides interpretability for radiologists and clinicians but also offers an additional layer of validation for the model, showing that its decisions are aligned with clinical expectations.

By integrating SHAP for clinical data and Grad-CAM for imaging data, this project achieves a comprehensive explainability framework, combining quantitative feature importance with spatial visualization. This dual approach ensures that both the tabular and imaging models are interpretable, increasing confidence in their predictions, and providing insights that are both clinically relevant and actionable. Ultimately, the use of these explainable AI methods strengthens the reliability of the predictive system, facilitates clinical validation, and supports informed decision-making in the diagnosis and management of lung cancer.

CHAPTER 4

IMPLEMENTATION

4.1 Explainability using SHAP and Grad-CAM

When the clinical dataset size was sufficient to form a sample of 5,000 patient records, with 18 attributes, the machines learning models were applied and systematically tested to determine which model would yield the best results to predict lung cancer. This data set contained very important patient data like age, smoking habits, lifestyle habits, respiratory symptoms, etc and was therefore appropriate in structured learning techniques. The regression models used were Logistic Regression, Random Forest, Gradient Boosting, AdaBoost, K-Nearest Neighbours (KNN) and Support Vector Machines (SVM). All the models were trained and tested on stratified train-test split, ensuring that the number of healthy and cancer-positive cases represented in the training and validation sets were equal. This was significant to balance the classes and prevent biased predictions.

Out of all the tested models, it was found that the most successful model is the Random Forest classifier, with the highest overall accuracy and balanced precision and recall values by class. It was best suited to this data, as it has a built-in capacity to model non-linear relationships in the data and it is also less susceptible to overfitting (through the use of decision trees) since it is an ensemble model. Although the results of Logistic Regression were reliable and easy to interpret, its predictive power was slightly lower than that of Random Forest, particularly in more complex interaction terms between variables. Conversely, other models, including SVM and KNN, demonstrated weaknesses when using this dataset, either because of inefficiency during their calculation, or because of reduced stability in the classification. SHAP (SHapley Additive exPlanations) was used to interpret the Random Forest model to provide contributions level at the feature level. SHAP has pointed out that the most influential attributes in the prediction of lung cancer were smoking history, sore throat, difficulty in breathing and oxygen saturation. Not only did these insights confirm previously known medical risk factors but they also enhanced confidence in the model by increasing the transparency of its decision-making process.

4.2 Deep Learning Models for CT Scan Data

The imaging dataset was subjected to deep learning techniques, and their main attention was on Convolutional Neural Networks (CNNs), which have so far shown impressive success in medical image classification problems. DenseNet121, ResNet50, and EfficientNet-B0 were also tested as these are the most popular architectures with high performance when dealing with complex image data. DenseNet121 has the highest overall performance, with an outstanding 99 percent accuracy, a high precision score, recall, and F1-score. The high efficiency of DenseNet121 might be explained by its dense connectivity structure that facilitates not only feature re-use but also effective gradient flow through layers. Such structural benefit allows the network to learn rich hierarchical properties, and it avoids vanishing gradient problems during training, which eventually enables it to learn subtle patterns in CT scans much better.

By contrast, ResNet50 also demonstrated promising results, which is attributed to its residual connections, though with comparatively lower precision and F1-scores, which reflects some limitations when dealing with the variability of the lung nodules appearances. Although computationally efficient and with a high accuracy of 97%, EfficientNet-B0 was relatively less predictively consistent than the two other models.

In order to enhance reliability and interpretability of CNN predictions, the Grad-CAM (Gradient-weighted Class Activation Mapping) was used. In this visualization, the areas on CT scans that were most useful in the decisions made by the model were emphasised to create heatmaps where they were most useful. More importantly, in malignant cases, these regions of interest were always associated with abnormal lung nodules, which validates the claim that the model was utilizing clinically relevant features and not random artifact. This not only confirmed the diagnostic usefulness of the predictions, but also resolved a key issue with the use of deep learning, which is the lack of transparency.

Grad-CAM enhanced clinical trust in the AI system by offering explainability using visual evidence, and this supports the claim that the system could be a valuable decision-support tool to help radiologists with the early-stage detection and classification of lung cancer.

4.3 Backend and Frontend Development

A more basic backend was created to facilitate the practical use, based on FastAPI with the help of which it is possible to deploy trained models in the form of RESTful predictions services. The backend processes the requests, preprocesses the input data, performs predictions with the help of the right models and gives out the results and interpretability insights. In the case of clinical data, the API will give predicted SHAP values and labels on the prediction of features. In the case of CT scans, it gives predicted classes and visual interpretability (Grad-CAM heatmaps). The user-friendly interface was built at the front end to enable users or clinicians to input patient records or CT scan images, display the results of the prediction and the logic the model used to make the choice. This integration shows the practical viability of the project and preconditions the construction of a fully-developed clinical decision support system in the future.

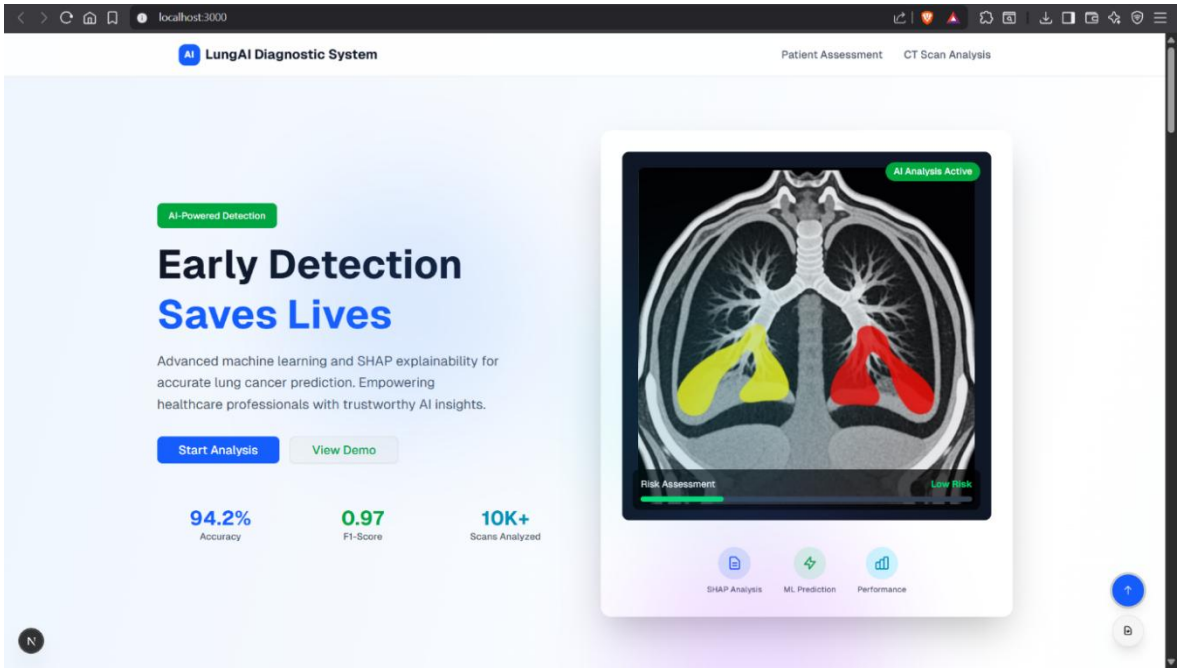

Fig 4.1 Home Page UI

Description: Figure 4.0 shows the home page view of the LungAI Diagnostic System with AI-powered early detection highlights, CT scan images, and analysis, demo, and performance insights options.

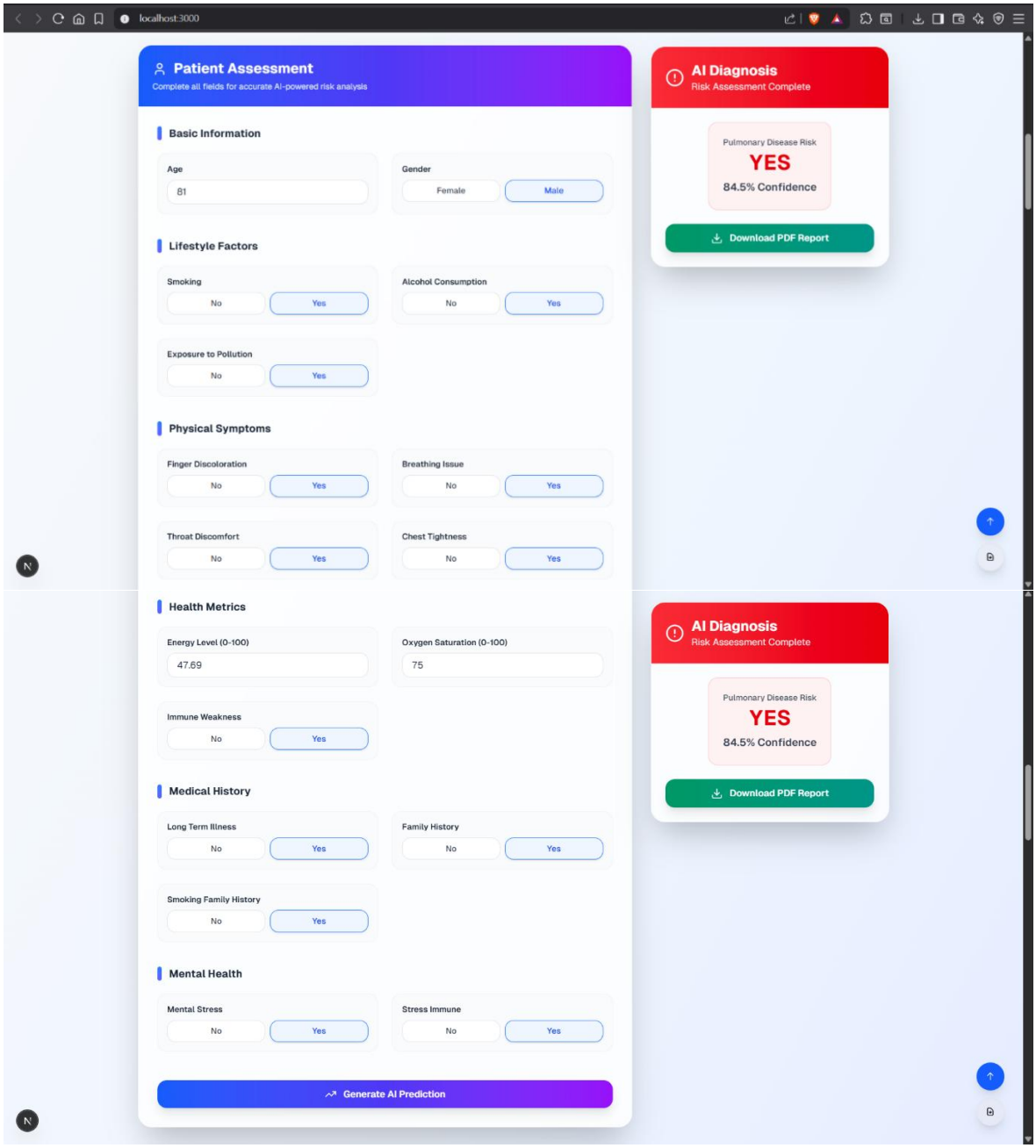

Fig 4.2 Pulmonary Disease Analysis UI

Description: Figure 4.1 demonstrates the pulmonary disease analysis interface, where the patient enters his or her personal, lifestyle, and medical data. The system then creates an AI based diagnosis with a downloadable report option.

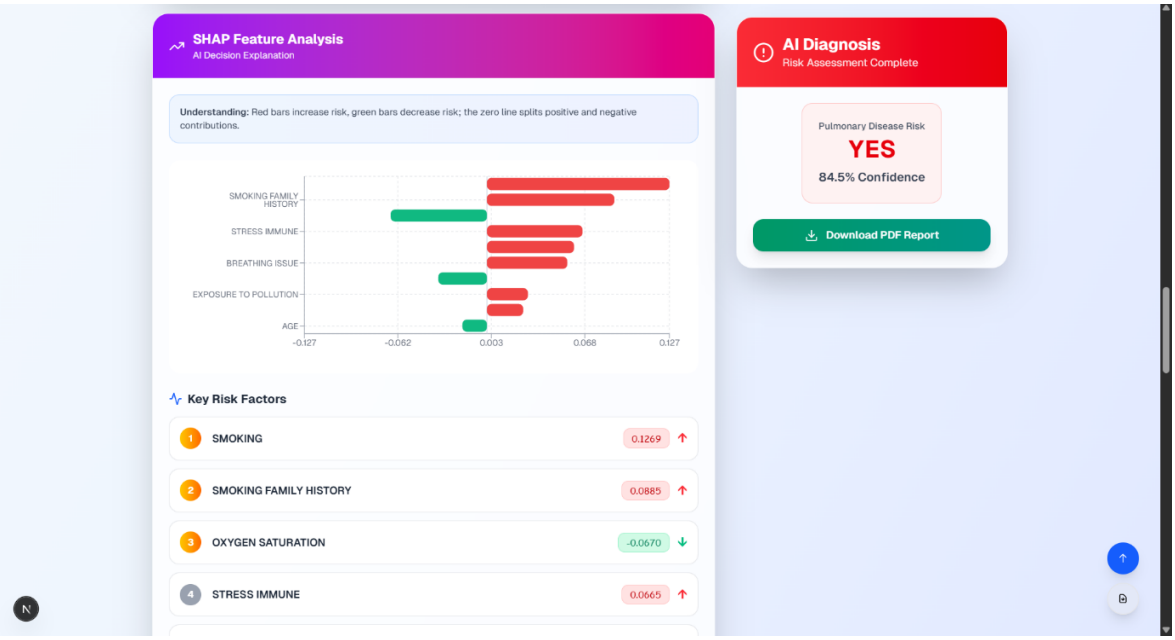

Fig 4.3 SHAP Analysis on Pulmonary Disease

Description: SHAP bar plot describes what characteristics were used to produce this prediction of pulmonary disease: red bars raise predicted risk and green bars lower it, and Smoking and Smoking Family History raise and reduce risk, respectively.

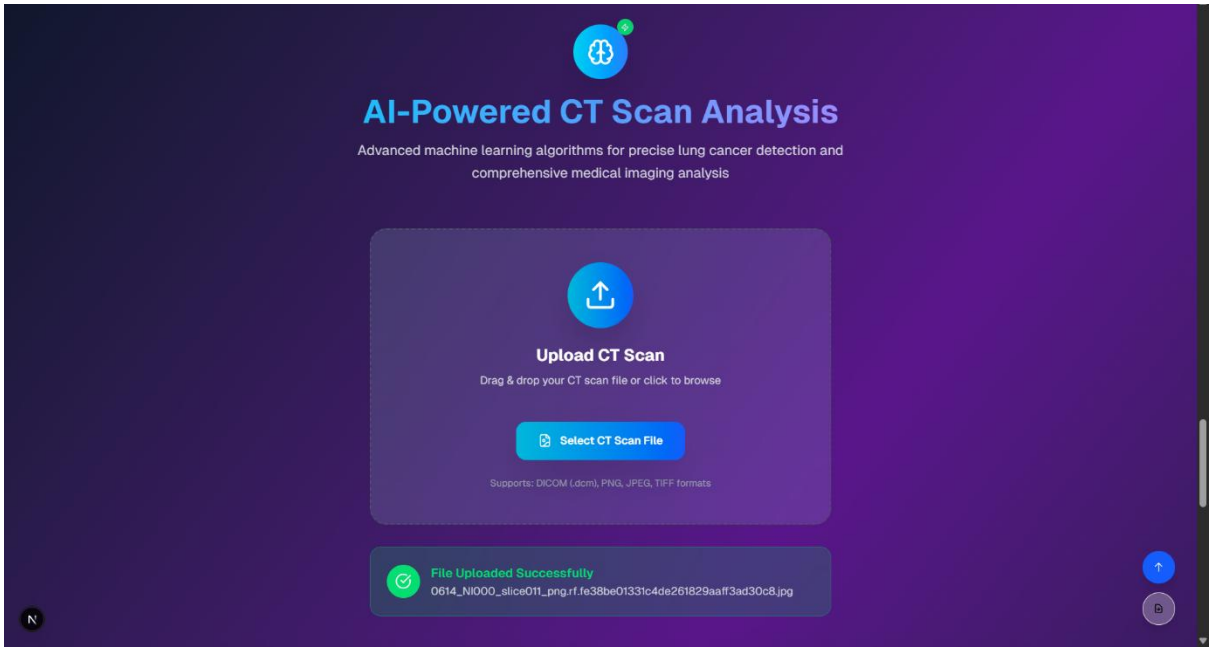

Fig 4.4 CT Scan Analysis

Description: An AI-based CT Scan Analysis interface has a centre card with support to drag-and-drop or button-pressed upload of DICOM, PNG/JPG, and TIFF scans to automatically evaluate them.

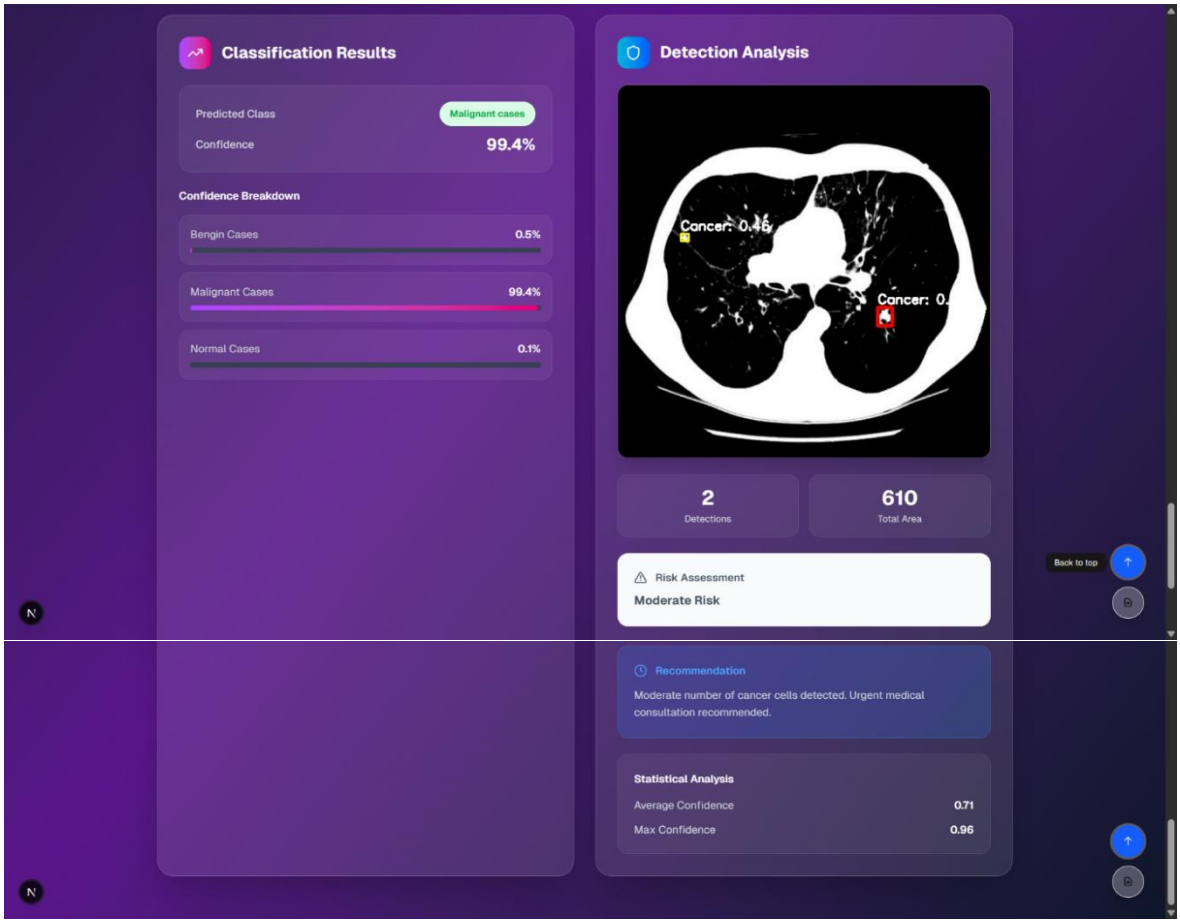

Fig 4.5 Result of CT Scan Analysis

Description: Figure 4.4 CT Scan Analysis Result depicts a medical AI interface that shows the results of the lung CT scan results with 99.4% accuracy of the malignant classification. The system identified 2 suspicious areas with a total of 610 area, which was proposed to have emergency medical consultation of the moderate-risk case.

4.4 Tech Stack

This project was implemented using a highly chosen combination of tools and technologies that combined to guarantee both efficiency, accuracy, and scalability. The main programming language was Python because it is a general programming language with a solid data science ecosystem. Regarding machine learning and deep learning tasks, the project heavily used scikit-learn, TensorFlow, and Keras, which allowed creating a powerful model, training it, and fine-tuning it. OpenCV was used in image preprocessing and augmentation, allowing to increase the variability and the quality of the CT scan data, minimizing the chances of overfitting. To explain results, SHAP (SHapley Additive exPlanations) was used to analyse the clinical data models at the feature level to determine the attributes that made the strongest contribution to the prediction. Concurrently, the CNN

models were subjected to Grad-CAM (Gradient-weighted Class Activation Mapping) to highlight important areas on CT scan images that the network relied on during the decision-making process, hence providing a transparent layer of interpretability to the medical experts.

On the deployment side, FastAPI was used to develop the backend due to its simplicity, speed, and capability to perform ML model inference effectively. The frontend was implemented with standard web technologies and a basic yet useful interface, and this interface was linked to the backend services through REST APIs to guarantee the proper flow of communication between the components. GitHub was utilized to follow the changes, administer versions, and organize activities to work together on the development. Also, the project planning and tracking was done in the form of Gantt charts which facilitated in dividing the work, establishing deadlines, and ensuring that milestones were being met on time. With this complete technology stack implemented, the project was able to achieve a predictive lung cancer system that placed the focus not only on accuracy and interpretability but also on scalability and deploy ability, making it a well-balanced solution and practical to use in real-world applications.

CHAPTER 5

RESULTS AND EVALUATION

5.1 Performance Metrics

To evaluate the performance of the developed models, standard classification metrics were used, including accuracy, precision, recall, F1-score, and ROC-AUC. Accuracy provided a measure of overall correctness, while precision and recall ensured that the model was not biased toward any particular class. The F1-score, being the harmonic mean of precision and recall, was particularly useful in balancing false positives and false negatives, which is critical in medical applications. ROC-AUC scores were also considered, as they provide insight into the trade-off between sensitivity and specificity. By employing these metrics, a holistic assessment of the models was carried out, ensuring that the best-performing systems were not only accurate but also clinically reliable.

5.2 Results on Textual Dataset

For the clinical dataset consisting of 5,000 records with 18 attributes, Random Forest outperformed other machine learning models. The model achieved high accuracy and balanced performance across precision and recall. The confusion matrix indicated that the Random Forest classifier correctly identified the majority of both cancer-positive and cancer-negative cases, with only 95 misclassifications overall. Logistic Regression also performed well but showed slightly lower recall, which is less desirable in a medical context where missing positive cases can be critical. SHAP analysis further revealed that smoking history was the single most influential attribute, followed by throat discomfort, breathing issues, and oxygen saturation. These results not only validated the predictive strength of the model but also aligned with established medical knowledge, reinforcing the reliability of the findings.

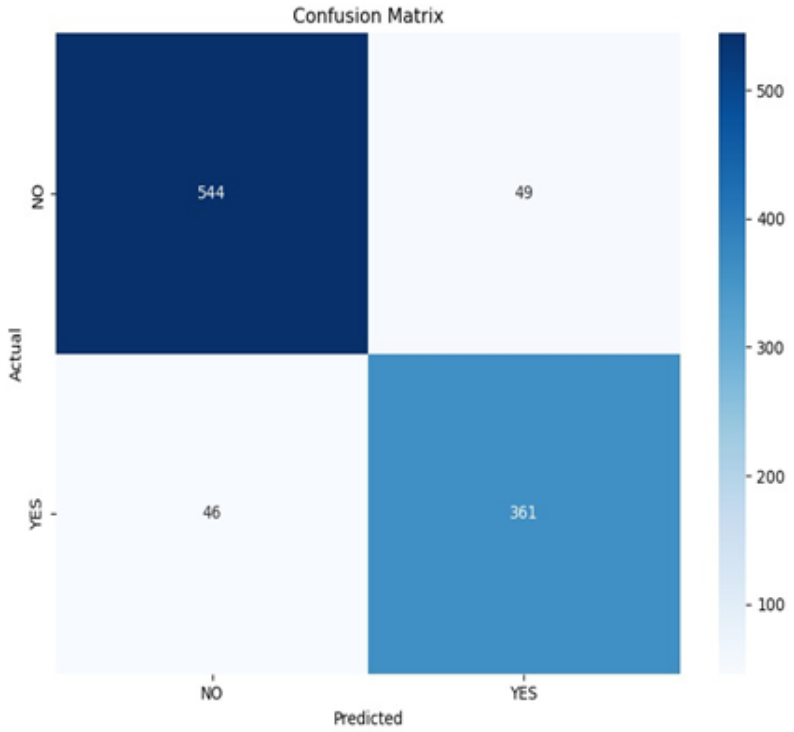

Fig 5.1 Confusion Matrix of Textual Dataset

Description: In Figure 5.0 a binary classification model is being used and the number of true negatives of 544 and true positives of 361 show that it is making correct predictions. The model is very accurate with just 49 false positives and 46 false negatives of very low misclassification errors.

5.3 Results on CT Scan Dataset

For the imaging dataset, deep learning models demonstrated exceptional performance, with DenseNet121 emerging as the best-performing architecture. It achieved an accuracy of 99%, precision of 99.45%, recall of 99.44%, and an F1-score of 99.44%. The confusion matrix showed near-perfect classification, with only two misclassifications across 220 total cases. ResNet50 also performed well, achieving similar accuracy but slightly lower precision and recall values, while EfficientNet-B0 achieved around 97% accuracy, making it competitive but less robust than DenseNet121. Grad-CAM visualizations provided additional interpretability by highlighting the regions in the CT scans that contributed most to the model's predictions. For malignant cases, the highlighted areas consistently corresponded to abnormal lung regions, confirming the clinical validity of the model's reasoning.

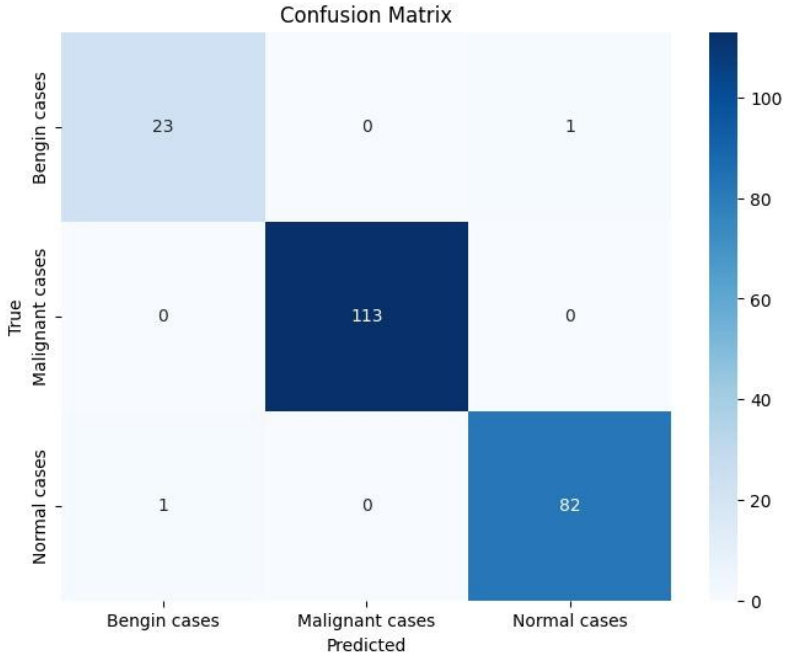

Fig 5.2 Confusion Matrix of CT-Scan Dataset

Description: Figure 5.1 shows a 3-class confusion matrix of CT-scan dataset with great performance with 23 benign, 113 malignant, and 82 normal cases correctly classified. Model has a near-perfect accuracy of 2 misclassifications in all the three diagnostic categories.

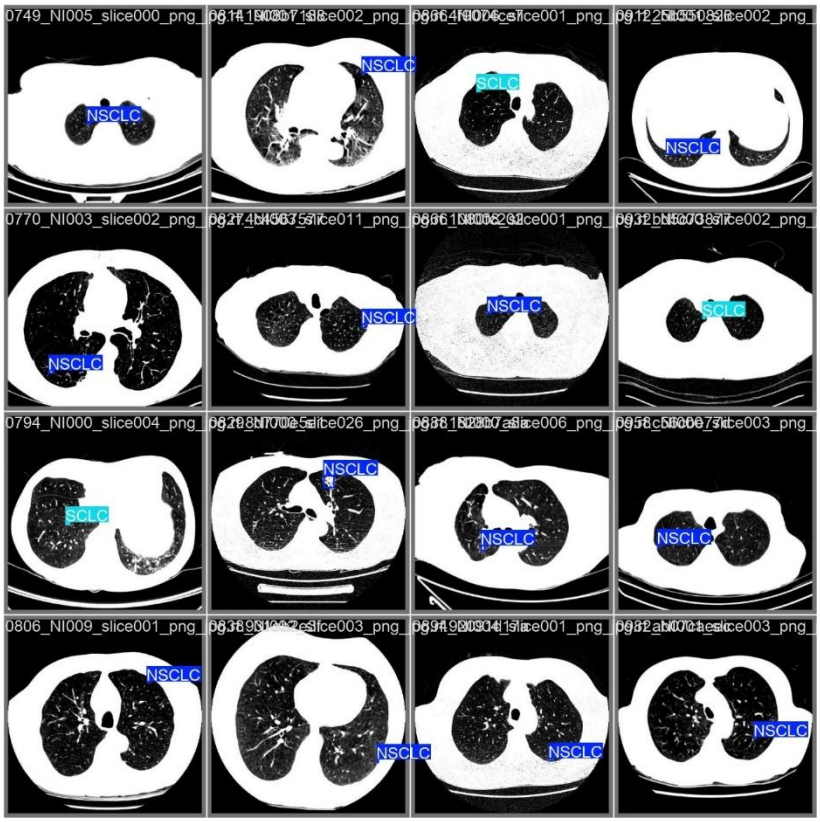

Fig 5.3 Results of CT-Scan images

Description: A grid of CT-scan lung images of the model with labels of NSCLC is shown in figure 5.3. The images illustrate different pathologies and abnormalities of the lungs observed in different patient cases in the dataset.

5.4 Comparative Analysis of Models

When comparing machine learning models for clinical data with deep learning models for imaging data, a clear distinction was observed in performance levels. While Random Forest provided strong and balanced results for textual attributes, the CNN architectures, particularly DenseNet121, significantly outperformed in image-based classification. This suggests that CT scan images contain highly discriminative features that deep learning models can capture effectively. However, the textual dataset offered valuable complementary insights, especially in identifying lifestyle and clinical factors such as smoking, pollution exposure, and long-term illness that may not be visible in imaging data alone. The results therefore emphasize the potential benefits of a multimodal system that integrates both textual and imaging data to achieve comprehensive predictions.

5.5 SHAP and Grad-CAM Explanations

The incorporation of SHAP and Grad-CAM greatly enhanced the interpretability of the results. SHAP values quantified the impact of each feature on model predictions, making it clear which clinical attributes were most responsible for a positive cancer classification. Smoking, throat discomfort, and breathing issues were consistently identified as the strongest predictors, aligning with established medical findings. Grad-CAM, on the other hand, visually highlighted abnormal regions in CT scans associated with malignancy, providing a direct link between the model's prediction and clinically observable patterns. Together, these explainability techniques transformed the models from black-box systems into interpretable tools, bridging the gap between artificial intelligence and clinical trust.

CHAPTER 6

DISCUSSION

6.1 Key Findings

The findings of this paper indicate the usefulness of combining machine learning and deep learning methods to predict lung cancer. A model with the highest performance in the textual data set was the Random Forest, with balanced accuracy, precision and recall, whereas DenseNet121 was the best in the CT scan image classification with almost perfect results across all measures. The joint SHAP-Grad-CAM added interpretability to the models, pointing out clinical features and image regions that had a strong impact on the predictions.

The present results confirm the relevance of using both quantitative clinical data and qualitative imaging data in order to reflect the complexity of the lung cancer diagnosis. The most significant clinical predictors were identified to be smoking, throat pain, breathing problems and exposure to pollution and CT scans were important sources of structural evidence of malignancy. Collectively, these results highlight the possibility of AI-based systems having a positive impact on the development and proper classification of lung cancer at an early stage of its progression.

6.2 Clinical Relevance of the Results

The clinical relevance of this project lies in its ability to provide not only accurate predictions but also interpretable insights that can assist healthcare professionals in decision-making. The high accuracy achieved by DenseNet121 demonstrates that deep learning can reliably distinguish malignant nodules from benign or normal cases, which could significantly reduce diagnostic errors and improve early detection rates. Meanwhile, the SHAP-based explanations for textual attributes allow doctors to understand which lifestyle and physiological factors contributed most to the prediction. This dual approach, combining predictive accuracy with transparency, bridges a critical gap between AI models and clinical adoption. Such a system can act as a decision-support tool, empowering doctors with an additional layer of evidence while ensuring that predictions are grounded in understandable and clinically relevant factors.

6.3 Limitations of Current Study

While the outcomes of this study are highly promising, several limitations must be acknowledged. The first limitation arises from the nature of the datasets used. Since the clinical and imaging data were obtained from publicly available repositories rather than local hospitals, the system's generalizability to real-world populations may be limited. Another significant limitation concerns interpretability. Although explainable AI techniques such as SHAP and Grad-CAM were integrated to make predictions transparent, there are still challenges.

SHAP explanations may become difficult to interpret as the number of features increases, and Grad-CAM heatmaps, while useful, do not guarantee absolute clinical precision. More importantly, as computer engineering students and not medical professionals, we are not qualified to definitively verify whether the highlighted regions in CT scans correspond to actual malignant tissues or benign structures.

While the models demonstrated high accuracy, only certified radiologists or oncologists can clinically validate whether these visual detections are correct. Furthermore, the project focused on analysing textual and imaging data independently, and the full integration of multimodal approaches remains a task for future development. Finally, the models were tested under controlled experimental conditions, and their performance in real-time clinical environments is yet to be validated. Addressing these limitations through collaboration with medical experts and real-world testing will be essential for the clinical deployment of this system.

CHAPTER 7

FUTURE WORK

7.1 Advanced Ensemble Methods

The other area that can be improved is the use of advanced ensemble techniques. Although Random Forest was incredibly successful with the clinical data and DenseNet121 was more successful on CT images, there is the danger of overfitting or underperforming in previously unexplained cases, as only one model is used. Single-model methods tend to fail when the properties of the data vary or when subjected to data variation that is not represented in training. This exposes them to decreased accuracy and reliability in actual implementations. Conversely, ensemble strategies are considered as the combination of the forecasts of a number of models so that the shortcomings of one model are offset by the strengths of another model. With this integration, the generalization is improved and the performance is more consistent over a wide range of data and patient population.

One can consider using ensemble methods including stacking, blending, and boosting in order to combine predictions of various machine learning algorithms or deep learning networks. An example of this is that, with stacking, the results of foundation learners such as Random Forest, XGBoost, or logistic regression can be combined via a meta-learner to produce a more adaptive and well-balanced model. On the same note, blending offers the flexibility of combining models based on weighted averages and boosting focuses on correcting errors committed by weaker learners. When these techniques are applied to clinical data, they can be used to dramatically enhance stability and robustness by making bias or variance dominate the outcomes less likely.

When applied to imaging data, DenseNet and ResNet may have complementary advantages since both architectures represent features in different manners and perform well in different situations. Combining them can provide increased accuracy as well as reduce error susceptibility in certain patient cases. Likewise, in the case of clinical data, a combination of Random Forest and gradient boosting algorithms may enhance predictive capabilities by exploiting the randomness of features, as well as, repeatedly correcting errors. These ensemble methods can minimize variance, misclassification and eventually improve the reliability of the system in practical medical conditions where robustness and accuracy are important.

CONCLUSION

This project shows that machine learning and explainable AI are effective to predict lung cancer using clinical and imaging data. Random Forest has demonstrated excellent performance on the clinical side, successfully capturing trends in patient records and lifestyle-related characteristics. In the case of imaging data, DenseNet121 performed better than other CNN models, achieving near-perfect accuracy in the classification of CT scans and thus proving its superiority in medical image analysis.

Clinical feature SHAP integration combined with CT image Grad-CAM introduced a much-needed interpretability dimension to the model, evidently indicating the attributes and regions that contributed to model predictions. Not only does this improve the reliability of results, but it also increases clinical trust, which makes the system more appropriate to possible applications in the real world.

The paper also highlights the significance of some of the critical patient variables, including smoking, throat pain, and breath-related complications that had become significant risk factors in predicting lung cancer. Simultaneously, the value of deep learning regarding the ability to reproduce the structural abnormalities of CT scans that are not readily identifiable with the standard analysis was also confirmed.

These encouraging findings still have certain limitations. The use of publicly available datasets creates issues related to generalizability because the data might not be entirely representative of the real-world diversity of patients. Furthermore, the lack of medical knowledge in medical interpretation of the results of CT scans limits the quality of clinical validation. It will be critical to address those challenges using larger and more diversified datasets and stronger collaboration with medical professionals to enhance the reliability and impact of the system.

REFERENCES

[1] S. N. Tisha and S. A. Ani, “Predictive Insights: Empowering Early Detection of Lung Cancer Using Machine Learning Excellence,” in 2024 IEEE Region 10 Symposium, TENSYMP 2024, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/TENSYMP61132.2024.10752136.

[2] S. Murthy Nimmagadda, K. Likhitha, G. Srilatha, and S. M. Sree, “Lung Cancer Prediction and Classification Using Machine Learning Algorithms,” in Proceedings - 2024 International Conference on Expert Clouds and Applications, ICOECA 2024, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 1012–1015. doi: 10.1109/ICOECA62351.2024.00176.

[3] Mohammad Shafiquzzaman Bhuiyan et al., “Advancements in Early Detection of Lung Cancer in Public Health: A Comprehensive Study Utilizing Machine Learning Algorithms and Predictive Models,” Journal of Computer Science and Technology Studies, vol. 6, no. 1, pp. 113–121, Jan. 2024, doi: 10.32996/jcsts.2024.6.1.12.

[4] S. P. Maurya, P. S. Sisodia, R. Mishra, and D. P. singh, “Performance of machine learning algorithms for lung cancer prediction: a comparative approach,” Sci Rep, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-58345-8.

[5] Joy Chakra Bortty et al., “Optimizing Lung Cancer Risk Prediction with Advanced Machine Learning Algorithms and Techniques,” Journal of Medical and Health Studies, vol. 5, no. 4, pp. 35–48, Oct. 2024, doi: 10.32996/jmhs.2024.5.4.7.

[6] K. Moon and A. Jetawat, “Predicting Lung Cancer with K-Nearest Neighbours (KNN): A Computational Approach,” Indian J Sci Technol, vol. 17, no. 21, pp. 2199–2206, May 2024, doi: 10.17485/IJST/v17i21.1192.

[7] R. K. Pathan, I. J. Shorna, M. S. Hossain, M. U. Khandaker, H. I. Almohammed, and Z. Y. Hamd, “The efficacy of machine learning models in lung cancer risk prediction with explainability,” PLoS One, vol. 19, no. 6 June, Jun. 2024, doi: 10.1371/journal.pone.0305035.

[8] M. Mamun, A. Farjana, M. Al Mamun, and M. S. Ahammed, “Lung cancer prediction model using ensemble learning techniques and a systematic review analysis,” in 2022 IEEE World AI IoT Congress, AIIoT 2022, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 187–193. doi: 10.1109/AIIoT54504.2022.9817326.

[9] S. Agarwal, S. Thakur, and A. Chaudhary, “Prediction of Lung Cancer Using Machine Learning Techniques and their Comparative Analysis,” in 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions), ICRITO 2022, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/ICRITO56286.2022.9965052.

[10] P. Chaturvedi, A. Jhamb, M. Vanani, and V. Nemade, “Prediction and Classification of Lung Cancer Using Machine Learning Techniques,” IOP Conf Ser Mater Sci Eng, vol. 1099, no. 1, p. 012059, Mar. 2021, doi: 10.1088/1757-899x/1099/1/012059.

APPENDICES

Appendix 1 Report Diary

Appendix 2 Review Card I

Appendix 3 Review Card II

Appendix 4 Review Card: Viva

Appendix 5 Review Card

Appendix 6 Invention Disclosure

Appendix 7 Research Paper

Appendix 8 Consent Letter